Dorian Gray and the Wilde Senses

by Signe Juel Nordentoft

A Corpus Stylistic Analysis of *The Picture of Dorian Gray*

Dorian Gray and the Wilde Senses: A Corpus Stylistic Analysis of Oscar Wilde's *The Picture of Dorian Gray*

Signe Juel Nordentoft

Aalborg University

Supervisor: Associate Professor, PhD. Marie Møller Jensen

June 2nd, 2020

Abstract

Oscar Wilde's novel *The Picture of Dorian Gray* (1891c) offers an intriguingly sensuous reading experience. In the study reported here, two approaches within corpus stylistics were combined with cognitive stylistics and foregrounding theory, pursuing three aims. Firstly, the goal was to investigate stylistic characteristics of the novel via a bottom-up corpus-driven methodological approach. Secondly, the colour motif was analysed in a top-down corpus-assisted approach. Thirdly, the two approaches were compared in terms of results yielded, and difficulties and advantages encountered. Throughout both analyses, the corpus methods yielded the quantitative data used as a stepping stone for the qualitative analysis, founded in cognitive stylistics and foregrounding theory. The corpus-driven analysis yielded innovative and rather dissimilar results, pointing towards multiple stylistic effects: the creation of dramatic effect; an effect of auditory stimulation; defamiliarization processes of beautiful entities; and a foregrounding of subjective intuitive feeling. The corpus-assisted analysis yielded focused results regarding the way in which the colour motif comes to be perceived as foregrounded. First of all, this analysis revealed that colour adjectives are used significantly more frequently in *The Picture of Dorian Gray* than in the three reference corpora consisting of Wilde's works, Gothic literature, and literary works of Wilde's contemporaries. It also revealed that the use of hyphenated compound adjectives is a unique characteristic of the novel and that a motif of possessions is attached to the colour motif, foregrounding dead objects in the text. Moreover, it was found that the novel's style is poetic and rich in the sense that the textual level is embellished and heavily adorned, making for an aesthetically powerful reading experience. The impression of a strong sensuous aspect was substantiated by the analysis, in which it was found that the text creates visual and tactile stimuli in the mind of the reader. In the comparison of the two approaches, it was found that the corpus-driven approach offers quite a few disadvantages, including a fragmented argumentation, disparate outcomes, and a detachment from the co-text. These issues have to be

amended by the stylistician at some cost of time but resulting in an analysis with high validity and originality. In the comparison, it was also found that the corpus-assisted approach had a clear advantage in its intrinsic focus, making for a targeted and unified argumentation but at the cost of possible spontaneity and serendipity. Consequently, it can be argued that in stylistics a combination of methods can be beneficial, though no exact estimate of the optimal distribution of the approaches can be offered.

Table of contents

List of Tables and Figures

Acknowledgements

I would like to voice my deep appreciation and gratitude to the following people:

To my supervisor, Marie Møller Jensen, for introducing me to the world of modern stylistics, helping me overcome an old fear of statistics, and for general guidance and inspiration.

To my fellow student, former co-author and personal friend, Katrine Lottrup, with whom I have discovered a world of literature and stylistics. This thesis is no doubt a result of many hours of tireless and ambitious work together.

To my fiancée and soon-to-be husband, Andreas, for always supporting and believing in me. You have made this process feel free and easy.

To my family, for constantly cheering me on, for offering me a study refugium and lending me their home offices for weeks.

1. Introduction

Oscar Wilde (1854-1900) is a Victorian coryphaeus, though a subversive one. In his short lifetime, the Irish writer produced plays that are still staged around the world to this day. But Wilde was not only a playwright – he wrote poetry, critical essays, reviews, short story fairy tales – and a single novel of merely 231 pages[1]; *The Picture of Dorian Gray* (1891c). In an age of extreme self-contemplation, epitomized by the growth of social media platforms such as *Instagram* and *Facebook*, Oscar Wilde's novel has much to offer a modern readership. It is a novel that debates the idea of an extreme hedonistic lifestyle; the value of aesthetics versus ethics; the bidding of our impulses; the commodification of the self and others; and the relationship between art and moral.

Oscar Wilde was a key figure in the Aesthetic movement and Dorian Gray was his perfect dandy aesthete. Under the doctrine 'art for art's sake', Wilde explored moral and ethical questions by writing a Gothic story. It is a story about a portrait painted by the sensitive artist Basil Hallward who is infatuated with his muse, the beautiful and young Dorian Gray. When Dorian, in a moment of mortal agony, wishes for the portrait to age instead of himself, his wish is mysteriously granted. This leaves the protagonist Dorian ever-young and untainted by his ever more immoral and debauched conduct, as the painting grows increasingly hideous in secret. The 'spell' spawns his fateful chase for personal gratification, an aesthetic pursuit of fulfilment of the senses at the cost of all morality, ushered by the middle-aged unscrupulous hedonist, Lord Henry Wotton. In the end, Dorian tries to rid himself of the portrait by tearing the canvas, whereby the physical stains of the painting are transferred onto his person, and he ends up murdering himself.

[1] My edition – Wilde (1891c).

The Aesthetic movement grew out of Aestheticism, a complex philosophical movement that "began by emphasizing the power of visual art and beauty to regenerate social life, and ended by insisting on the complete independence of art from life and on the alienation of the artist from society" (Dowling, 1998). The Aesthetic movement is sometimes referred to as the cult of beauty, and "no one did more to popularize the Aesthetic movement than Wilde, partly in a spirit of self-parody, but partly too as the living symbol of a serious commitment to Aestheticist values" (ibid.). In the Aesthetic movement in general and in *The Picture of Dorian Gray* specifically, the senses hold a special and valued position as "the dandy-aesthetes of the fin-de-siècle period above all honoured their senses" (Calloway, 1997:34). Wilde's *Intentions* (1891a) offers a confirmation:

> Aesthetics are higher than ethics. They belong to a more spiritual sphere. To discern the beauty of a thing is the finest point to which we can arrive. Even a colour-sense is more important, in the development of the individual, than a sense of right and wrong (Wilde, 1891a:222)

Holland (1997) dubs him a "musician of words and painter of language", a description that appears to be in line with the style used in the novel. *The Picture of Dorian Gray* offers a reading experience in which the senses are activated and stimulated. The novel exudes sensuous stimuli and Wilde used language to create effects of physical stimulation through the words on the pages. Such is the impression evoked by the novel, and given the emphasis on *effect* and *language*, stylistics offers a well-suited theoretical framework for further investigating such an impression.

Much has been written about Wilde's person – about his clothes, his interior decorating, his book covers, his friends, lovers, and enemies[2]. Yet not much is said of the *style* of this writer, nor in fact much about his works that does not involve a biographic analysis of the

[2] For example, see Ellmann (1988) or Calloway (1997) whose studies of Wilde are largely biographic.

author. Though he surely led a fascinating life, his works almost appear to suffer under it, being *de facto* eclipsed by their scandal-ridden author. Therefore, I set out to investigate *how* Wilde writes, how he achieves *effects*, and why and how reading his novel has an impact on the reader. Not many stylisticians have made endeavours to analyse Wilde's works and I have found no published works concerning the style in *The Picture of Dorian Gray*. In the traditional field of literary criticism, however, there are several good examples of in-depth analyses of Wilde's prose. I will return to these critics' conclusions and compare them to mine in a separate section in the discussion chapter.

In a digital era, the study of literature is increasingly informed by new progress in computational methods. Ready-to-use software programmes are easily accessible and literary works are digitized, facilitating the analysis of style in the fairly new and continuously expanding discipline of *corpus stylistics*. The present study bears some resemblance to the study *Floral Foregrounding: corpus-assisted, cognitive stylistic study of the foregrounding of flowers in Mrs. Dalloway* (Jensen, Lottrup & Nordentoft, 2018). The present thesis observes much of the same structure as the article, as it was then concluded that the combination of corpus stylistics and cognitive stylistics led to a stronger insight regarding the construction of effects created by language but also regarding the reception, that is, the impact a text has on a reader. *Cognitive stylistics* is also used in the present thesis as the branch offers valuable terminology for analysing the creation and perception of effects – a terminology that also merges well with *foregrounding theory*, as was concluded in the previous study. The previous study was founded on a top-down corpus-assisted approach, whereas this thesis sets out to test the combination of the latter *and* a bottom-up corpus-driven approach[3]. I use both approaches in order to generate interesting analyses of Wilde's novel, but also in order to compare the methods and assess the

[3] The approaches are introduced and discussed in section 3.1.

outcome of combining the inductive and the deductive[4]. Therefore, my analysis is tripartite and is concerned with the following research questions:

In the corpus-driven analysis I seek to analyse Wilde's style and the effect that his style creates in the mind of the reader. In the corpus-assisted analysis, this goal is still pursued but with the focus on the colour motif and testing of the following hypothesis[5]: In *The Picture of Dorian Gray*, colours hold a unique position and impact the reading experience. Moreover, I seek to compare the use of a corpus-driven versus a corpus-assisted corpus stylistic approach.

Having introduced the purpose(s) of the thesis, I will briefly outline the structure. In chapter 2, the field of stylistics is first positioned in its academic context and important discussions within the field are introduced. Subsequently, the specific theoretical framework is introduced, starting with cognitive stylistics, then introducing foregrounding theory, and finally corpus stylistics. In chapter 3, all the methodological considerations appertaining to corpus stylistics (also general but mostly thesis-specific) are presented, such as the building of corpora, the use of software and statistics, and hypothesis testing. In chapter 4, the first corpus-driven part of the analysis is carried out with a brief conclusion before moving on to the second corpus-assisted part of the analysis in chapter 5. In chapter 6, I conclude on and compare the results of the two analyses, before evaluating the findings. Finally, I assess the two approaches in a methodological discussion.

[4] See section 6.3.

[5] The hypothesis is refined and made operational in section 3.4.

2. Theory

In the following theory chapter, I begin by introducing the field of stylistics broadly, explaining what stylistics is and what characterizes the discipline. Subsequently, I outline the history of stylistics from its earliest progenitors to its current state. The historical account is included in order to position the field in the wide landscape of literary studies and to point out the traditions, schools, and conventions that have informed stylistics. This account is followed by an outline of the contemporary state of the field and the shared underlying assumptions and principles ruling the discipline – essentially what could be labelled a theory of stylistics.

Once this framework has been established, the contemporary field of cognitive stylistics (spurred by a cognitive turn in the humanities (Stockwell & Whiteley, 2014:2)) is introduced along with the somewhat affiliated schema theory and ICM theory, which offer invaluable terminology for approaching *the readerly experience*. The notion of ICM is introduced as it is a prerequisite for the subsequent introduction of foregrounding theory, which offers practical methods for pinning out what comes into focus in the course of a reading process. The theory section is concluded by an assessment of the balance between quantitative and qualitative methods and the debates regarding this issue.

2.1. Stylistics

Stylistics is an approach to literary texts primarily, in which the meaning-making is investigated through means of a linguistic theoretical framework. For this reason, stylistics can also be referred to as *literary linguistics* (Burke, 2014b:1). Simpson defines the nature of stylistics as "a method of textual interpretation in which primacy of place is assigned to *language*" (2014:3. his italics). This refers to that linguistic affiliation, and he goes on to specify that "while linguistic features do not themselves constitute a text's 'meaning', and account of linguistic features nonetheless serves to ground a stylistic interpretation and to help explain why,

for the analyst, certain types of meaning are possible" (ibid.:3). The kinship between linguistics and stylistics is unambiguous. Stylistics is a field that champions scientific principles and ideals of objectivity. Although this objectivity often proves illusory according to Stockwell and Whiteley who qualify the field as an "intersubjective discipline" (2014:5). They go on to characterize stylistics as a discipline that is "*progressive, systematic, transparent, replicable, evidential* and *textually grounded*" (2014:4. their italics). It is progressive because approaches that prove inadequate are abandoned and almost never revisited. It is systematic for three reasons. Firstly, because "features of language are viewed within a generally consistent theory of language", secondly because "the terms of the analytical framework are clearly set out first and then applied rigorously", and thirdly because "the objects of investigation (…) are available for investigation in the world outside the domain of stylistics" (ibid.:4-5). It is transparent (or aims to be) because good stylistic analyses rely on common technical terminology and because "obfuscation and deliberate obscurity are not well regarded" in the practice of stylistics (ibid.:5). It aims towards being replicable, though readings of texts probably always diverge to some extent. To achieve a degree of replicability "analytical explanations are offered in an open and transparent articulation precisely so that later readers and stylisticians can see the working of the analysis (…)" (ibid.:5). Finally, stylistics is evidential and textually grounded because "stylistics arguments are only presented for verification if they are accompanied by data from the literary work or reading" (ibid.:5). While Stockwell and Whiteley propose five characteristics of stylistics, Carter suggests that a good stylistic analysis must be "open, evidenced and retrievable" (2010:68). Likewise, Simpson points out "the three Rs" for proper stylistics, stipulating that stylistics should be *rigorous, retrievable,* and *replicable* (2014:4-5). Carter and Simpson essentially agree since there is very little deviation between their tripartite criteria. These principles do, whether formulated as an abbreviated three or an expanded five criteria, constitute a part of what we might call *the theory of stylistics* which I will return to.

But before moving on to the slightly complicated task of formulating a theory of stylistics, I will position the field and discipline of stylistics in a historical context.

2.1.1. Stylistics from Antiquity to present time

Stylistics is a discipline that is rooted in classical Antiquity's rhetoric and poetics, and much of the linguistic terminology stems from these practical disciplines. In traditional rhetoric, stylistics can be linked to the third of Aristotle's five canons, which is *elocutio* (originally *léxis* in Greek), the oral expression in which the stylization process takes place. In the tradition of rhetoric, style was thus a sub-element in the art of persuasion, rather than a textual 'feature' to be studied in its own right (Burke, 2014a:20-21). However, the field was largely unexplored and moribund until the Russian formalists showed a renewed interest in the discipline at the beginning of the 20th century – an interest that did not spread to Europe until these formalists' works were translated in the 1960's (Burke & Evers, 2014:32). Thus, stylistics, reinvented as an academic discipline, did not really exist until the early 1900's (ibid.). Bradford explains that the two most influential groups, the formalists in Russia and Europe, and the New Critics in Great Britain and North America even "remained within mutually exclusive geographical and academic contexts" until the 1950's (1997:12).

The Russian formalists are among the most important seminal linguistics scholars and include Viktor Šklovsky, Vladimir Propp, and Roman Jakobson among others (Wales, 1989:184). Jakobson, a founding member of the Prague School created a link between the Russian formalists and the European structuralists, and when the formalist movement was suppressed by Stalinism in the 1930's, the structuralists continued the studies of stylistics (Burke, 2014b:2 and Burke & Evers, 2014:31). According to Fialho and Zyngier, it is possible to accurately date the formalists' bridging between linguistics and literature (2014:330-331). They attribute this event to Roman Jacobson's seminal paper which he first delivered in 1958 at a conference in Indiana – a paper that introduced "'poetic language' as communicative language

that, differently from other communicative acts, focuses on the message for its own sake" (ibid.:331). The formalists' primary focus concerned the idea of 'literariness' and they sought to define how such structures as parallelism or linguistic deviation serve to make a text *literary* (Nørgaard, Montoro, Rocío & Busse, 2010:2). It should be mentioned, however, that the formalists and the New Critics did not consider literary style to be confined to literary texts, as they recognized the wide use of literary stylistic devices in non-literary forms such as advertisements, conversations, political speeches, etc. (Bradford, 1997:13; Simpson, 2004:3). The New Critics, being often less mentioned inheritors of the methods and goals of stylistics, were a group of teachers and writers from North America and Great Britain (Bradford, 1997:12). What they had in common with the formalists was the desire "to define literature as a discourse and art form and to establish its function as something that can be properly studied" (ibid.:12). This is in line with Stockwell and Whiteley, as they define stylistics as "the proper study of literature" (2014:1).

Jumping forward to the state of present-day stylistics, the most common qualifier seems to be *interdisciplinarity*. Burke suggests that "contemporary stylistics goes far beyond the rhetoric, poetics, formalism, structuralism and functionalism of the past to embrace corpus, critical, cognitive, pedagogical, pragmatic, gender, multimodal and, most recently, neuroscientific approaches" (2014b:2). The growth in the field of stylistics is in part due to its continuous enrichment in methods, originating from "theories of discourse, culture, and society" (Simpson, 2014:2), and according to Wales, the 'openness' of the stylistic tool-kit results in its ever-expanding nature, "while remaining true to its core" (2014:35).

Indeed, a great many branches of stylistics exist today: cognitive, corpus, feminist, systemic-functional, critical, functional, historical, multimodal, pragmatic, pedagogical, etc. Each form has its own underlying theoretical assumptions (not to mention methods), which is why a description of a theory of stylistics can become a lengthy affair. However, Toolan suggests

there are a number of common underlying assumptions and that, despite the influences from several large disciplines, it is possible to articulate a theory of stylistics (2014:13-14).

What stylistics offers, is an ability to pinpoint and isolate proof of what the literary critic may intuitively sense in a text. Verdonk articulates the objective of stylistic analysis as offering "precise linguistic descriptions, which can substantiate otherwise impressionistic interpretations of literary texts" (2004:126-127). This brings me to the connection between literary criticism and stylistics, where divergent attitudes are found. Wales explains that the question of the affiliations or disciplinary kinship of stylistics is uncertain and polemic:

> Given the traditional emphasis on linguistic levels, however, and a deep knowledge of how language works, some would argue that stylistics is itself a sub-discipline of linguistics. Others would argue, given its practice of close reading, its goals and main subject matter or genres, that it is a sub-discipline of practical criticism, itself a sub-discipline of hermeneutics, the art of interpretation (2014:36)

In the 1970's, Widdowson, an EFL stylistician, argued for a strong kinship between these fields, and for the perception that stylistics and literary criticism exist on a continuum. He continued to suggest that stylistics was "particularly valuable as a preparation for literary study" (Hall, 2014:244). This attitude indicates a perception of a hierarchical relation between stylistics and literary criticism, in which the latter holds a superior position. More recently, Burke stated that stylistics is a field that "confidently has one foot in language studies and one in literary studies" (2014b:2), and he goes on to argue that "stylistics encourages literary criticism to be about more than just opinions" (ibid.:3), indicating an association between the two fields, although the exact relation appears less precise. Stockwell and Whiteley problematize this assumption of a kinship between stylistics and literary criticism, as they argue that the latter has transformed "into a form of cultural studies and historiography" (2014:3). They go on to argue

that stylistics should *replace* literary criticism as the central field in English departments, and moreover, they "insist that any approach to literary study that does not engage closely with the language in which the literary work appears is by definition indirect, distracted, partial and improper" (ibid.:1). However, at the core, literary criticism and stylistics do engage with the same subject matter – *literature*. This notion brings me to the theory of stylistics, in which the notion of literature plays a pivotal role.

2.1.2. A theory of stylistics

One of the clearest attempts at formulating an up-to-date and comprehensive theory of stylistics is found in Toolan's contribution to *The Cambridge Handbook of Stylistics* (Toolan, 2014). Quoting the OED definition, he defines *theory* as consisting of "the knowledge or statement of the facts on which it [the subject red.] depends, or of its principles or methods as distinguished from the *practice* of it" (Toolan, 2014:13. my insertion. his italics). Therefore, he offers eleven statements "that most stylisticians take as foundational knowledge or facts" (ibid.:16). I offer a summarized paraphrase of these facts:

1. In stylistics, it all starts and ends with language. Language is central.
2. Literature holds a special position in most cultures. Leisure time is an important prerequisite for readers of literature to even exist.
3. Literature can take on many forms and is special. It is often complex and sophisticated.
4. Literature depends on and challenges linguistic forms, which is also one reason why it is interesting to investigate literature.
5. Literature is the result of a writer's critical choices.
6. Context is important and must never be neglected by the stylistician.
7. Most societies ascribe a special value to literature. Literature thus holds a powerful position in most societies.
8. Exhaustive stylistic analysis is impossible.

It cannot be determined when a stylistic analysis is complete or incomplete – this question depends on the purposes of that analysis.

9. Stylistic analyses have the ability to produce extraordinary insights or adjustments in the way we think about language and literature.
10. Stylistic analyses negotiate assumptions about literature.
11. Literary language is not radically different from other linguistic products. Literature is different because it is treated differently and because it has a radically different purpose than other linguistic acts (list paraphrased from Toolan, 2014:16-20)

In Simpson's introduction to modern stylistics, he observes many of the same assumptions regarding language and literature and their role in stylistics (2014:2-5). It is apparent now that questions of linguistics (language in use) and of literature are the most dominating in formulating a theory of stylistics and these cores assumptions should be kept in mind throughout the analysis.

2.2. Cognitive stylistics

As outlined above, interdisciplinarity is a central characteristic of modern stylistics. One of the most significant interdisciplinary advancements in the field is what is sometimes referred to as *the cognitive turn* (Simpson, 2014:40). This turn, prompted by advancements in the fields of cognitive psychology and Artificial Intelligence, took place during the 1990's and greatly affected most research in the humanities (Stockwell & Whiteley, 2014:2). The term 'cognitive' has to do with human cognition, that is, the conscious, mental processing of stimuli from the world that takes place in the human brain. Cognitive stylistics contribute with a readerly aspect in stylistic analysis and "supplement, rather than supplant, existing methods of analysis" (Simpson, 2014:40). Stockwell and Whiteley argue that "when stylisticians have drawn on insights from cognitive science, we have been able to offer analyses of readerly knowledge and experience, feelings and emotions, imagined worlds, metaphors, allegories, and the valuations

of social significance and personal affect" (2014:3). A new array of methods, models, and terminologies have thus enriched the stylistician's toolbox considerably. One of the chief theoretical improvements made possible by the cognitive turn has to do with the question of literary language. As outlined in section 2.1.2., in item 11, literary language is not considered to be essentially different from all other language – a claim that was polemical and frequently debated in stylistics. As Simpson asks – how can it be claimed that literature is linguistically *different* from other forms of discourse if the underlying assumption is that it is not? (2014:40). Having changed the focus from the writerly dimension to a readerly one, this problem is however solved because, as Simpson says, "literature is perhaps better conceptualized as a way of reading than as a way of writing" (ibid.:40). Therefore, 'literature' should be understood as a specialized reading strategy that the reader adopts. Taking this as foundational knowledge, it is then possible to address two of the core issues in cognitive stylistics: "what do people do when they read?" and what happens to people when they read?" (Burke, 2006:218). These questions are important in the analyses of the present thesis.

According to West, cognitive stylistics is distinct from other approaches as it constitutes a proper "theory of literature, one that has an explicit *object of study*, a principled *methodology* and a clear and significant *purpose*" (2016:110. his italics). Accordingly, the object of study, he says, is "readerly experience" (ibid.:110). West explains that the readerly experience emerges from the interaction of two forces. Firstly, "the words on the page, (…) the text's semantic, syntactic, and sonic (or phonetic) features, which act as stimulus to evoke complex thoughts and emotions in the reader" (ibid.:110). Secondly, "the reader's cognitive faculties, which inevitably intervene in, and shape, her or his experience of the text" (ibid.:110). The methodology is clear, he argues, since it consists of "combining a microscopic analysis of the literary text's linguistic features (this is the *stylistic* element) with an account of what is known

from cognitive psychology about the ordinary workings of the human mind (this is the *cognitive* element)." (ibid.:110). He claims that the purpose of cognitive stylistics is the exploration of human cognition rather than the exploration of certain texts, writers, or genres, and there are fundamental questions which cognitive stylisticians seek to answer:

> What is literature? What distinguishes it from other uses of language? What accounts for the breathless fascination that we have for literature? Where does literature come from, and what function does it play in human life and in the evolutionary development of the species? (West, 2016:111)

While these are not the questions asked in this thesis, it is important to stress the scope of cognitive stylistics in order to understand the discipline's potential. This is one of the subbranches where the intersubjectivity is most apparent – while we do not all conceptualize the world in the same way, there are obviously some similarities, and these similarities are interesting as a starting point for synthesizing and generalizing ideas about literature. This is where *schema theory* must be introduced.

2.2.1. Schema theory and the idealized cognitive model

One of the key notions within cognitive stylistics is the term *schema*. Originating from advances in (gestalt) psychology and work in Artificial Intelligence, a schema (sometimes referred to as *frame* or *script*) "is a cognitive structure which provides information about our understanding of generic entities, events, and situations, and in so doing helps to scaffold our mental understanding of the world" (Emmott, Alexander & Marszalek, 2014:268). This is a central term because it helps us to understand how we as readers can fill in the inevitable gaps in a text by using schemas from our long-term memory. Nørgaard et al. explain that "schema theorists claim that meaning is not only contained in the text; meaning needs to be built up by the reader using the text in negotiation with their own background knowledge" (2010:8). An

important distinction is made by Rumelhart and Ortony who argue that there are two complementary and inter-dependable processes involved in reading: "bottom-up or stimulus-driven processes and top-down or conceptually-driven processes" (1977:128). Top-down processing has to do with a reader's prior knowledge (outside the text), whereas bottom-up processing refers to the textual elements or "linguistic triggers" (McIntyre, 2014:152). These processes are much in line with West's (2016) definition of the readerly experience as introduced above.

One of the most quoted examples is Stockwell's 'restaurant schema', "which gives an understanding of what restaurants are, what they look like, what sorts of things one would expect to find in them, how to go about ordering food, paying, and so on" (2003:255). In schema theory, Rumelhart and Norman have identified three processes that affect schemata: "*accretion, tuning, and restructuring*" (1978:38. their italics). Accretion is the "daily accumulation of information", a gradual and usual expansion of our knowledge (ibid.:38). Tuning is the minor adjustment of schemata, and restructuring is the creation of an entirely new schema, which may be prompted by the dividing, combining, or altering of existing schemata. The notion of *sensory schemata* is of special value to this thesis, as it can help explain how a reader is aware of the senses when reading. Emmott et al. position that "there is now substantial evidence that, as we read, brain areas which would be activated for real physical and emotional responses in the real world are activated for their imaginary equivalents in stories" (2014:273). This is important when trying to explain how the sensory (particularly the visual as the focus in this thesis is largely on *colour*), the reading experience, and the literary correlate.

Even though schema theory is considered to be somewhat outdated (see Stockwell 2003:253), I introduce it here as it is a framework for understanding the notion of the *idealized cognitive model*, or the ICM, as it was coined by Lakoff (1987:68)[6]. An ICM is an organized mental structure – a prototypical image based on our accumulated experiences and knowledge,

[6] In accordance with Lakoff's conventions, ICMs are rendered in small capitals.

and as such, it functions as a mental short-cut because "we do not (…) need to have a fully elaborated *textual representation* of a concept in order to set in motion a *cognitive representation* of that concept"(Simpson, 2014:43. his italics). This idea of a short-cut is in line with Stockwell's qualification of what it is we do with these ICMs: "Cognitive models are what cause prototype effects and our sense of basic categories" (2002:33). The notion of ICM is thus closely linked to prototype theory and an often-mentioned example is the concept of A BACHELOR (Fillmore, 1982). The concept does not simply denote 'an unmarried man' since we would not tend to think of unmarried men like the Pope, a seventeen-year-old or Tarzan as bachelors. Croft and Cruse explain the example like this:

> There is a sense in which the meaning of bachelor really is just 'adult unmarried male'. It is just that the BACHELOR is profiled against a frame that does not accommodate the variety of actual social statuses found in the real world. The frame for BACHELOR represents an idealized version of the world that simply does not include all possible real-world situations (2004:28)

ICMs are not neat and ordered structures but rather they develop and change through time, in different contexts and for different purposes. Some ICMs may even be better conceptualised as clusters of more ICMs. Lakoff gives the example of the MOTHER ICM (a cluster ICM), which he argues consists of multiple ICMs: BIRTH, GENETIC, NURTURANCE, MARITAL and GENEALOGICAL (Lakoff, 1987:74-76). All these ICMs do not necessarily apply to all kinds of mothers, and deviation is often lexically indicated (e.g. using an adjective + noun form or a conventional compound), like for instance 'stepmother', 'foster mother', 'birth mother', 'genetic mother', or 'unwed mother' (Croft & Cruse, 2004:31). This serves to show that ICMs can be more or less complex and indicates that culturally charged ICMs may tend to be more complex. The notion of ICM is valuable in relation to cognitive stylistics but also in relation to yet another framework of this thesis – that is the important notion of *foregrounding*.

2.3. Foregrounding theory

The notion of foregrounding originally stems from Prague structuralists Havránek and Mukařovský (Emmott & Alexander, 2014:329). The term has to do with the creation of *relief* in a text, by which certain textual elements appear foregrounded against a less noticeable background. The theory is rather broad and flexible, and as Emmott and Alexander explain, "foregrounding is recognized as occurring at all linguistic levels" (2014:329). They go on to list examples of common types of foregrounding which are "sound play, unusual graphic patterning, excessive lexical and pronominal repetition, unusual word choices, highly creative metaphors, parallelism, and breaches of the usual discourse structure" (ibid.:329). Nørgaard et al. argue that "foregrounding is a relative concept because it can only be measured if norms and conventions are established and observed and if these are related to complex contextual features" (2010:95). Thus, an element cannot be foregrounded before the 'background' is properly established, which can be done in multiple manners. In the case of *The Picture of Dorian Gray*, an example of one such norm could be the frequent use of predicators with almost every noun, as exemplified in this short extract:

> About half past eight I passed by an absurd little theatre, with great flaring gas-jets and gaudy play-bills. A hideous Jew, in the most amazing waistcoat I ever beheld in my life, was standing at the entrance, smoking a vile cigar (Wilde, 1891c[7]:48. my underlining)

Once Oscar Wilde has established such a norm, any passage that does *not* contain any such predicators would stand out – it would be foregrounded. Here it is important to make the distinction between *internal* and *external* foregrounding (Simpson, 2014:53-54). The passage above is externally foregrounded because it contains so many adjectives, which is not 'usual'

[7] Henceforth, when referring to *The Picture of Dorian Gray*, I will use Wilde's initials: OW. References to all other works by Wilde will still include year published.

– it is not a norm. An extract from the novel with few or no predicators would be an example of so-called *internal* foregrounding, as the norm to have many adjectives is one established *inside* the text, or as Simpson puts it "a kind of deviation within a deviation" (2014:53). Emmott and Alexander suggest that foregrounding may be used for various purposes, such as "highlighting specific key points, producing thematic meaning, prompting an emotional response, and yielding iconic effects" (2014:329). Some stylisticians argue that the use of foregrounding in literary texts is more structured and valuable, but this assumes that literary language is rather different from other types of discourse (ibid.:330). Like Emmott and Alexander, the notion of foregrounding is here used to denote "any type of language use which may be assumed to prompt attention, regardless of whether it has literary value" (ibid.:330).

According to Simpson, foregrounding may occur if the text shows *repetition*, *parallelism*, and/or *deviation* (2014:52). Foregrounding through repetition and parallelism, he says, can be understood as a "more of the same" type of foregrounding, as opposed to foregrounding as "a deviation from a norm" (ibid.:52). The concept of deviation from a norm is further explicated below, in the guise of what is often referred to as 'defamiliarization'.

2.3.1. Defamiliarization

Foregrounding is related to the equally important notion of *defamiliarization* (the common translation of the original 'ostranenie', meaning 'making-strange'), which was first introduced into literary criticism by Russian formalist Šklovsky in 1917 (Gavins, 2014:196). It refers to the process of representing the known as if it were in fact unknown. The goal of defamiliarization is "to increase both the effort and the duration of a perceiver's [or reader's] experience of a text, a purpose that is in explicit opposition to *habitualisation* or *automatisation,* the aesthetic deadening that results from an over-familiarity with language" (Emmott & Alexander, 2016:290. their italics). Defamiliarization is related to the very nature and purpose of art, which becomes apparent in this famous, humorous, yet ominous passage by Šklovsky:

> Habitualization devours work, clothes, furniture, one's wife, and the fear of war. (…) art exists that one may recover the sensation of life; it exists to make one feel things, to make the stone *stony*. The purpose of art is to impart the sensation of things as they are perceived and not as they are known. The technique of art is to make objects 'unfamiliar', to make forms difficult, to increase difficulty and length of perception because the process of perception is an aesthetic end in itself and must be prolonged. *Art is a way of experiencing the artfulness of an object: the object is not important* (2012:12. his italics)

Like foregrounding, defamiliarization is an important term in relation to cognitive stylistics as it has to do with the readerly experience and the reader's cognitive efforts, which is evident from the above quote. Moreover, Mukařovský relates the theories of foregrounding and defamiliarization (and its opponent automatisation) to the field of schema theory by stating that "automatization schematizes an event; foregrounding means the violation of the scheme" (1964:19). Furthermore, cognitive stylistics can be linked to the theory of foregrounding, given a parallelism to two important terms, entrenchment, and salience, which I will now introduce.

2.3.2. Entrenchment and salience

A parallel can be observed between terms from foregrounding theory and cognitive linguistics, which becomes instrumental in this thesis:

> From a cognitive stylistics point of view, the deviation from the norm type of foregrounding could be connected to the notion of salience, whereas entrenchment is more closely related with more of the same foregrounding. This is a rough pairing up of the theories as they are not developed to complement each other (Jensen et al., 2018:39)

Some terms demand clarification. The first term is *entrenchment*. In most language use, the reader's cognitive processing (decoding or encoding language) hardly ever reaches a conscious

level of awareness – at least not when it comes to familiar entities (Schmid, 2007:118). Little to no processing time is needed to encode or decode the concept of a 'dog', as this exists in "some kind of prepackaged, ready-made format" (ibid.:118). This prepackaged format is what is referred to as an ICM in this thesis. The almost non-existent need for cognitive effort has to do with the fact that the concept of a 'dog' is deeply ingrained in our long-term memory – it is entrenched. But not all concepts are equally entrenched as some may require more effort, either because they are concepts that have been defamiliarized or because the reader's ICM of that concept is not well established. One such less well-entrenched concept could be a DUCKBILL – an animal that we rarely encounter (here in the West), which is why the ICM for it is likely to be less established and 'prepackaged', and as Schmid states, "cognitive units come to be entrenched and their activation automated to the extent that they have been used before" (ibid.:118). This also entails that a less well-entrenched concept can become more entrenched (or re-entrenched) during the reading process. If a text had frequent mentions of a duckbill, then the concept would cease to require cognitive effort on the reader's part, although it may be difficult to establish how fast this happens. Moreover, that which is entrenched can become salient through frequent repetition or notable parallelism.

While the notion of entrenchment is related to an individual's cognition and personal experience (one person's duckbill could be another person's dog), it is also possible to elevate the notion of entrenched concepts to the level of societies, or linguistic communities, since "the frequency of occurrence of concepts or constructions in a speech community has an effect on the frequency with which its members are exposed to them" (ibid.:119). Bearing this in mind, it thus becomes possible to make statements about what is and is not *typically* entrenched, without losing sight of such a statement being relative, and culturally and contextually dependent.

The second term is *salience*. According to Schmid, there are two main types of salience; *cognitive salience* and *ontological salience* (ibid.:119). The first type concerns speech (or reading) events, where salience may be (temporarily) activated in two ways: "the activation of a concept may be controlled by a conscious selection mechanism, whereby the concept enters a person's focus of attention and is being processed in current working memory" or, "a concept may be activated through *spreading activation*, which occurs when the activation of one concept (e.g. DOG) facilitates the activation of others (e.g. BARK, TAIL WAGGING, FUR, POODLE, ALSATIAN, COLLIE, etc.)" (ibid.:119). He concludes that a concept is cognitively salient "if it has been loaded, as it were, into current working memory and has thus become part of a person's center of attention" (ibid.:119).

The second type, ontological salience, has to do with the way we normally perceive the world, and thus relates to a more permanent type of salience as "by virtue of their very nature, some entities are better qualified to attract our attention than others and are thus more *salient* in this sense" (ibid.:120). For example, a person is more salient than the pavement she is walking on. Langacker suggests that there are some fundamental principles when it comes to defining what we usually perceive as salient, because "a whole is more salient than its parts; a physical object is more salient than an abstract entity; and a person has maximal cognitive salience" (1991:171). It should be kept in mind that there is an interplay between the degrees of entrenchment and salience. To offer an example of this interplay, I will return to Oscar Wilde's "vile cigar", from the previously given excerpt (OW:48). The ICM of a CIGAR is fairly entrenched. 'A vile cigar', however, is a less well-entrenched form and the unexpected predicator 'vile' creates an effect of defamiliarization, making it cognitively salient[8]. In a hypothetical example where Oscar Wilde keeps referring to a cigar as a 'vile cigar', the form would become less

[8] Had it been a 'pink cigar', then it would have been ontologically salient as well.

salient as the reading progresses, and while losing salience, the concept would become re-entrenched.

Having now introduced some key concepts within cognitive stylistics and foregrounding theory, the foundation for much of the following qualitative analysis, I will move on to the theoretical scaffold of the quantitative analysis, which is *corpus stylistics*.

2.5. Corpus stylistics

A corpus is – simply put – a collection of words. Corpora vary drastically in size and content, from a single novel to huge collections of all types of language in use, sometimes containing millions of words, or 'tokens' as they are traditionally referred to. Corpus stylistics is a relatively new field that approaches literary text using computer-assisted quantitative methods and corpora (Mahlberg, 2014:139). The field experienced considerable growth throughout the 2000's, partly due to the accessibility of computer-readable texts provided by *Project Gutenberg*, and to the accessibility of free and specialized software for this purpose (Mahlberg, 2007:21-22). Theoretically, it does not deviate from the common assumptions in stylistics as listed in section 2.1.2., but it is, qua its specialized methods, a field that must be positioned within the wider traditions of stylistics. Corpus stylistics originates from and is inspired by the methods of corpus linguistics, a field that is concerned with the analysis of *patterns* in large corpora of data, and whereas corpus linguistics tends to focus on general patterns and tendencies in language, "a corpus stylistic analysis embraces the language of individual texts by providing frameworks against which these features can be identified, in terms of tendencies, intertextual relations etc." (Nørgaard et al., 2010:10). Thus, the object of analysis is one of the major differences between corpus linguistics and corpus stylistics, but they are both sub-disciplines in the wide field of digital humanities (Mahlberg, 2014:378).

Corpus stylistics is sometimes used synonymously with computational stylistics due to similarities in methods, even though there are differences between these fields (Archer,

2007:245-246). Whereas computational stylistics is linked to the disciplines of stylometry, stylometrics, statistical stylistics, and stylogenetics, corpus stylistics mainly observes two directions or approaches: corpus-assisted and corpus-driven (ibid.:245-246). The theoretical and methodological concerns and disputes generated by these approaches will be outlined in the methods chapter (section 3.1.). In this chapter, I will position corpus stylistics in relation to cognitive stylistics and foregrounding theory, in order to explain how these fields are worth combining. I will also describe the goals and advantages of using corpus stylistic methods. The particulars of this approach such as corpus design, the statistical significance of findings, the chosen concordancer and the e-text provider will all be reserved for the methods chapter.

Mahlberg's definition of corpus stylistics focuses on the field's kinship with corpus linguistics, as she highlights that "a crucial aspect of corpus stylistics is the fact that it combines methods and principles from both corpus linguistics and literary stylistics" (2014:378-379). She further stresses this kinship by positioning that "we approach a text both as an example of language and as a work of art" (ibid.:379). While corpus linguistics and corpus stylistics are not the same disciplines, there is theoretical common ground due to the shared origin. Mahlberg draws on corpus linguistics when she proposes a corpus stylistic circle in which "corpus methods open up an additional perspective on the linguistic analysis", which in turn would lead to greater literary appreciation (2016:140). This close connection between the *literary* and the *linguistic* is also evident in Fischer-Starcke's definition of the two main goals in corpus stylistics:

> 1. to study how meaning is encoded in language and to develop appropriate working techniques to decode those meanings, and
> 2. to study the literary meanings of texts (2010:1)

Fischer-Starcke goes on to define success criteria for including corpus linguistic approaches in stylistic practice. The merging of methods is thus successful when "literary insights or, more

generally, new and additional information on the data are gained that could not or have not been generated without electronic analyses" but also when "already known information on the data or previous interpretations of it can be supported or refuted by way of electronically generated data" (ibid.:20). These criteria are important in order to explain the validity of the theoretical framework, and when evaluating the results of the analysis, which is done in section 6.2. In her work on Dickens' *Great Expectations* (1861), Mahlberg concludes that a corpus stylistic approach to a novel is fruitful because it assists in revealing otherwise 'invisible' patterns as well as systematizing and describing intuitively present patterns (2007:31). The study of patterns is also, partially, what is at the core of foregrounding theory and as Nørgaard et al. position:

> The focus in stylistics on how a text means and what makes it distinctive in terms of norms allows for a productive interplay between corpus linguistic and stylistics, especially with regard to the theory of foregrounding, which discusses aspects that account for patterns and structures such as deviation and parallelism (2010:9)

Both parallelism and deviation may thus be approached in a text using corpus methods, which is part of my motivation for including it here. However, there are limits to the descriptive and analytical potential of corpus stylistics, and it is not to be considered as an automated process (Mahlberg, 2007:31). Mahlberg's conclusion to her *Great Expectations* analysis is somewhat circumspect, as she points out that "although associations between form and meaning are made visible with the help of corpus tools, there is no one-to-one relationship between form and meaning" (ibid.:31). The stylistician's interpretative role is thus of great importance if corpus methods are to be used.

Based on this productive relation between corpus stylistics and foregrounding theory, the theoretical structure of this thesis is both methodologically motivated and theoretically motivated. This leads me to the final component of the theoretical framework of this thesis: the balance between the quantitative and the qualitative approach.

2.6. Quantitative and qualitative approaches

It is evident from the introduced theoretic framework that this thesis is based on both qualitative approaches (from cognitive stylistics and foregrounding theory), as well as quantitative approaches (from corpus stylistics). This is an important issue in the polemic, as the balance between the quantitative and the qualitative is by no means set in stone. Stubbs introduces the controversy between literary scholars and stylisticians (and literary linguists for that matter) in the following way:

> Literary scholars are often sceptical of quantitative analysis, especially if this implies that literary style can be reduced to statistics about low-level language features (usually words and grammar). They object that deeper thematic and symbolic patterns must be interpreted, not only against linguistic norms but also against their historical and cultural background (2014:48)

Stubbs thus understands the feud as one that concerns the possibilities of interpretation inside and outside the text itself. However, to pour oil on troubled waters it should be pointed out that it is rare to encounter a literary critic who does not work in a text-oriented manner, often using close-reading as the primary approach. In this perspective, the quantitative and qualitative approaches are both very text-oriented and meticulous. Indeed, most stylisticians recognise the need for a contextualised interpretation, like Jeffries and McIntyre, who state that it "remains important to recognise that stylistics, whatever its detractors say, does not 'read off' meanings from linguistic features, so the discovery of a significant (statistical) difference between two texts, though interesting does not in itself constitute an interesting finding" (2010:172). Like

most modern stylisticians, they also recognise that "the question of why this difference is there often requires more detailed analysis both of *the co-text* and *the context*" (ibid.:172. my italics).

C.P. Snow's 1959 vision of a third culture, in which the humanities and the natural sciences become able to use and understand each other's methods, is an oft-mentioned ideal when defending the use of scientific methods in the humanities (Stubbs, 2014:48; Fialho & Zygnier, 2014:329-330). The resistance and protests against this idea are mentioned as explanations as to why scientific methods have not become more integrated into the humanities than they are today. However, it should be noted that the number of scholars who have in fact embraced scientific methods is enormous. The rise of digital humanities, literary linguistics, and cognitive approaches, just to mention a few, seems to indicate that scientific methods have been largely accepted and adopted in the humanities. Though this may be the case, it remains true that the mere counting of words, however elaborate, does not offer any literary insight until qualitative interpretations are introduced. This is also the basis for the present thesis. The purely quantitative analysis of texts is not an ideal in stylistics – not even in the early days of modern stylistics, when "stylistics features of a text were restricted to the narrow linguistic elements" (Stockwell & Whiteley, 2014:1). Furthermore, it should be specified that with the 1990's 'digital turn' in linguistics, the quantitative approaches have been extensively refined, which should also lead to a partial silencing of the, now somewhat dated, criticism of stylistics and its potentials (ibid.:2).

Though this chapter may seem to suggest that quantitative methods are an intrinsic part of stylistic analysis, this is not the case. Fialho and Zyngier have analysed published papers from the journal *Language and Literature* to determine what the current distribution of methods is (2014:332). Looking at 95 papers from 2007 to 2011, they found that only 25 included quantitative methods. This serves to show the current proportionality in the field. In the current thesis, quantitative approaches are used both as a driving force (the corpus-driven analysis) and as a

method for investigating the hypothesis (the corpus-assisted analysis). In both cases, it is held true that mere quantification does not offer insight, and thus both analyses will be based on a mix of quantitative and qualitative approaches.

The complete theoretical framework of the present thesis is now thoroughly outlined, positioning stylistics in the wider field of linguistics, and introducing the key theoretical points relevant for the thesis' purpose. In the next chapter, I will provide details of the different methods employed in this thesis, mostly relating to the specific technicalities of corpus stylistic discipline.

3. Methods

In the following chapter, I will introduce a number of methodological steps that form the practical foundation for this thesis. As foreshadowed, the methods primarily appertain to corpus stylistics which was introduced in section 2.5., and therefore it is essential to outline and discuss these steps in order to live up to the demand for replicability, as introduced in section 2.1. Having just introduced the 'macro-methods' in the final theory chapter, I move on to the 'micro-methods', meaning the specific corpus stylistic approaches used in the two analyses and the important considerations linked to them. I discuss the corpus-assisted and the corpus-driven approaches along with several polemic questions appertaining to these. Then I move on to the thesis-specific methods, all relating to the discipline of carrying out corpus stylistic analyses such as the question of designing the corpora, using a concordance software and a tagging software, and the issues concerning corpus annotation, and testing results for statistical significance.

3.1. Corpus-driven and corpus-assisted approaches

One of the most important aspects of this thesis' methodological structure is its combination of top-down and bottom-up approaches, which relate to the corpus stylistic approach specifically. The objective of this structure is, as described in the introduction, to assess whether such a structure is beneficial and whether it can provide new or different insights regarding stylistic characteristics of a novel. However, before proceeding to investigate this question, the methods must be defined and delimited.

The distinction between these approaches, commonly referred to as corpus-assisted[9] (top-down) and corpus-driven (bottom-up), was originally introduced in corpus linguistics by Tognini-Bonelli (2001). Soon, the distinction was embraced by stylisticians as well, and today,

[9] Sometimes corpus-based.

they remain the dominant types of approaches in the discipline. Tognini-Bonelli defines the corpus-assisted methodology as one "that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study" (2001:65). In other words, a corpus-assisted analysis sets out to test *a priori* made observations and the subsequently formed hypothesis – ideas about patterns and tendencies arise from the stylistician *before* any quantitative analysis has been conducted. This approach, she argues, is thus best suited to "quantify existing categories" (ibid.:65), that is, to offer quantitative support to qualitatively (or intuitively) made observations. This is in line with Fischer-Starcke's second criterion for the successful use of corpus stylistic methods[10] (2010). In order to remain methodologically valid, a theory (or hypothesis) should be clearly formulated, so that the analysis can remain focused, or as Tognini-Bonelli puts it, "the theory would have to be put into an explicit form so that those aspects of corpus patterning that it covered could be distinguished from those where the theory did not cover, or was at variance with, the evidence" (2001:65). What she criticizes about the corpus-assisted analysis is the potential for skewing the results by 'looking for the evidence' as it were[11]. She also occupies a sceptical position regarding the potentials of the corpus-assisted method, as she argues that corpus evidence is only used as "an extra bonus" to reinforce already made claims, rather than being controlling and pivotal for the analysis (ibid.:66).

However critical Tognini-Bonelli is of the corpus-assisted approach, it must be kept in mind that she is first and foremost a linguist and not a stylistician. Her critique may be warranted in linguistics but in stylistics, the corpus-assisted method remains instrumental in testing hypotheses about literary language. Another defence of the corpus-assisted approach is that the

[10] The second criterion is that "already known information on the data or previous interpretations of it can be supported or refuted by way of electronically generated data" (Fischer-Starcke, 2010:20).

[11] For an exemplification of this methodological danger, see Tognini-Bonelli (2001:15).

'test specimen' is not language use in general as is often the case for linguists, but rather a *specific and specialized* example of language use, a writer's particular articulation, demarcated by the first letter and the last period of the *literary novel*. In that light, a corpus-assisted approach and pre-defined hypotheses should be viewed as less controversial and categorical in their scope. The corpus-assisted approach offers significant advantages when it comes to the integration between literary criticism and stylistics because it recognizes the possible validity of *intuitive* findings. However, the corpus-assisted approach appears to have limited potential when it comes to *innovative* findings – that is, patterns that have not previously been detected by the stylistician before the introduction of corpus methods.

The second approach is bottom-up and inductive – the corpus-driven approach. For an analysis to be corpus-driven, the analyst must 'go in blind', so to speak. No presuppositions about the data can inform the analysis, and the analyst should have no ideas, however vague, regarding the outcome of the analysis. This approach is celebrated by Tognini-Bonelli, who claims that the method has the potential to bring about a "qualitative revolution" (2001:84). She claims that the corpus-driven approach is superior because the data as a whole is respected and because the methodology is clear-cut and unambiguous: "observation leads to hypothesis leads to generalization leads to unification in theoretical statement" (ibid.:85). Tognini-Bonelli praises this approach as she states that

> (…) every step taken in this direction seems to lead the scholar to uncover new grounds, posit new hypotheses and not always support old ones. This is where the change of attitude indirectly brought about by the computer becomes the most threatening for the linguistic status quo; the evidence that comes to light has to be either rejected by argument or respected – it cannot be ignored (2001:84)

Yet, it should once again be highlighted that Tognini-Bonelli is a linguist and not a stylistician. Not all inductively made observations regarding Oscar Wilde's novel will be equally interesting since the aim is not a comprehensive description of his language use, as this will not say much about the stylistic characteristics and effects. However, stylisticians are turning towards the corpus-driven methodology as well. In her book on Charles Dickens' style, Mahlberg applies corpus-driven methods[12] and phraseology to make innovative observations regarding Dickens' language and characterization (2012). Working inductively, she claims that "the types of patterns retrievable with the help of corpora are often not consciously accessible by our linguistic intuitions" (2012:1). While it is true that new types of patterns may be observed when working bottom-up, it would be erroneous to claim that this approach is entirely objective and dispassionate. A certain degree of filtering of the results is probably inevitable. However, this process of filtering is not certain to block out the innovative potential of the method – one corpus observation may constitute an innovative finding, regardless of the analyst's choice not to focus on a second (and potentially equally interesting) observation. This sieving is inevitable unless the goal is to conduct a (truly) comprehensive linguistic analysis of an author's work, which is certainly not the case here. Comparing Tognini-Bonelli's definition of the corpus-driven approach and Mahlberg's application of it, it becomes apparent that Mahlberg's work is not entirely corpus-driven and inductive. Mahlberg has a pre-determined focus on e.g. characterization and there are several examples of guiding presuppositions that heavily inform her delimitation. For example, she introduces her second chapter like this: "The present study begins with the assumption that clusters illustrate textual patterns that are relevant to the creation of characters in readers' minds" (2012:26). This means that the methodology could either be more flexible than Tognini-Bonelli leads to believe, or that in stylistics, a methodological def-

[12] According to Douglas Bieber (Mahlberg, 2012:i).

inition different from the one found in linguistics should be adopted. In this project, it is assumed that stylistics is a discipline too different from linguistics to apply the same standards for being corpus-driven. This entails that in a corpus-driven stylistic analysis, the ideal of 'going in blind' is not observed as a rigid framework. Some degree of filtering should be accepted, particularly if the conditions for the filtering are clearly outlined beforehand, thus observing Simpson's three Rs (see section 2.1.).

In this thesis, the corpus-driven approach is used for the first part of the analysis and the corpus-based approach is used for the second part of the analysis, as outlined in the introduction. The question that arises is thus: is it methodologically possible to conduct an inductive analysis while also conducting a deductive analysis? From a scientifically rigid and categorical point of view, the answer would be no. However, the actual application of the corpus-driven method in stylistics seems to suggest that there is room for flexibility – that the application of the method is not in fact conditioned by the analyst's ability to go in blind. The goal of the present thesis is, besides analysing Oscar Wilde's only novel, to investigate whether a combination of these methods is valuable, what each method contributes and possibly suggest which method yields the most interesting and compelling results. But before doing so, it must be clearly stated that such a combination *is* methodologically possible because of the partial flexibility of the corpus-driven approach in stylistics. Having established this, I will now move on to introducing the corpora used in this thesis along with the practical and methodological issues they give rise to.

3.2. Corpus design

In this thesis, four corpora are used: *The Picture of Dorian Gray* by Oscar Wilde (henceforth abbreviated to *PDG* and referred to as my target text rather than a corpus); a corpus of Oscar Wilde's collected works (henceforth *Wilde*); a Gothic corpus, which consists of *generically*, but not periodically, similar texts (henceforth *GothLit*); and finally a corpus of texts that are

contemporary with Wilde's but not generically similar (henceforth *ContempLit*[13]). They are all specialized corpora, as all of them are designed to represent special types of language varieties (McEnery, Xiao & Tono, 2006:15). The texts in *GothLit* and *ContempLit* can all be viewed as canonized. No official measure of the texts' canonization has been used – rather the selection of texts has been informed by Fischer-Starcke's corpus design (2010:27-30) and surveys of celebrated literature such as *Great Writers Inspire* (Thompson, 2020), provided by the University of Oxford. I have also sought information and inspiration on Penguin Books' website, under their 'classics' category ("Read - Penguin Books USA – Classics", 2020).

The design of the corpora is one of the most important methodological aspects in corpus stylistics. This has to do with the central role of the corpora used, and Tognini-Bonelli emphasises this concern:

> When it comes to a corpus-driven approach, the issue of the representativeness of the corpus can be seen in its true importance; since the information provided by the corpus is placed centrally and accounted for exhaustively, then there is a risk of error if the corpus turns out to be unrepresentative. At present, until we know a lot more about the effect of selections on the overall picture, it is imperative to be explicit about how a corpus is constructed, and to review the relevance of a particular corpus if there is reason to query the data it provides (2001:89)

However, in corpus *stylistics*, this is true both for corpus-driven and corpus-based analyses. Consequently, the following chapter will be an elucidation of the manner in which the corpora in this thesis are designed and composed. A number of methodological concerns regarding corpus design will be outlined and discussed as well. Given that the goals of the two analyses are different, this must be reflected in the choice of corpora. For the corpus-driven analysis,

[13] Inspired by Fischer-Starcke, 2010.

using a reference corpus is optional as this analysis is *intra-textual*, which means it is "an analysis that examines a particular piece of language data in order to extrapolate information pertaining specifically to that data" (Jeffries & McIntyre, 2010:183). However, in order to evaluate and assess the findings in a greater context, the use of a reference corpus is appropriate and valuable. The second analysis is the contrasting *inter-textual* type. An inter-textual analysis "compares the linguistic features of the target text with those of a control text or collection of texts", that is, a reference corpus (ibid.:183). Here, the use of reference corpora is pivotal. The reason I include three reference corpora in the analysis of the target text has to do with the validity of my results. For me to conclude that some pattern is significantly unique to *PDG*, I must make sure that it is not, in fact, common in all of Wilde's writings (an authorial fingerprint so to speak), that it is not a common trait in Gothic literature (a genre-specific trait) and that it is not a common trait in fin de siècle or Victorian writings (a period-specific trait). A general corpus such as the BNC[14] has not been included because it is an underlying assumption that literary language use is special (see section 2.2.) and thus not directly comparable to language use in general.

3.2.1. Considerations when Building a Corpus

A key aspect of a corpus is its representativeness – if it is not representative, it is an archive, or a random text collection, rather than a corpus (McEnery et al., 2006:13). For a corpus to be representative, it must be *balanced*, that is, it must include an appropriate "range of text categories" (ibid.: 16). In the present case, the corpora can be considered to be balanced[15] as they all consist of novels, and *Wilde* consists of all available texts of the author, regardless of the genre. However, balance is, and remains to this day, "an act of faith rather than a statement of fact, as at present, there is no reliable scientific measure of corpus balance" (ibid.:16). In an

[14] *The British National Corpus* of 100+ million words

[15] The purpose of the analysis being to establish whether *PDG* displays uniquely foregrounded patterns

attempt to ensure greater balance, I decided that no author should be represented more than once in the same corpus, even though this does make the corpora smaller. Moreover, it must be mentioned that the distribution is not symmetrical with regards to the works' year of publication in either corpora. I will return to this issue in section 3.2.2.

Now, a definition of the term *representativeness* is in order, and I will draw on Biber's widely accepted definition: "Representativeness refers to the extent to which a sample includes the full range of variability in a population" (1993:243). This leads to the question of defining the population, and in Biber's words "an assessment of this representativeness thus depends on a prior full definition of the 'population' that the sample is intended to represent, and the techniques used to select the sample from that population" (ibid.:244). He goes on to outline two aspects of defining the target population: "(1) the boundaries of the population — what texts are included and excluded from the population; (2) hierarchical organization within the population — what text categories are included in the population, and what are their definitions" (ibid.:244). For example, the population that *GothLit* is intended to represent is Gothic literature from 1764 to 1911[16] in the form of novels from Great Britain and The United States. The full population would thus be all texts that live up these selection criteria. However, the boundary of the population has to do with accessibility in this case. Only texts that are available from *Project Gutenberg* or *The Oxford Text Archive* (https://ota.bodleian.ox.ac.uk/), are included. This is an issue and constitutes a methodological weakness. However, previous research applying roughly the same methodology as this thesis[17] presents the same weakness. Another aspect to consider when evaluating a corpus' representativeness is the degree of *saturation* (also termed 'closure') (McEnery et al. 2006:15-16). McEnery et al. explain that "the corpus is said to be saturated at the lexical level if each addition of a new segment yields approximately

[16] For an explanation of the choice of dates, see section 3.2.2.

[17] E.g. Fischer-Starcke, 2010:27-30

the same number of lexical items as the previous segment" (ibid.:16). This means that the texts included in one corpus must be approximately of the same length but determining whether that is the case relies on personal judgment. However, McEnery et al. are sceptical of this attitude towards the data:

> In our view, it is problematic, indeed it is circular, to use internal criteria like the distribution of words or grammatical features as the primary parameters for the selection of corpus data. A corpus is typically designed to study linguistic distributions. If the distribution of linguistic features is predetermined when the corpus is designed, there is no point in analyzing such a corpus to discover naturally occurring linguistic feature distributions. The corpus has been skewed by design (2006:14).

The texts included in the corpora are of quite different lengths, as can be seen from the word token count in tables 1 and 2. However, I have chosen not to alter the length of texts by sampling, as this would possibly skew the natural occurrences of certain linguistic features. McEnery et al. argue that a corpus' representativeness is "highly idiosyncratic" (ibid.:9). Therefore, it is my estimate that I have achieved so-called *cumulative representativeness*, though this entails a limitation regarding the ability to generalize the results beyond the corpora of the thesis (ibid.:10).

This thesis engages exclusively with older texts, which is also methodologically important because it means that the corpora are historical or diachronic. Rissanen points to three problems one must consider when working with diachronic corpora: "[the problems] could be called "The philologist's dilemma", "God's truth fallacy", and "The mystery of vanishing reliability"" (2018:16). The first problem has to do with the risk of losing contextual knowledge as all texts in a corpus are unlikely to be read by the corpus stylistician. Correcting this problem by reading all texts included is a next-to-impossible task here. However, an awareness of these

contextual conditions can mend the issue to some degree. The second issue, "God's truth fallacy", has to do with the authority ascribed to the corpora, and Rissanen argues that "an authoritative corpus may easily create the erroneous impression that it gives an accurate reflection of the language it is intended to represent" (ibid.:10). This issue is somewhat averted by properly establishing the validity issues of one's conclusions, outlining the issues of the corpora's balance and representativeness and offering a clear description of the corpora. The last problem concerns the size of historical corpora – due to their finite number, such corpora tend to be rather small which means "it may be difficult to maintain the reliability of the quantitative analysis of less frequent syntactic and lexical variants" (ibid.:11). In order to escape this issue, conservative statistical corrections will be applied to the quantitative results.

The fact that most texts are from *Project Gutenberg* warrants some attention as well. *Project Gutenberg* was started in 1971 by Michael Hart, then a student at the University of Illinois (Hart, 2010). He envisioned the creation of a database of freely accessible e-books, which has grown into a database of over 60,000 free e-books. However, two aspects of *Project Gutenberg*'s work procedures pose minor issues in relation to corpus stylistics. Firstly, they have a principle of minimum regulation, working against "dogmatism, perfectionism, elitism" (Hart, 2004). This means that there is a risk that the volunteer transcribers make mistakes which are then transferred to the corpora. The second issue has to do with accessibility, as Hart states that "we choose etexts we hope extremely large portions of the audience will want and use frequently" (ibid.). They are thus defining the contents based on popularity, which may cause important texts to be omitted. However, this issue has been overcome in the present thesis by

supplementing the use of *Project Gutenberg* with *The Oxford Text Archive*, a less user-friendly database that has a different range[18].

Another aspect to consider is the text types included. The fact that the texts used for the three corpora are all literary poses a problem since "writers tend to innovate, and so literary texts cannot be considered 'sample texts'" (Zyngier, 2008:173). However, this issue can be overcome:

> literary texts may still be studied empirically against some sort of norm. Whether this norm is a collection of other texts from the same period or region, or whether it is the collected works of the same author, we may use the methodology of corpus linguistics for more objective analysis (Zyngier, 2008:173).

In the case of e.g. *GothLit*, the norms are a textual genre, a specific original language, and a specific medium – texts included must be Gothic, they must originally be written in English and they must be novels.

3.2.2. The corpora

In the target text *PDG*, the paratext has been omitted so that it is not considered by the corpus software. The text has been retrieved from *Project Gutenberg*. It consists of 80,744 word tokens.

The *Wilde* corpus consists of the 17 texts listed below, all retrieved from *Project Gutenberg*. It consists of 526,467 word tokens, making it a rather small corpus. Moreover, this is a closed corpus, as Wilde is deceased and it is a unique corpus (Balossi, 2014:49)

[18] More text types besides fiction are included – they define the database as "a repository of full-text literary and linguistic resources" (OTA Home, 2019).

Title	**Text type**	**Tokens**
The Importance of Being Earnest	Play	21,041
An Ideal Husband	Play	32,031
De Profundis	Letter	17,844
Lady Windermere's Fan	Play	21,236
A Woman of no Importance	Play	23,356
Salomé: A Tragedy in One Act	Play	13,417
The Canterville Ghost	Short story	11,817
Poems with the Ballad of Reading Gaol	Poetry	39,148
The Soul of Man under Socialism	Essay	14,613
A House of Pomegranates	Collection of fairy tales	34,641
Reviews	A collection of reviews	164,176
The Duchess of Padua	Play	22,949
Intentions	Collection of essays	60,332
The Happy Prince and Other Stories	Collection of fairy tales	16,640
Vera; Or the Nihilists	Play	19,505
Impressions of America	Travel description	2,435
A Florentine Tragedy – La Sainte Courtisane	Play	11,286
	Total:	**526,467**

Table 1. *Wilde* corpus contents.

The *GothLit* Corpus consists of texts that are traditionally considered to be Gothic, and while most are retrieved from *Project Gutenberg*, two are retrieved from *The Oxford Text Archive* which provides corpora and single texts. The first text marks the beginning of the Gothic tradition and the year 1911 is set approximately where a new attitude in popular Gothic fiction begins (Gothic fiction, 2012). Again, the paratext has been removed (*Project Gutenberg*'s

terms of use, prefaces, translator's notes, introductions by others than the author, content lists, etc.). *GothLit* consists of 1,550,134 word tokens.

Title	**Author**	**Year Published**	**Tokens**
The Castle of Otranto	Horace Walpole	1764	39,864
The Old English Baron	Clara Reeve	1777	54,243
The Mysteries of Udolpho	Ann Ward Radcliffe	1794	145,135
The Monk	Matthew Lewis	1796	137,748
Northanger Abbey	Jane Austen	1817	52,512
Frankenstein; Or, The Modern Prometheus	Mary Wollenstonecraft Shelley	1818	76,398
The Fall of the House of Usher	Edgar Allan Poe	1839	7,217
Jane Eyre	Charlotte Brontë	1847	188,480
Wuthering Heights	Emily Brontë	1847	119,398
A Christmas Carol	Charles Dickens	1848	29,597
The Woman in White	Wilkie Collins	1859	250,467
Carmilla	Joseph Sheridan Le Fanu	1872	28,053
The Strange Case of Dr. Jekyll and Mr. Hyde	Robert Louis Stevenson	1886	25,960
She	H. Rider Haggard	1886	69,275
An Occurrence at Owl Creek Bridge	Ambrose Bierce	1890	3,788
The Yellow Wallpaper	Charlotte Perkins Gilman	1892	6,251
Grim Tales	Edith Nesbit	1892	24,251
The Little Room and Other Stories	Madeleine Yale Wynne	1895	21,779
Dracula	Bram Stoker	1897	162,562
The Turn of the Screw	Henry James	1898	43,429

The Willows	Algernon Blackwood	1908	19,875
The Door in the Wall and Other Stories	H.G. Wells	1911	43,852
		Total:	**1,550,134**

Table 2. *GothLit* corpus contents.

The last corpus is the *ContempLit* which consists of texts from writers that are roughly contemporary with Oscar Wilde (1854-1900), as they range from 1834 to 1910. This estimate is inspired by Fischer-Starcke's practice where she defines the contemporaries of Austen as those authors writing 35 years prior to Austen's birth and 42 years after ditto's death (2010:27-28). In this thesis, the estimate is more restricted (my maximum deviation was 20 years prior to, and 10 years after Wilde's death). The result is a corpus of texts ranging from 1834 to 1902. Some writers appearing in the *GothLit* corpus also appear in the *ContempLit*, but different texts have been chosen. Again, the paratext has been removed. The *ContempLit* corpus consists of 3,244,788 word tokens.

Title	**Author**	**Year Published**	**Tokens**
The Last Days of Pompeii	Edward Bulwer-Lytton	1834	152,594
Mr. Midshipsman Easy	Frederick Marryat	1836	142,367
Sartor Resartus	Thomas Carlyle	1836	79,888
Oliver Twist	Charles Dickens	1837/39	161,508
Windsor Castle	William Harrison Ainsworth	1842	117,100
Vanity Fair	William Makepeace Thackeray	1847	310,381
Agnes Grey	Anne Brontë	1847	69,283
Mary Barton	Elizabeth Gaskell	1848	165,404
Shirley	Charlotte Brontë	1849	32,506
Alton Locke, Tailor and Poet	Charles Kingsley	1850	155,087

The Scarlett Letter	Nathaniel Hawthorne	1850	61,879
Moby Dick	Herman Melville	1851	150,104
The Heir of Redclyffe	Charlotte Younge	1853	238,178
The Warden	Anthony Trollope	1855	73,312
Tom Brown's School Days	Thomas Hughes	1857	112,288
Alice's Adventures in Wonderland	Lewis Carroll	1865	27,330
Little Women	Louise May Alcott	1868/69	190,979
Lorna Doone	R.D. Blackmore	1869	50,480
Middlemarch	George Eliot	1871/72	97,995
Erewhon	Samuel Butler	1872	82,618
The Portrait of a Lady (vol.1)	Henry James	1881	38,048
Adventures of Huckleberry Finn	Mark Twain	1884	86,756
Flatland	Edwin Abbott Abbott	1884	32,712
The Man Who Would Be King	Rudyard Kipling	1888	16,110
Tess of the D'Urbervilles. A pure Woman	Thomas Hardy	1891	153,077
New Grub Street	Georg Gissing	1891	190,896
The Adventures of Sherlock Holmes	Sir Arthur Conan Doyle	1892	105,759
The Time Machine	H.G. Wells	1898	32,808
McTeague	Frank Norris	1899	50,691
To Have and to Hold	Mary Johnston	1899	27,356
Heart of Darkness	Joseph Conrad	1902	39,294
		Total:	**3,244,788**

Table 3. *ContempLit* corpus contents.

Returning to the issue of normal distribution, it should be noted that the balance is somewhat affected by the fact that I have not achieved a symmetrical or normal distribution across time

in either corpora. In *GothLit* the extremities are at 1764 and 1911, warranting a median of 1856. Mine is set at 1859, which is not a notable deviation. However, the distribution is quite random as becomes apparent in the graph below.

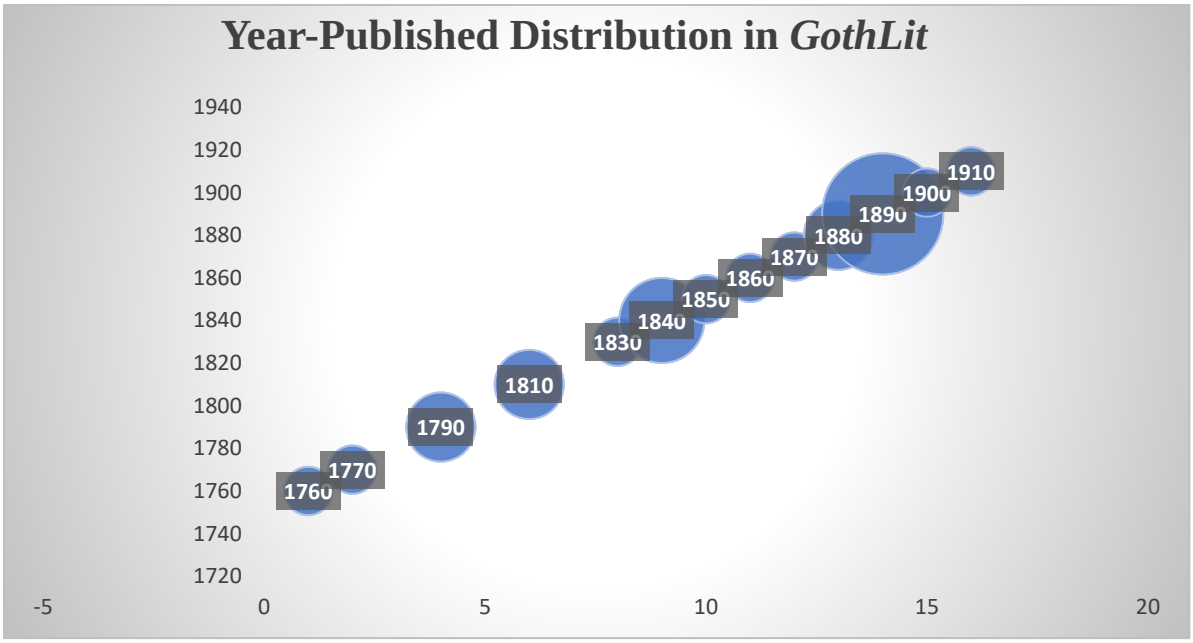


Figure 1. Year-published distribution by decades in *GothLit*.

In *ContempLit*, the extremities are set at 1834 and 1902, warranting a median at 1866. In my case, it is 1861, which is not far from the ideal median. Once again, the graph below shows that the distribution is rather random, albeit less so than in *GothLit*.

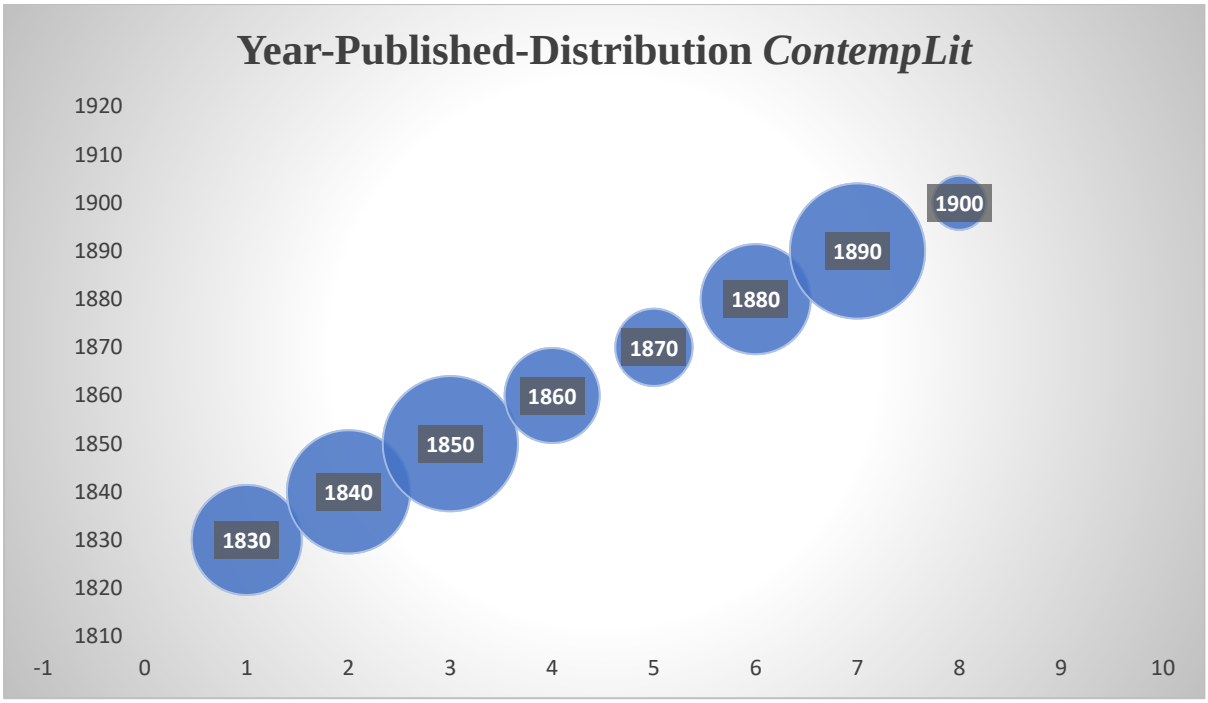


Figure 2. Year-published distribution by decades in *ContempLit*.

Even though this poses a potential issue regarding the balance of the corpora, the problem should be viewed in light of external conditions such as accessibility and canonization. It appears some centuries simply were not very productive of celebrated/canonized literature for some unknown reason, which is why, when working with diachronic corpora, the lack of balance should be accepted to some extent. Having now accounted for the corpora, their design, and potential validity issues, I will move on to the software programmes used in this thesis.

3.3. Software – AntConc, TagAnt, and corpus annotation

In modern corpus linguistics and stylistics, the use of computer software is both pivotal and instrumental. The corpus stylistic analysis is conducted using the concordance software, *AntConc*, provided by Laurence Anthony, professor of Educational Technology and Applied Linguistics at Waseda University (Anthony, 2015). It is a tool that has been specifically designed for the purpose of using corpus approaches as Anthony defines it as a "freeware, multi-platform tool for carrying out corpus linguistic research" (Anthony, 2014). *AntConc* is a concordancer, which is a programme that can search a corpus for different elements such as single

words, lemmas[19], frequent clusters of words (two words or more that appear together, also called n-grams), which words they appear with in their co-text, collocates and more. Boiled down to the essentials, the software can find two things: "the frequency of linguistic units (words, phrases, and lexicogrammatical structures) and their distribution (within an individual text or across a corpus)" (Stubbs, 2014:53). *AntConc* currently has seven different tools or functions: concordance tool, concordance plot tool, file view tool, clusters/n-grams, collocates, word list, and keyword list. In this thesis, different tools will be used for the two analyses. For the corpus-driven analysis, all tools could have been relevant, but I have decided to limit the analysis to the use of the concordance tool, the concordance plot tool, the word list, and the keyword list. The exact functions of these tools will be introduced *ad hoc* in the analysis.

For the corpus-assisted analysis, only the word list tool is used. This has to do with the design of the analysis, which is informed by a working hypothesis. Since predetermined types of words are the target of the analysis, the word list tool is indispensable for locating the 'colour words' which will then serve as data for descriptive statistics[20]. However, the function of this tool should not be underestimated as it is really a *word-frequency list*, sorting all words alphabetically or by frequency (depending on the settings) – a task that cannot be completed without the use of this tool.

In corpus stylistics, corpus annotation is an important tool. Annotation can be done in a number of ways, and in this thesis, the corpora have been annotated at the lexical level, using part-of-speech tagging (POS tagging or morpho-syntactic annotation), which is the most widespread type of annotation (McEnery et al., 2006:34). In this thesis, the corpora have been tagged, using the software programme *TagAnt*, equally provided by Laurence Anthony. That a

[19] *Lemma* is defined like this: "A lemma is a group of all inflectional forms related to one stem that belong to the same word class" (Brezina, 2018:40). When used, they are set in small capitals.

[20] See section 3.4.

corpus is POS tagged means that each word has been assigned a tag denoting its general grammatical word class. This enables a number of more sophisticated searches, such as verbs and adjectives frequently occurring together, or the calculation of *lexical density*, that is the ratio of lexical words (also called content words) to grammatical verbs, which is indicative of a text's level of complexity (the higher the content word ratio, the more complex) (Balossi, 2014:47). *TagAnt* is built on the programme *TreeTagger* developed by Helmut Schmid but has a more intuitive user interface (Anthony, 2019b). Anthony has set up 58 POS tags in *TagAnt* – e.g. VV which denotes a verb in its base form like 'take' or 'enter'[21], or JJ which denotes an adjective like 'green'. Here is an example of a short extract from the untagged (a) and the tagged (b) versions of *PDG*:

a) The studio was filled with the rich odour of roses
b) The_DT studio_NN was_VBD filled_VVN with_IN the_DT rich_JJ odour_NN of_IN roses_NNS

As can be seen, each word (or token) is followed by an underscore and a tag denoting word class. However, the process of tagging is not completely unproblematic.

Two processes that precede the POS tagging must first be accounted for: *sentence segmentation* and *tokenization* (Lu, 2014:40). Sentence segmentation has to do with the way sentence boundaries are demarcated, and as Lu points out, it is a process that is "relatively straightforward, because, in the overwhelming majority of cases, sentence boundaries are marked by a small set of punctuation marks, such as a period, a question mark, an exclamation mark, and a right close quotation mark, among others" (ibid.:40-41). However, some instances of punctuation should not be counted as a boundary by a tagger, such as the period in some abbreviations like 'Mr.'

[21] The full list has been included as appendix A. See Anthony 2019a.

or in acronyms. As is apparent from the image below, *TagAnt* treats 'Mr.' as a noun (specifically a proper noun, NP) without confusing the period for a sentence period, meaning that the issue does not exist with my tagged corpora.

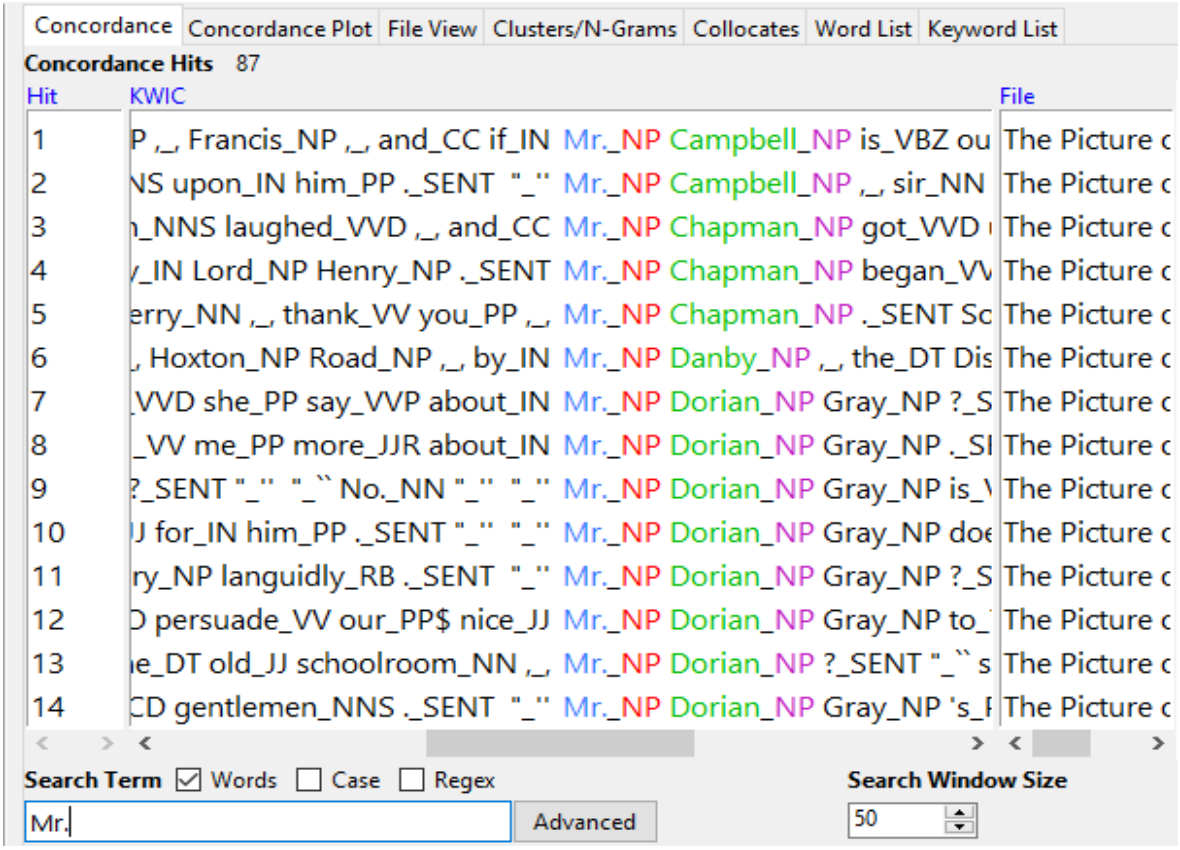


Figure 3. Tagging and sentence boundaries.

The second process to observe is the tokenization. Lu advises that "given that we need to tag each and every token in the text, it is also necessary to determine where each token starts and ends and to separate the sentences in the text into individual tokens accordingly" (ibid.:41). In the vast majority of cases, a token coincides with a word delimited by blank spaces, but there are some problematic instances such as contracted forms ('don't', 'won't' etc.) or hyphenated forms (ibid.). Using the *TagAnt* tagger, contracted forms are divided into two tokens as illustrated below, and so are hyphenated forms (the hyphen being tagged with ':' marking a general joiner.

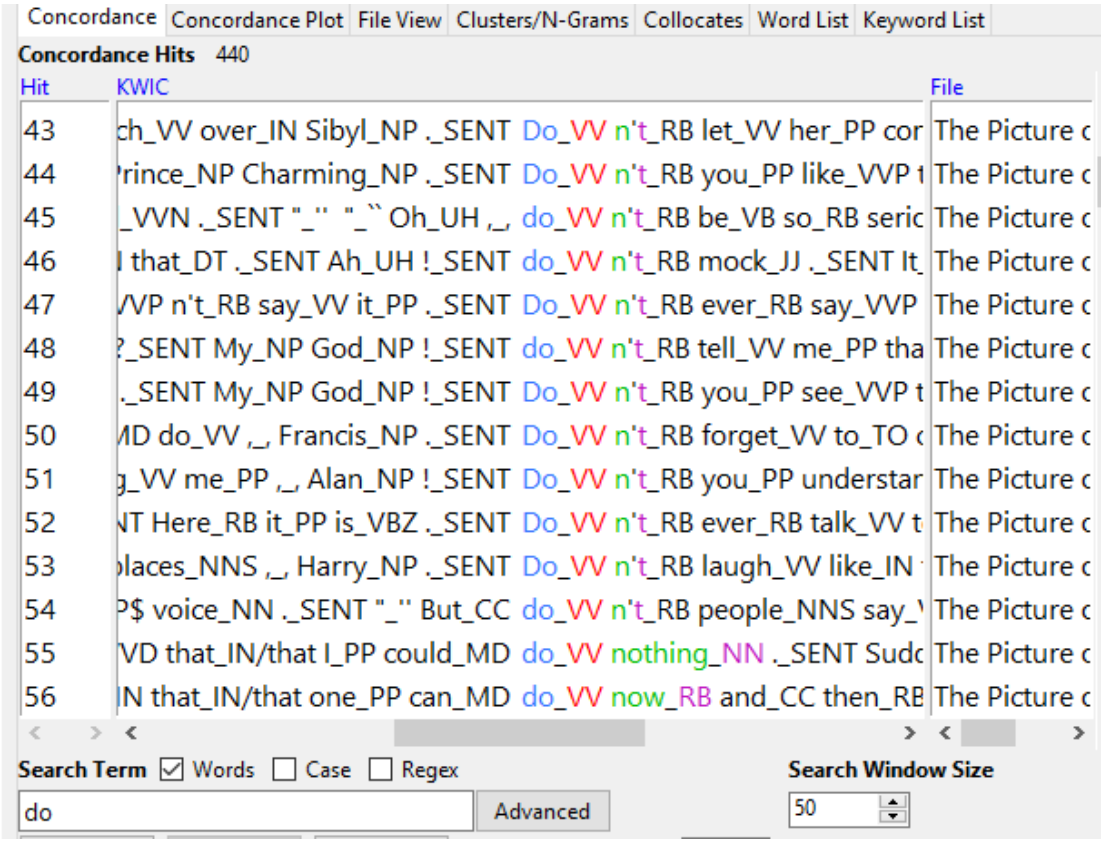


Figure 4. Contracted forms and tagging

Having now accounted for the software used and the important aspects of annotation, I will move on to introducing the particular methods for statistics warranted in this type of corpus stylistic thesis.

3.4. Corpus stylistics statistics and hypothesis testing

One of the great advantages of working with quantitative data is the possibility to perform statistical calculations that allow you to further substantiate claims about observations. Therefore, I will introduce some of the terms and methods that can be used in corpus stylistics. The methods are roughly the same as used in Jensen et al. (2018:42-44). In many cases, and particularly in the corpus-based analysis, *AntConc* is used as a counting tool, tallying the number of occurrences or raw frequencies, of a word, lemma or tag (the linguistic variables). I will use both descriptive and inferential statistics in this thesis, as this is warranted when working with

corpus-driven and corpus-based methodologies: "while descriptive statistics are useful in summarizing a dataset, it is inferential statistics that are typically used to formulate or test a hypothesis" (McEnery et al., 2006:54).

I will begin by introducing the statistics terms related to the corpus-based analysis, as this is slightly more complex because it involves hypothesis testing. The goal of the analysis is to investigate whether or not *PDG* contains more language denoting visual experience of colour – a goal that has been operationalized by singling out words denoting colours (see next section). This goal is turned in to a hypothesis that can be tested – the (tripartite) hypothesis being that:

- There is a difference between the number of colour adjectives in *PDG* and *Wilde* (H_1).
- There is a difference between the number of colour adjectives in *PDG* and *GothLit* (H_2).
- There is a difference between the number of colour adjectives in *PDG* and *ContempLit* (H_3).

Therefore, the null hypothesis is that:

- There are no significant differences between the number of colour adjectives in *PDG* and the three reference corpora (H_0).

When testing a hypothesis, the goal is to find out whether the null hypothesis should be rejected or not (Field, 2009:27). H_0 is the turning point and hypothesis testing cannot determine whether or not the real hypotheses of interest, H_1, H_2, and H_3, should be accepted or rejected – the rejection or non-rejection of H_0 will determine whether there is strong or weak evidence against the other hypotheses (ibid.:55-56). Working with hypothesis testing, it is important to be aware of the risks of so-called Type I and Type II errors. Type I error, also known as a *false positive* "occurs if an investigator rejects a null hypothesis that is actually true in the population" (Banjeree, Chitnis, Jadhav, Bhawalkar & Chaudhury, 2009:131). The opposite is a Type II error, or *false negative*, which "occurs if the investigator fails to reject a null hypothesis that is actually false in the population" (ibid.). Precisely as in Jensen et al., I "conduct a large number of

planned comparisons to check the hypothesis"[22] (Pallant, 2007:206). However, when looking extensively for indications that can serve to reject the null hypothesis, the risk of producing Type I errors is increased (ibid.). One way of avoiding this issue is to apply a Bonferroni correction, which I will explain later. It must be kept in mind that neither hypothesis can be fully accepted nor rejected. As McEnery et al. put it, when making claims in inferential statistics "one can never be 100 per cent sure" (2006:55). Therefore, they continue, "one must state the 'level of statistical significance' at which one will accept a given hypothesis" and in corpus linguistics, this is done by choosing a *p* value. The *p* value (or calculated probability) is a measure of the strength of the evidence against the null hypothesis and "the general practice is that a hypothesis can be accepted only when the level of significance is less than 0.05 (i.e. $p < 0.05$)" (ibid.:55). To clarify, this means that "if the p-value is smaller than (0.05), conventionally we would reject the null hypothesis and say that the agreement is not due to chance" (Brezina, 2018:91).

In order to calculate the significance of an observed difference between two corpora, I use Paul Rayson's (2011) UCREL log-likelihood wizard[23]. Log-likelihood, or LL, is a formula used in maximum likelihood theory (or MLT) (Field, 2009). The principle of MLT is to use the collected data (e.g. two sets of specific word frequencies in two corpora) to create a model in which there is a maximized probability of obtaining the same data. Then this model is compared to the probability of obtaining the same results under the null hypothesis (ibid.:690). To offer an example of the log-likelihood calculation I use the word 'horrible' which occurs 48 times in *PDG* (Corpus 1) and 120 times in *GothLit* (Corpus 2). These numbers and the total number of tokens in each corpus are filled in:

[22] Each tested word is a planned comparison

[23] http://ucrel.lancs.ac.uk/llwizard.html

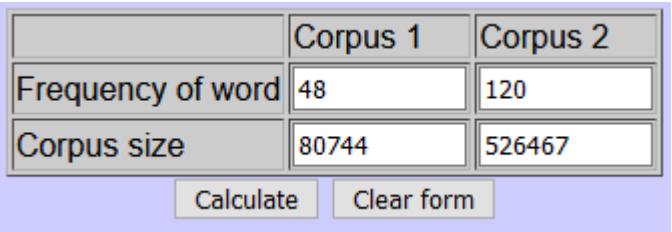


Figure 5. The log-likelihood wizard.

Then I click 'calculate', and get a result bar in which the LL must be above 3.84 – this value follows the *p* value at the alpha level, as Anthony demonstrates in the *AntConc* 2014 help file:

> 95th percentile; 5% level; $p < 0.05$; critical value = 3.84
>
> 99th percentile; 1% level; $p < 0.01$; critical value = 6.63
>
> 99.9th percentile; 0.1% level; $p < 0.001$; critical value = 10.83
>
> 99.99th percentile; 0.01% level; $p < 0.0001$; critical value = 15.13
>
> (Anthony, 2014)

The LL-wizard displays the result like this:

Item	O1	%1	O2	%2	LL	%DIFF	Bayes
Word	48	0.06	120	0.02 +	26.92	160.81	13.60

Figure 6: Log-likelihood wizard result bar.

McEnery and Hardie explain how to interpret Rayson's (2011) result bar:

- O1 and O2 are *observed frequencies*, the numbers you entered
- %1 and %2 are the observed frequencies in normalised (percentage) form
- The + sign indicates that the word is more frequent, on average, in Corpus 1 (a minus sign would indicate it is more frequent in Corpus 2)
- The LL score is the log-likelihood, which tells us whether the result can be treated as significant (McEnery & Hardie, 2011)

In this example, the LL is 26.92 meaning that the word 'horrible' is *significantly* overused in *PDG* compared to *GothLit*.

In the corpus-based analysis, I conduct a number of planned comparisons of colour adjectives in order to test the null hypothesis, H_0. This means that there is a risk of obtaining a Type I error, which occurs "when we think there is a difference between our groups, but there really isn't" (Pallant, 2016:228) since I am deliberately looking for the evidence that may speak against H_0. In order to avoid this issue, a multiple-comparison correction can be applied at the alpha level known as the Bonferroni adjustment (ibid.:230). The adjustment is applied by dividing the alpha level, 0.05, by the number of planned comparisons – 24 words are tested in this thesis. This means that the new critical value must be adjusted to $p < 0.002$ at the adjusted alpha level. There is one degree of freedom and the calculation can be found in appendix B. The *new critical value* is 9.62 at the adjusted alpha level, which means that the log-likelihood must be above 9.62 in the corpus-based analysis in order to be viewed as significant. This adjustment is quite conservative, meaning that the risk of producing Type II errors is somewhat increased, that is, "when we fail to reject a null hypothesis when it is, in fact false" (ibid.:228). Nevertheless, I have judged that when positioning this kind of hypothesis, it is better to risk a false negative than a false positive, thus strengthening the position of any significant observations in the analysis. It should be kept in mind that the Bonferroni adjustment is only actively used in the corpus-based analysis[24].

In addition to the log-likelihood ratio tests, I also include *Bayes Factor* when testing the null hypothesis. It is a measure that is included in Rayson's (2011) wizard (see fig. 6) and it has been argued that it "provides a clearer estimate of the amount of evidence present in the

[24] In the corpus-driven analysis, there are some calculations in which *AntConc* applies a Bonferroni correction automatically.

data" (Jarosz & Wiley, 2014:2). Many researchers in the field of digital humanities have argued that null hypothesis significance testing should be combined or substituted with Bayesian methods because the *p* value can arguably be said to be somewhat arbitrary and based on traditions (Wilson, 2013; Dienes & McLatchie, 2018). Jarosz and Wiley explain that "the Bayes factor is a ratio that contrasts the likelihood of the data fitting under the null hypothesis with the likelihood of fitting under the alternative hypothesis" (2014:3). That is expressed like such:

$$BF_{01} = \frac{\text{likelihood of data given } H_0}{\text{likelihood of data given } H_1}$$

(Jarosz & Wiley, 2014:3)

The Bayes Factor is interpreted in the following way:

> You can interpret the approximate Bayes Factor as degrees of evidence against the null hypothesis as follows:
>
> 0-2: not worth more than a bare mention
>
> 2-6: positive evidence against H_0
>
> 6-10: strong evidence against H_0
>
> > 10: very strong evidence against H_0
>
> For negative scores, the scale is read as "in favour of" instead of "against"

(Rayson, 2011)

In the corpus-driven part of the analysis, *AntConc* is used in a number of ways but one of the main notions that needs clarification is 'keyword'. It is a word (or rather a word type) that occurs significantly more in one corpus compared to a reference corpus. *AntConc* can generate a list of keywords and rank them from most significant to least significant (never insignificant). The above example essentially shows that 'horrible' is a keyword in *PDG*, and the keywords

are all calculated using *AntConc*'s default setting which is log-likelihood. In *AntConc*, the statistical measure can either be set as log-likelihood or chi square test, and I use log-likelihood as "it does not assume that data are normally distributed" (McEnery et al., 2006:55). However, it should be kept in mind, as Brezina warns, that

> The term 'keywords' might be slightly misleading because it suggests that there is a single set of words that characterize a particular corpus. However, (…) the keyword list is a result of multiple decisions in the process starting with the selection of the reference corpus and finishing with the choice of the particular statistic (2018:86)

Nevertheless, the keyword list generated by *AntConc* can still be a valuable point of departure for further analysis, as will become apparent in chapter 4. The methods chapter is thus concluded and having now accounted for the most pivotal aspects of statistical analysis, I will move on to the first part of the analysis, in which I explore the potentials of a corpus-driven approach.

4. Corpus driven analysis

The corpus-driven analysis is characterized by the bottom-up approach in which the text is explored with no prior conception of the possible results. In the first analysis, I start out by conducting large quantifications in *AntConc* that can serve as a steppingstone for qualitative analysis. Then I select the most striking and interesting results and, where possible, analyse them along with their immediate co-text. This process is, as explained earlier, a subjective one in which I, as the analyst, make critical choices – yet they are not based on intuition and *ad hoc* observations as they are all founded in the quantitative data and its scores. It is a process that ensures that the results focused on are also interesting for the literary interpretation and for the analysis of the style in *PDG*. Finally, I summarize the findings in a chapter conclusion.

4.1. Detecting patterns in a keyness analysis of tags

Commencing the, initially, somewhat directionless corpus-driven analysis, I will conduct a comprehensive counting of the tags in all corpora, in order to be able to locate elements that are significantly more frequent (or infrequent) in *PDG* compared to the three reference corpora. To do so, I use Anthony's 58 tags[25] as a starting point, adding a clear direction to the analysis, without narrowing it down too much. As can be seen in appendices C.1, C.2, and C.3, each tag's occurrence in each corpus is tallied – these are the raw frequencies. The next step was to calculate the log-likelihood for each tag in order to locate any statistically significant deviations, and these were the results:

Tag		*Wilde*		*GothLit*		*ContempLit*	
CC	coordinating conjunction	46.35	-	58.11	-	92.82	-
CD	cardinal number	0.00	-	26.55	-	0.25	+
DT	determiner	84.87	-	18.51	-	49.32	-
EX	existential there	16.26	+	72.06	+	54.27	+
FW	foreign word	4.52	-	7.36	+	0.09	+

[25] See appendix A

IN	preposition/subord. Conj	18.01	-	10.26	-	10.17	-
IN/that	complementizer	24.07	+	31.58	+	50.46	+
JJ	adjective	49.18	-	85.72	+	3.52	+
JJR	adjective, comparative	7.95	-	14.75	-	25.28	-
JJS	adjective, superlative	19.84	-	2.62	-	17.12	-
LS	list marker	23.40	-	28.03	-	7.18	-
MD	modal	12.85	+	0.38	-	0.02	+
NN	noun, singular or mass	281.59	-	88.83	-	191.25	-
NNS	noun plural	94.35	-	4.34	+	30.27	-
NP	proper noun, singular	1285.61	-	6171.11	+	88.41	-
NPS	proper noun, plural	13.03	-	3.58	+	0.51	+
PDT	predeterminer	0.65	+	0.24	-	4.12	-
POS	possessive ending	9.80	-	10.06	-	29.26	-
PP	personal pronoun	1176.93	+	26.05	+	636.78	+
PP$	possessive pronoun	25.35	+	126.42	-	1.40	-
RB	adverb	14.06	+	20.10	-	55.96	-
RBR	adverb, comparative	0.31	+	5.05	-	4.20	-
RBS	adverb, superlative	12.09	-	2.56	+	0.57	+
RP	particle	67.91	+	2.45	+	5.06	-
SENT	end punctuation	106.97	+	798.15	+	1496.74	+
SYM	symbol	1331.28	-	402.77	-	74.62	-
TO	*to*	47.74	+	71.45	-	3.55	-
UH	interjection	0.09	+	27.15	+	44.25	+
VB	verb *be*, base form	0.12	-	1.72	+	0.01	-
VBD	verb *be*, past	535.57	+	0.83	+	1.43	+
VBG	verb *be*, gerund/participle	1.01	+	0.38	+	0.16	+
VBN	verb *be*, past participle	41.23	+	10.19	+	3.28	+
VBZ	verb *be*, pres 3rd p. sing.	181.16	-	243.36	+	184.21	+
VBP	verb *be*, pres non-3rd p.	0.00	-	173.07	+	220.07	+
VD	verb *do*, base form	invalid		invalid		invalid	
VDD	verb *do*, past	invalid		invalid		invalid	
VDG	verb *do*, gerund/participle	invalid		invalid		invalid	
VDN	verb *do*, past participle	invalid		invalid		invalid	
VDZ	verb *do*, pres 3rd per.sing	invalid		invalid		invalid	
VDP	verb *do*, pres non-3rd p.	invalid		invalid		invalid	

VH	verb *have*, base form	25.47	+	5.50	+	3.20	+
VHD	verb *have*, past	735.64	+	41.13	+	57.97	+
VHG	verb *have*, gerund/participle	8.16	+	0.48	-	0.59	-
VHN	verb *have*, past participle	0.15	+	0.00	-	1.05	-
VHZ	verb *have*, pres 3rd per.sing	62.45	-	25.18	+	18.97	+
VHP	verb *have*, pres non-3rd p.	0.34	+	31.41	+	49.01	+
VV	verb, base form	27.04	+	7.12	-	0.51	+
VVD	verb, past tense	1444.07	+	12.79	-	35.32	+
VVG	verb, gerund/participle	106.74	+	1.85	-	25.27	-
VVN	verb, past participle	49.21	+	20.88	-	1.27	-
VVP	verb, present, non-3rd p.	51.22	+	219.43	+	205.91	+
VVZ	verb, present 3d p.sing.	356.91	-	57.54	+	10.23	+
WDT	wh-determiner	3.97	+	4.56	-	0.80	-
WP	wh-pronoun	4.19	+	96.92	+	66.28	+
WP$	possessive wh-pronoun	0.79	-	5.27	+	9.49	+
WRB	wh-adverb	13.37	+	2.25	+	5.24	+
:	genreal joiner	369.78	-	1853.16	-	912.97	-
$	currency symbol	invalid		invalid		invalid	

Table 4. Log-likelihood for all tags in *PDG* compared to the reference corpora.

All results below 3.84 have been marked yellow and they should not be considered since they are statistically insignificant at the 5 % level. All *patterns* have been marked with a green colour (to be read horizontally, thus comparing results across the corpora). These are the results that are worth further inspection as they indicate patterns that are statistically significant. The un-marked results will not be analysed, for though they may be statistically significant, there is no coherent pattern across the three reference corpora. These results may thus only say something about the difference between *PDG* and a single reference corpus (which may be interesting in its own right but not in the context of this thesis' goal).

One of the most interesting patterns revealed by this counting is the relative underuse of nouns (the NN-tag) in *PDG* compared to all three reference corpora. However, the result is not easy to interpret as there may be several reasons why this is the case, which demands larger

and more targeted corpus linguistic analyses such as the Longman Corpus analysis by Biber, Johansson, Leech, Conrad and Finegran (1999). Referring to their analysis, I do suggest that the notably high overuse of personal pronouns (PP) in *PDG* could be caused by the fact that a substantial part of the novel is made up by dialogue. As Biber et al. (1999) conclude, pronouns are used much more frequently in conversations and common-speech than in writing, or as Säily and Siirtola put it, "pronouns (…) are found to be more common in speech than in writing, especially in conversational as opposed to task-oriented speech, and more common in imaginative than in informative writing" (2011:169). There also seems to be a pattern regarding coordinating conjunctions (CC) and prepositions/subordinate conjunctions (IN) and general joiners (:). These word types are significantly underused in *PDG* which may indicate that the text has a low degree of explicit cohesion – a result that fluctuates with the highly significant overuse of sentences (SENT) in *PDG*. The latter indicates that the sentences are rather short in *PDG* compared to the reference corpora. However, both these results could also potentially be explained by Wilde's extensive use of dialogue in *PDG*.

Some of the word classes that give rise to the question of style are adjectives and adverbs, as they may indicate a degree of descriptive language. The results above show that adjectives are more frequent in *PDG* than in *GothLit* and *ContempLit*, though the result for the latter is statistically insignificant. Since the results are rather inconsistent, I have collapsed the counts for all adjectives and adverbs (JJ, JJR, JJS, RB, RBR and RBS) and calculated new log-likelihoods on this basis. This is the result:

	Wilde		*Goth-Lit*		*ContempLit*	
Log-Likelihood	8.18	-	10.08	+	18.89	-

Table 5. Log-Likelihood scores for descriptive language markers.

The numbers show that adverbs and adjectives are more frequent in *PDG* compared to *GothLit* but not compared to the other reference corpora. Consequently, it is not possible to make any generalizations regarding the degree of descriptive language in *PDG* based on this type of counting. Seeing as it is problematic to interpret the results in any general terms, I will move away from the tags and on to another approach that may generate a better starting point for a qualitative analysis: analysing the keywords.

4.2. Keyword analysis

Though the initial survey of tags and word classes did not yield much insight into *PDG*, there are other ways of approaching the corpus-driven analysis. Mahlberg (2012) and Fischer-Starcke (2010) both use keyness in their analyses, a measure that can determine how *key* a word is compared to a reference corpus. For example, the names of protagonists are often keywords (their keyness is high) since they are often more or less unique to one novel. Thus, I have conducted a keyword search in *AntConc*, which provided the results in appendix C. The un-tagged corpora have been used for this purpose. The search was conducted using a statistic threshold where the result is significant at $p < 0.05$ and a Bonferroni correction is automatically added by *AntConc*. This is a short extract from the appendix C, illustrating its structure.

Keyness *Wilde*					
Rank	Freq.		Keyness (LL4)	Effect (DICE)	Keyword
1	414	+	1637.68	0.0102	dorian
2	236	+	750.53	0.0058	henry
3	833	+	738.78	0.02	had
4	199	+	731.86	0.0049	gray
5	178	+	667.43	0.0044	harry

Table 6. Excerpt from the initial keyword analysis.

The three tables represent the keywords of *PDG* compared to each reference corpus. They are ranked in descending order beginning with the highest keyness and their frequency is listed in the second column. In the third column, the keyness score (calculated by log-likelihood) is listed and in the fourth column the effect size is listed. The effect size is based on the Dice

coefficient by default - a statistic used to determine the similarity of two samples (Gries, 2009:200). Keyness is a measure that tests for significance, that is, whether or not there is evidence to support that an effect exists. Yet it can say nothing of the *size* of this effect (Baron et al., 2006:53). This is why the effect size score can be used together with the keyness score to assess the overall importance of the word. According to Cohen, an effect size score of "0.2 can be considered a 'small' effect size, 0.5 represents a 'medium' effect size and 0.8 a 'large' effect size" (Walker, 2008). As can be seen in appendix C, the effect sizes are all well below 0.2 which just means that none of the results are detectable to the naked eye (ibid.). They must therefore be considered in relation to the given spread which is between 0.0376 (highest) and 0.0001 (lowest).

As is often the case, the words with the highest keyness are the names of the characters and words that are specific to the theme of the novel (e.g. 'portrait', 'picture', and 'canvas' which are highly key compared to all reference corpora). As the initial survey showed, pronouns are also key in *PDG* – a tendency that also shows in a keyword list. For the full list, see appendix C.

4.3. Key verbs and their co-text

Looking closer at the results, a number of verbs appear interesting. Below, I have listed them with their rank, frequency, LL, and effect.

Rank	Freq.	Keyness (LL4)		Effect (DICE)	Keyword
18	119	+	195.77	0.0029	seemed
28	42	+	97.23	0.001	murmured
30	95	+	87.27	0.0023	cried

Table 7. Key verbs in *PDG* compared to *Wilde*.

Rank	Freq.	Keyness (LL4)		Effect (DICE)	Keyword
225	119	+	24.51	0.0029	seemed
34	42	+	119.42	0.001	murmured
64	95	+	75.47	0.0023	cried

Table 8. Key verbs in *PDG* compared to *GothLit*.

Rank	Freq.	Keyness (LL4)		Effect (DICE)	Keyword
105	119	+	50.75	0.0029	seemed
37	42	+	109.64	0.001	murmured
109	95	+	48.04	0.0023	cried

Table 9. Key verbs in *PDG* compared to *ContempLit*.

The verbs above are all found significantly more in *PDG* compared to the reference corpora, making them key, and in the following, I will analyse the three most frequent verbs in their co-text ('seemed', 'murmured', and 'cried').

The frequent use of the verb 'seem' in the past tense could indicate that the notions of 'impression' and 'intuition' are important in the novel. However, in order to make such a claim, the key word must be analysed with its co-text. When doing so, it becomes apparent that 'seemed' is a verb that occurs with metaphoric and descriptive language on several occasions and that it regularly collocates with 'to be' ('seemed to be' occurs 10 times), 'to have' (5 times), 'to him' (27 times), and 'to me' (8 times). There are occurrences of 'seemed' where no metaphor co-occurs (e.g. "she seemed not to listen to him" (OW:86)), but it is interesting that Wilde *often* uses the verb to initiate metaphoric language, as it then becomes explicitly subjective – the metaphor (often simile) is linked to a character's personal inner life. In the following, I list a number of examples of Dorian Gray's experience of the world (as rendered by the narrator): "and horror seemed once more to lay its hand upon his heart" (OW:206), "thin blue petals of flame they seemed, rimmed with white fire" (OW:90), "birds that were singing in the dew-drenched garden seemed to be telling the flowers about her" (OW:93), "the fresh morning air seemed to drive away all his sombre passions" (OW:93) and "the trees seemed to sweep past

him in spectral procession" (OW:215). These examples indicate that Dorian Gray experiences the world in a highly impressionist manner, but they are also, with the exception of the second example, instances of *personification*. 'Horror', 'birds', 'flowers', 'morning air', and 'trees' are all personified and animated.

Pursuing this observation, I found more examples of personification with the verb 'seemed' by comparing the results in the concordance window with the 'file view' window, where the text appears in full length. These are the examples I found:

(a) "The honey-sweet and honey-coloured blossoms of a laburnum, whose tremulous branches seemed hardly able to bear the burden of a beauty so flamelike as theirs" (OW:1)

(b) "The heavy scent of the roses seemed to brood over everything" (OW:24)

(c) "The woodwork creaked and seemed to cry out as if in pain" (OW:163)

(d) "His cool, white, flowerlike hands, even, had a curious charm. They moved, as he spoke, like music, and seemed to have a language of their own" (OW:21)

(e) "And why was the red stain larger than it had been? It seemed to have crept like a horrible disease over the wrinkled fingers" (OW:229)

(f) "Time seemed to him to be crawling with feet of lead" (OW:171)

(g) "The wood seemed to him to have become suddenly alive with faces" (OW:209)

In the first example (a), the personification revolves around the branches' ability to bear a burden "of a beauty so flame-like" – that is, an emotional and aesthetic burden, not a physical one[26]. In the second example (b), the verb 'brood' suggests a chicken-like behaviour which points to an animation or a personification of 'the heavy scent'. In example (c), the personification of the woodwork is unquestionable, whereas it is opaquer in (d). The predicator 'flowerlike' introduces one simile quickly followed by a second one ("they moved like music"), but

[26] Had the burden in question been merely physical, then this would not have been an instance of personification.

the surprising finale in which they are attributed a language of their own functions like a personification – the hands become creatures in their own right, detached from the man and are as such defamiliarized, an effect I will return to later. In (e), the 'blood stain' is animated as a disease, making for a weaker personification given that viruses and bacteria are probably not viewed as sentient and animated beings by most people, though it must be recognized as a living entity (even in Wilde's time). Example (f) is perhaps slightly cliché (TIME MOVES LIKE A SLOW PERSON), but the precise transformation of what we usually call 'clock hands' into 'feet of lead' adds a new and interesting defamiliarization, turning the original ICM up-side down. Personifications may be cliché but when they are not, they work to create a strong defamiliarization of an entity as Wilde demonstrates. The final example, (g), is not a straightforward personification as it is unclear what the vehicle is, whereas the tenor, 'the wood' (here meaning 'small forest'), is quite tangible. But the presence of multiple faces adds a dimension of animation that works to defamiliarize an otherwise simple concept.

Though the verb 'seemed' constrains the personification (or animation) to an imagined one, or at least a subjective experience, the reader has to go through and understand the same process as the character when reading. The metaphoricity is thus working doubly to add to the poetic quality and to construe inner life characterization at the same time. Using the verb 'seem' marks the subject and his subjectivity[27], and even more so with the addition of 'to him' or 'to me'. The personifications remain personal experiences of the character – whether they are poetic-aesthetic impressions of the world or mental disturbances remains unclear, which may well be an intentional effect, given the supernatural element in the novel. This is an interesting observation but also one that would perhaps be unlikely to arrive at, had it not been for the corpus-driven, quantitative data analysis.

[27] It should be noted that this is sometimes made ambiguous by the self-conscious narrator, whose subjectivity tends to merge with that of the characters.

The list of interestingly frequent verbs also counts ‘murmured’ and ‘cried’ – two modes of communication or expression that seem almost opposite – one toned down or almost muted, the other highly expressive and normally associated with a high volume. In order to make sense of this observation, I used the concordance plot tool, and this is what I found:

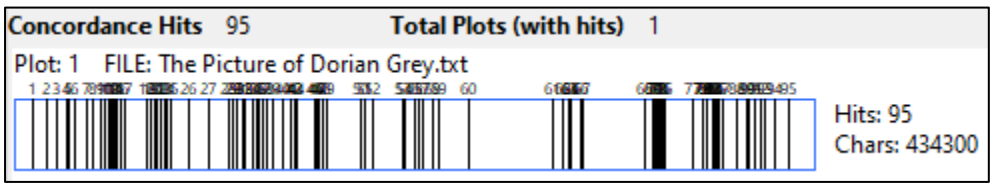


Figure 7. Concordance Plot for ‘cried’.

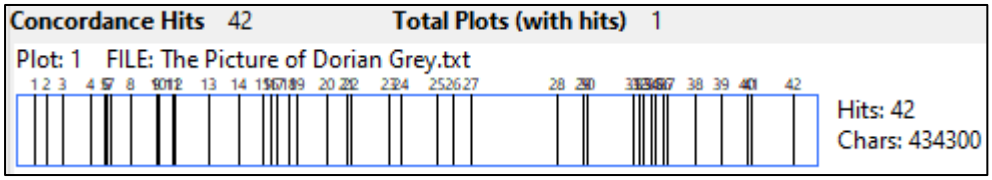


Figure 8. Concordance Plot for ‘murmured’.

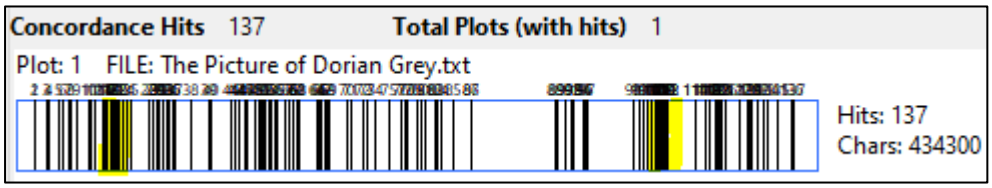


Figure 9. Concordance Plot for ‘cried/murmured’.

Each line indicates an occurrence of the search word and helps visualize the use of these words. What is most interesting is that the words appear to have a similar concordance plot which means they appear together. This was unexpected given the semantic difference between the words. I have examined two places in the text further[28]. Upon closer inspection, the words first appear noticeably together towards the end of chapter 2, when the painter Basil Hallward reveals his painting of Dorian to the latter. Reading through the passage (OW:25-27), it is striking that, at first, the ‘murmured’ is reserved for Dorian Gray and the ‘cried’ is used by Basil Hallward and Lord Henry, reflecting Dorian’s enchanted state of mind when he sees his portrait and contrasting it with the excitement of Lord Henry and Basil Hallward. Dorian starts using

[28] The yellow, marked areas in figure 9.

'cried' when Basil threatens to destroy the painting, and the verb is used by him exclusively until the end of the chapter. The change from 'murmured' to 'cried' thus marks an important change in Dorian's state of mind but also in his character – from being a young, somewhat timid boy he becomes more expressive and 'loud', marking the beginning of his hedonistic lifestyle. The second interesting co-occurrence of the verbs that I will address is in chapter 15, when Dorian Gray and Lord Henry are at a dinner party with the old Lady Narborough. The dialogue here appears very spirited which may be partially due to the verb choice 'cried'. Here are some examples of the use of the verbs in the co-text:

> "I believe he is in love," **cried** Lady Narborough, "and that he is afraid to tell me for fear I should be jealous. He is quite right. I certainly should."
>
> "Dear Lady Narborough," **murmured** Dorian, smiling, "I have not been in love for a whole week; not, in fact, since Madame de Ferrol left town."
>
> "How you men can fall in love with that woman!" **exclaimed** the old lady. "I really cannot understand it" (OW:182. boldface added).

In this example, Dorian's reply appears jaunty and playful but also idiosyncratic compared to Lady Narborough's 'cried' and 'exclaimed', setting the two characters apart from each other. The shifts between the vociferous and boisterous lines and the calm, subdued replies add an effect of shifting volume. These shifts add energy to the dialogue because the reader can sense the atmosphere and has their senses activated – quite literally by having the hearing sense stimulated by the words on the page. Yet, that is not the only effect that these verbs contribute. Looking at more examples another effect appears: (a) "" How can you, Harry!" cried Dorian." (OW:183); (b) ""Isn't he incorrigible?" cried Dorian, leaning forward in his chair" (OW:182); (c) ""*Fin de siècle*," murmured Lord Henry" (OW:182); (d) ""Ah! what a cynic you are!" cried the old lady, pushing back her chair and nodding to Lady Ruxton" (OW:185). What these shifts in volume also add is *dramatic effect*, which is not surprising given that Wilde's authorship

includes many stage dramas. The reader can thus process the dialogue in more detail, since volume specifications are added, often also accompanied by the character's movements and interactions such as in (b) and (d). In conclusion, the frequent use of the verbs 'murmured' and 'cried' suggests the inclusion of a sense dimension, that of *hearing* – thus widening the reading experience. It also adds dramatic effect – both because the volume shifts a lot in the treated excerpts, but also because the volume specifications guide and inform the reading process. These are interesting results that point towards a dramatic style, substantiated by evidence in the text that the top-down approach has yielded. Having analysed the most interesting verbs, I will now move away from them and focus on another class of words that are significantly overused in *PDG* – adjectives.

4.4. Key adjectives and their co-text

A number of the key words from the initial survey are adjectives, and below I have listed some of the most interesting results with a high ranking and high LL-score:

Rank	Freq.	Keyness (LL4)		Effect (DICE)	Keyword
42	48	+	65.18	0.0012	horrible
50	36	+	52.55	0.0009	mad
77	32	+	38.23	0.0008	hideous
88	43	+	32.81	0.0011	dreadful
147	25	+	21.89	0.0006	monstrous

Table 10. Key Adverbs and Adjectives in *PDG* compared to *Wilde.*

Rank	Freq.	Keyness (LL4)		Effect (DICE)	Keyword
42	48	+	99.73	0.0012	horrible
57	32	+	83.5	0.0008	hideous
72	25	+	69.51	0.0006	monstrous
88	43	+	58.1	0.0011	dreadful
101	36	+	48.54	0.0009	mad

Table 11. Key Adverbs and Adjectives in *PDG* compared to *GothLit*.

Rank	Freq.	Keyness (LL4)		Effect (DICE)	Keyword
22	48	+	167.66	0.0012	horrible
32	32	+	118.91	0.0008	hideous
39	43	+	98.49	0.0011	dreadful
62	25	+	77.66	0.0006	monstrous
86	36	+	59.52	0.0009	mad

Table 12. Key Adverbs and Adjectives in *PDG* compared to *ContempLit*.

Since these words are key, they are important to the style of the novel. They are all adjectives denoting the ugly and visually unpleasant (with the exception of 'mad'), which is a theme in the novel, and therefore their presence is expected. However, out of all 184 occurrences of 'hideous', 'mad', 'horrible', 'dreadful', and 'monstrous', only eight refer directly to the portrait: (1) "Yes, that would serve to wrap the dreadful thing in" (OW:121); (2) "No one would ever look upon the horrible thing" (OW:125); (3) "He saw in the dim light the hideous face on the canvas grinning at him" (OW:159); (4) "the hideous lines that seared the wrinkling forehead or crawled around the heavy sensual mouth" (OW:130); (5) "For weeks he would not go there, would forget the hideous painted thing" (OW:142); (6) "Yes, he would be good, and the hideous thing that he had hidden away would no longer be a terror to him" (OW:228); and finally (7) "Was it to become a monstrous and loathsome thing, to be hidden away in a locked room" (OW:107). The use of adjectives denoting the ugly is thus not mostly due to the portrait's role in the novel, which is a surprising observation. To detect what the adjectives then do refer to, the concordance tool can be used once again as it displays the three words that follow each of the adjectives (in my settings). Thus, it aids the visualization of the words that co-occur with the pejorative predicators.

A survey of these adjectives shows that they co-occur with vastly different words with no apparent pattern emerging. There is, however, one lemma that several of the keywords co-occur (or collocate) with, which is THING/S. The lemma THING/S collocates with 'horrible' five times, 'hideous' three times, and 'dreadful' eight times. Given that a main motif of the novel

is the visualization of moral decay and depravation, it is not surprising that these negative words describing unpleasant visions are frequent – it would be expected given the portrait's pivotal role in the story. However, as was just observed, this is not in fact the primary use of the adjectives and THING/S only refers to that portrait three times (see examples 1, 2, and 3 above). Upon closer inspection, a surprising function of the adjectives appears. The pejorative adjectives are often used to exaggerate and to add a certain camp and dramatic attitude – particularly when the adjective is used to predicate an entity that is rarely thought of as 'horrible', 'monstrous', etc. There are many examples of this, but these are among the most striking: (a) "horrible fascination" (OW:148), (b) "horrible laughter" (OW:190), (c) "horrible sympathy" (OW:107), (d) "monstrous as orchids" (OW:128), (e) "monstrous butterflies" (OW:68), (f) "monstrous dahlia" (OW:81), (g) "monstrous lutes" (OW:136), (h) "hideous greeting" (OW:194), (i) "hideous Jew" (OW:48), (j) "dreadful orchestra" (OW:50), (k) "dreadful wax image" (OW:153), (l) "mad adoration" (OW:56), (m) "mad delightful follies" (OW:169), (n) "mad letter" (OW:227), (o) "mad melodramatic words" (OW:72). This surprising use of the adjectives appears to defamiliarize the beautiful, neutral, or even mundane entities that they relate to. This effect of defamiliarization initially causes a foregrounding of the atypical noun phrases which is due to the deviation from a norm established *outside* the novel. Normally, these pejorative predicates would be expected to collocate with negatively loaded nouns (to some degree). Yet, since the adjectives are continuously used in this manner throughout the novel, a new norm (or style) is established within the text and the effect of foregrounding wears off as the reading progresses.

While some examples are more unusual than others, it is clear from these examples that many of the nouns that are pre-modified by the adjectives are thought of as inherently positive; 'laughter', 'butterflies', 'orchestra' usually denote the beautiful and the positive. It may also

function to further disclose and reinforce the characters' attitude towards life and the surrounding world – one that is often exploitative, pessimistic, and wildly hedonistic. The juxtaposition of the positive and the negative creates an effect of a distorted and *baroque* aesthetic and an *exaggerated*, *affected* style. It also adds an idiosyncrasy to the language in the novel by simultaneously eroding and confounding the semantic boundaries between the ugly and the beautiful – an important motif in the novel if not a main theme.

4.5. Striking keywords – 'Thing', 'Things', and 'Something'

Returning to the word 'thing', it is striking how often this word and similar words ('things', 'anything', 'something') are used. While 'thing' is only key in *PDG* when compared to *GothLit* and *ContempLit*, it does have a high LL score here.

Rank	Freq.		Keyness (LL4)	Effect (DICE)	Keyword
25	126	+	100.81	0.0031	something
90	69	+	32.41	0.0017	everything
102	128	+	28.84	0.0031	things

Table 13. The lemma THING/S and keywords with the suffix 'thing' in *PDG* compared to *Wilde*.

Rank	Freq.		Keyness (LL4)	Effect (DICE)	Keyword
31	128	+	127.1	0.0031	things
35	120	+	110.84	0.0029	thing
39	69	+	105.89	0.0017	everything
53	126	+	85.2	0.0031	something

Table 14. The lemma THING/S and keywords with the suffix 'thing' in *PDG* compared to *GothLit*.

Rank	Freq.		Keyness (LL4)	Effect (DICE)	Keyword
43	128	+	95.64	0.0031	things
47	126	+	92.14	0.003	something
52	120	+	83.54	0.0029	thing
74	69	+	69.23	0.0017	everything

Table 15. The lemma THING/S and keywords with the suffix 'thing' in *PDG* compared to *ContempLit*.

In the following, I will analyse the words ‘thing’, ‘things’, and ‘something’ and their immediate co-text because of their slight semantic similarity, which could point to a larger pattern of meaning. What is noticeable is the high frequency of these words and particularly the way they are used in the novel. For example, the word ‘thing’ or ‘things’ is often used to denote “an object that one need not, cannot, or does not wish to give a specific name to” (Thing - Meaning of Thing, Lexico, 2020). Similarly, ‘something’ is defined as “a thing that is unspecified or unknown” or it is “used in various expressions indicating that a description or amount being stated is not exact” (Something – Meaning of Something, Lexico, 2020).

In several instances, the meaning of ‘thing’ or ‘things’ is exactly this indeterminable object:

(a) “A mist makes things wonderful” (OW:212)

(b) “The colours faded wearily out of things” (OW:104)

(c) “Of course, it is sudden; all really delightful things are” (OW:75)

(d) “Beautiful sins, like beautiful things, are the privilege of the rich” (OW:79)

(e) “Facts fled before her like frightened forest things” (OW:41)

This choice of words points to a ‘shortness of words’, language’s inability to provide precise and accurate signs to denote this ‘thing’ in question, like in (a). In other instances, a more accurate word could have been used, like in (e) where ‘things’ probably could refer to animals, but the fact that ‘things’ was used instead could also be interpreted as an *indifference* towards the surroundings (and language), a detached sensing of the world in which impression is more important than precise expression and subjective experience is ascribed primacy of place. In (c) and (d), ‘things’ seems to denote a vast, abstract concept of beauty materialized, the sum of all paintings, flowers, symphonies, summer days, vases, butterflies etc. Indeed, this use of ‘things’ is rather frequent in the novel, as ‘things’ collocates with ‘beautiful’ eight times, with ‘delightful’ two times and ‘wonderful’ two times. These collocations point to the theme of

beauty but also subtly indicates Dorian's immoral materialistic and consumerist approach to all beauty, even reducing people (mostly women) to 'things'. At the same time, it may reveal beauty's innately indefinable character and thus the true inability to offer a more precise lexical choice. In any case, the convincing keyness of the word in all three comparisons indicates that there may be a motif concerning 'objects' or in a larger sense, *consumerism*.

Furthermore, from the readerly perspective, the reader must rely on her own experience of beauty, as it is unlikely that there is a well-defined ICM for BEAUTIFUL THINGS, a concept that is invariably ambiguous and dependent on personal experience. Another word that has a high keyness compared to all reference corpora, is 'something' with 126 occurrences. This word is often used to describe an intuition, a subjective evaluation or impression that cannot be crystalized into words:

(a) "There was something about Dorian that charmed everybody" (OW:122)

(b) "There was something about him, Harry, that amused me" (OW:48)

(c) "You had brought me something higher, something of which all art is but a reflection" (OW:86-87)

(d) "There was something in his face that made one trust him at once" (OW:16)

(e) "There was something in his look that had made her feel afraid" (OW:65)

(f) "There was something in his low languid voice that was absolutely fascinating" (OW:21)

In these examples, the uncertainty of interpretation is clear, and the ambiguity leads to a precarious, ambiguous, and ominous atmosphere – a feeling of the Gothic, brought about by the lexical choice and its frequent use. At other times, 'something' is used to introduce similes, where 'something' thus means something along the lines of 'similar characteristics', creating a strong link between *impression* and *observation*: (a) "There is something of a child about

her" (OW:53); (b) "There was something of the fawn in her shy grace and startled eyes" (OW:82); and (c) "There is something of the martyr about her" (OW:111).

These examples show how the use of 'something', like 'thing' and 'things', foregrounds impressionist sensing over linguistic expression, adding an uncertainty, a vast room of interpretation demanding of the reader that they accept the insecure and unreliable atmosphere this brings about. A repetitive pattern emerges as 'thing/s' and 'something' are words that are used significantly more frequently in *PDG* than in the reference corpora and that their LL score is high. Therefore, they are important words to examine. Yet, it seems that although there is a certain repetition, this does not foreground any entity per se, as 'thing' and 'something' are not categories that one may classify[29]. Rather they are abstracts and as lexical choices, they serve to foreground the importance of subjectivity, intuition, sense, perception, and impression and their frequent use often leaves the reader with nothing but clues and hints, transferring the incomprehensibility of beauty and horror from Dorian Gray's experience to the reader's experience.

4.6. Conclusion of the corpus-driven analysis

In the initial corpus-driven analysis, many interesting conclusions were enabled by the tallying of tag and word occurrences in *PDG* and the reference corpora. However, the first approach in 4.1. remained a distant reading with limited possibilities for interpretation. While it can be interesting in its own right to observe that there is a significant underuse of nouns and a ditto overuse of personal pronouns, it remains a somewhat isolated observation as the inclusion of co-text analysis is an impossibly time-consuming task here. The analysis of the keywords

[29] In order to avoid any confusion, it must be clarified that the *words* 'thing' and 'things' have been conflated when I write 'thing/s'. When 'thing/s' is rendered in small capitals, it is to be understood as a *lemma*, not an ICM per my previous explanation.

yielded more interesting results in the context of this thesis' goal, albeit the results point in many different directions, as could be expected when including many observations in a corpus-driven analysis. The contrast between 'cried' and 'murmured' adds dramatic effect and makes the dialogue vibrant and energetic. The pejorative adjectives add to the dramatic style and serve to defamiliarize a number of concepts, most interestingly beautiful and benign concepts, thus blurring the line between beauty and ugliness. However, some of the results surprisingly pointed in the same direction. The overuse of the verb 'seem' as well as the vague words 'things' and 'something' all appear to foreground subjectivity, intuition, and ambiguity, making for the Gothic style of the novel and the ominous atmosphere. The pejorative adjectives were also found to co-occur rather often with 'thing/s', linking the observations together – the style is not only one that foregrounds ambiguity and subjectivity – it also does so by using negative and ominous adjectives.

5. Corpus assisted analysis

The corpus-based analysis differs from the corpus-driven analysis because the former is based on testing an *a priori* made assumption. In this part of the analysis, I work from an existing assumption that there is a particular textual foregrounding of a sensuous and aesthetic dimension in *PDG*, one that particularly concerns the textual representation of colour. This assumption is based on the initial reading of the novel as stated in the introduction, and it is essentially the type of impressionistic interpretation that stylistic analysis can substantiate and explain according to Verdonk (2004:126-127). This is, in part, done with a starting point in null hypothesis significance testing (NHST) as introduced in section 3.4. In the present chapter, I test the projected hypotheses (or rather, I test the null hypothesis), and subsequently interpret the results using the theoretical frameworks of cognitive stylistics (see section 2.2.) and foregrounding theory (see section 2.3.). With a foundation in these theories, I analyse the results of the NHST and move on to further qualitative interpretation of the most interesting observations made here. Here, I include the readerly aspect and textual aspects of foregrounding.

5.1. Testing the hypothesis – A quantitative analysis

The corpus-based analysis is carried out in much the same manner as the study *Floral Foregrounding: A corpus-assisted, cognitive stylistic study of the foregrounding of flowers in Mrs Dalloway* by Jensen et al. (2018), an article that is based on hypothesis testing, using log-likelihood tests of lemmas. One of the major differences between the current thesis and the article is that I do not use the term 'lemma' in this part of the analysis since this is not required by my test design. In order to test my hypothesis about *PDG*, it is imperative to locate and delimit the words that have to do with *colour*. Since I am including all colour adjectives (in so far as they

exist in *PDG*), understanding how they may be categorized is fruitful for the subsequent qualitative analysis. Taylor presents anthropologists Berlin and Kay and their work *Basic Color Terms* (1969), in which they position that basic colour terms

(a) are not subsumed under other terms. *Crimson* and *scarlet* are not basic terms in English, since they are varieties of *red*. *Orange* is a basic term, since it is not subordinate to any other colour term

(b) are morphologically simple. Terms like *bluish*, *bluish-green* and *chocolate-coloured*, even *golden*, are excluded

(c) are not collocationally restricted. *Blond*, which describes only hair, is not a basic colour term

(d) are of frequent use. Rare words like *puce*, and technical words like *xanthic*, are excluded (Berlin & Kay, 1969; cited in Taylor, 1989:8)

In the present thesis, I deliberately include adjectives that cannot be said to be basic colour terms since these are stylistically important to consider as well. I also include terms that are not morphologically simple, like 'gilded' and compound adjectives. However, collocationally restricted terms are not included and neither are technical words. The reasons for not including these are that I am not interested in, for instance, the number of blond characters in a corpus or in words that have little to no power to evoke visual sensations in the reader because they are unlikely to be recognized. Since I only include words that exist in *PDG*, there is a heightened risk of falsely rejecting the H_0 (Type I error), because I do not include colour adjectives that *do* occur in the reference corpora but not in *PDG*. For example, there are seven occurrences of the word 'azure' in *Wilde*, three occurrences in *GothLit*, and 14 occurrences in *ContempLit* – and no occurrences of 'burgundy' in either corpus. The reason I do not include these is that a log-

likelihood ratio cannot be measured with no occurrences in the primary corpus, *PDG*. However, this risk is balanced out both by the conservative alpha level and the inclusion of the Bayes factor (see 2.5.).

In a tagged text, words denoting colour are almost all adjectives (marked with the tag 'JJ'), and so, in order to avoid Type I errors, I delimited the search to include adjectives only, thus avoiding issues of polysemy. This also explains why 'words' are preferred over 'lemmas' in this part of the analysis. By issues of polysemy, I am referring to issues like 'orange' meaning the colour orange but also the fruit, 'violet' potentially denoting a flower, 'white' being a character surname in *The Heir of Redclyffe*, or 'green' denoting a grass field. This was done by including '_JJ' after the search word. As *TagAnt* is not perfect at distinguishing between noun forms and adjective forms, a few errors must be expected here, giving rise to potential Type I and Type II errors. The special compound adjectives (e.g. 'rose-coloured') were located by searching for '#coloured'[30] and manually removing wrongly included results such as "the_DT coloured_JJ marbles".

I listed all occurrences of adjectives denoting colour or relating to colour and these words form the basis for this corpus-based analysis. I also included three words that do not specifically denote colours but are more loosely *related* to colour terminology ('iridescent', 'bright' and 'pale'), as they have to do with nuance and colour quality. Moreover, I included compounded colour adjectives (hyphenated) ending in 'coloured' such as 'peach-coloured' or 'moss-coloured'. These were the results, sorted by frequency in *PDG*:

[30] The wildcard # means 'any one word'.

Search term	*PDG*	*Wilde*	*GothLit*	*ContempLit*
white	58	258	481	865
red	41	183	214	464
black	30	147	346	752
blue	28	87	155	361
green	24	128	129	325
yellow	22	97	86	182
scarlet	18	41	20	105
pale	15	82	197	329
purple	15	75	36	85
gilt	14	10	5	36
bright	12	94	187	440
golden	9	69	73	124
pink	9	23	18	96
brown	8	53	75	193
grey	7	64	83	95
crimson	5	39	31	71
silver	5	48	24	59
gilded	4	31	1	13
violet	3	7	2	5
orange	3	2	9	25
vermilion	3	7	0	3
mauve	2	3	0	3
iridescent	2	1	2	0
compound colour adjective	19	29	29	71
Total	**298**	**1320**	**1722**	**3837**

Table 16. Frequency of colour adjectives in *PDG* and the reference corpora.

What becomes apparent is that ‘white’, ‘red’, and ‘black’ are some of the most frequently used colour adjectives in *PDG*, but surprisingly also in all three reference corpora. This is interesting because it then means that these particular basic colour terms must hold a special position in language, or at least in literary writing, though limited to the included works in the corpora. It is also possible to conclude that this tendency is due to the fact that these colours are, generally speaking, more entrenched colours than e.g. ‘mauve’, which will be the case for basic colour terms. This is supported by psychology professor, Heider, whose research “strongly suggests

that focal colours[31] are perceptually and cognitively more salient than non-focal colours" (Taylor, 1989:11). Similarly, 'blue', 'green', and 'yellow' occur more often than less entrenched and non-basic forms such as 'crimson' and 'vermilion', which is an expectable result. Among the surprisingly frequent forms we find such colour adjectives as 'scarlet', 'purple', and 'gilt'. This is an interesting result which may suggest that Taylor is possibly right in positioning that

> for some speakers, terms like *mauve, lavender, lime, burgundy* seem to have basic level status, for others not. Interesting in this connection is Robin Lakoff's (1975) claim that women tend to employ a more precise and more differentiated colour vocabulary than men. If this claim is true, women might in general possess a larger number of colour terms than men (1989:8-9)

Lakoff's theory surely is interesting but whether a difference in the size of basic colour vocabulary is gendered or not seems difficult to conclude when considering these results, as the corpora are composed of texts written by both women and men. However, the result points toward a necessity for re-thinking such terms as 'scarlet', 'purple', and 'gilt' as possibly being *near-basic* colour terms. Overall, the results show a surprising similarity to Berlin and Kay's colour hierarchy, based on colour terminology from 98 different languages. Their radical claim was that the 98 languages all selected their basic colour terms from an inventory of 11 colours, a process that proved not to be random. When a language had only two colour terms (the lowest number observed) these would be 'black' and 'white'. If a third existed, it would be 'red'. The fourth and fifth would be 'yellow' or 'green', the sixth would be 'blue' and the seventh 'brown'. The remaining basic colours (in their classification of focal colours) 'grey', 'orange', 'purple' and 'pink' would not be ordered in a particular manner. They visualized this cross-linguistic tendency like such:

[31] Focal colours are largely the same as 'basic colour' – for a detailed explanation of the slight difference, see Taylor, 1989:9.

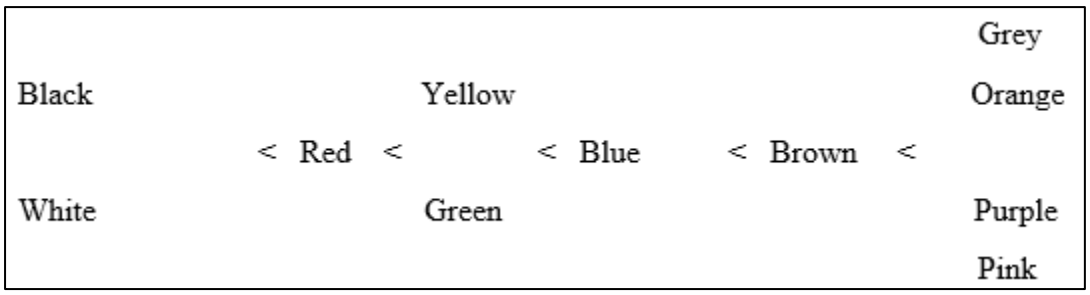


Figure 10. Berlin and Kay, implicational colour hierarchy, adapted from Taylor (1989:10).

Taylor explains how this hierarchy should be interpreted in the following way: "The existence in a language of a category to the right of an arrow implied the existence of all the categories to its left; the reverse implication does not necessarily hold" (1989:10). Although somewhat debated among anthropologists, it seems this hierarchy is valid, considering the present results, when it comes to the 'hierarchy of usage' in English literature. What can then be seen as a more surprising result is the relatively rare use of 'orange', a basic colour term than one would expect to occur more often than it actually does (n=3 in *PDG*, n=2 in *Wilde*, n=9 in *GothLit*, and n=25 in *ContempLit*). This could suggest that the colour orange holds a somewhat different position in literature compared to speech (Berlin and Kay base their research on speech). These observations are interesting because they can say something about the frequency of use of colour adjectives in (some) literature, but in order to test the hypotheses, pairwise comparisons are carried out and the log-likelihood is calculated for each colour adjective (all compared to *PDG*).

5.1.1. Null hypothesis significance test

Having calculated the log-likelihood and Bayes factor scores for the selected colour adjectives, it is now possible to conduct a null hypothesis significance test. To recall, the hypotheses are as follows:

- There is a difference between the number of colour adjectives in *PDG* and *Wilde* (H_1).
- There is a difference between the number of colour adjectives in *PDG* and *GothLit* (H_2).

- There is a difference between the number of colour adjectives in *PDG* and *ContempLit* (H_3).

Therefore, the null hypothesis states that:

- There are no significant differences between the number of colour adjectives in *PDG* and the three reference corpora (H_0).

A total of 24 colour adjectives have been tested for log-likelihood, relative overuse or underuse[32], and Bayes factor. These are the results:

	Wilde			*GothLit*			*ContempLit*		
Search term		LL	Bayes		LL	Bayes		LL	Bayes
white	+	6.38	-6.94	+	29.38	15.07	+	40.55	25.54
red	+	4.43	-8.88	+	51.81	37.51	+	43.22	28.20
black	+	1.91	-11.41	+	6.23	-8.07	+	5.58	-9.44
blue	+	10.15	-3.17	+	27.45	13.14	+	24.67	9.65
green	+	0.78	-12.54	+	24.43	10.13	+	19.64	4.62
yellow	+	2.52	-10.79	+	31.79	17.49	+	33.01	18.00
scarlet	+	11.75	-1.57	+	57.66	43.35	+	36.60	21.58
pale	+	0.38	-12.94	+	1.81	-12.50	+	4.40	-10.61
purple	+	0.83	-12.49	+	32.03	17.73	+	31.18	16.16
gilt	+	26.75	13.43	+	62.76	48.46	+	46.58	31.56
bright	-	0.38	-12.94	+	0.46	-13.84	+	0.10	-14.92
golden	-	0.22	-13.10	+	4.77	-9.54	+	7.17	-7.85
pink	+	4.86	-8.46	+	21.56	7.25	+	10.22	-4.80
brown	-	0.00	-13.31	+	3.07	-11.23	+	1.72	-13.30
grey	-	0.79	-12.53	+	1.31	-12.99	+	5.71	-9.31
crimson	-	0.15	-13.17	+	4.19	-10.11	+	3.79	-11.22
silver	-	0.75	-12.56	+	5.83	-8.47	+	4.99	-10.03
gilded	-	0.11	-13.21	+	19.14	4.84	+	11.83	-3.18
violet	+	1.89	-11.43	+	11.51	-2.80	+	11.97	-3.05
orange	+	5.95	-7.37	+	5.45	-8.85	+	4.47	-10.55
vermilion	+	1.89	-11.43	+	invalid	invalid	+	14.14	-0.88
mauve	+	2.20	-11.12	+	invalid	invalid	+	8.29	-6.73
iridescent	+	4.54	-8.78	+	6.68	-7.62	+	invalid	invalid
compound colour adjective	+	20.50	7.18	+	52.71	38.41	+	52.00	36.98
Total	+	33.44	20.12	+	278.09	263.35	+	265.23	250.22

Table 17. Log-likelihood and Bayes factor of colour adjectives compared to *PDG*.

[32] Overuse in *PDG* is indicated by the + sign and overuse in the reference corpus is indicated by the – sign.

Now that the quantitative analysis has been carried out, it is time to interpret the results. First, the log-likelihood will be assessed for each reference corpus and subsequently the Bayes factor, and any extraordinary results that warrant further explication. Initiating the interpretation of the quantitative data with the *Wilde* corpus (or the H_1 versus the H_0), what is most notable is that there is a substantial difference between the log-likelihood for each individual search term and for the total. This indicates that while *PDG* does display more colour adjectives in total, this result could largely be due to single cases of significant overuses, such as is the case with 'blue', 'scarlet', 'gilt', and compound adjectives, the only words that have an LL score above 9.62. What is then surprising is that 'gilt' has a positive Bayes factor of 13.43, indicating a very strong evidence against H_0[33], whereas the Bayes factor for 'blue' and 'scarlet' are negative (in favour of H_0). This also serves to show the conservative nature of the Bayes factor. The compound colour adjectives also have a positive Bayes factor of 7.18, indicating strong evidence against H_0. As the Bayes factor is rather conservative, it would be prudent to balance the result with the LL score. However, the *individual* scores seem to largely speak against H_1, and the H_0 cannot be rejected on this basis. Only one result, namely 'blue' and its factor -1.57, is neutral/not worth mentioning. What must then be considered is the fact that the *total* yields a log-likelihood score and a Bayes factor that very much speak against H_0, with an LL of 33.44 and a Bayes factor of 20.12. Since the hypothesis is aimed at the *totality of adjectives* in the corpora, this is a very significant log-likelihood result and, in this light, the Bayes factor indicates very strong evidence against H_0. In the case of *Wilde* versus *PDG*, it thus seems that H_1 should be

[33] 0-2: not worth more than a bare mention

2-6: positive evidence against H_0

6-10: strong evidence against H_0

> 10: very strong evidence against H_0

For negative scores, the scale is read as "in favour of" instead of "against"

preferred over H_0. The numbers indicate that there is in fact a difference between the number of colour adjectives in *PDG* and *Wilde*. The most astonishing single result concerns the compound colour adjectives. These seem to be uniquely overused in *PDG* compared to Wilde's remaining authorship, and at this stage it can be said to be a special characteristic of the writing style in the novel.

Moving on to H_2, colour adjectives in *GothLit* compared to *PDG*, a survey shows that 12 out of 22 valid search terms, just more than half, display an LL score above 9.62, making them significant observations of difference. All search terms are relatively overused in *PDG* and two results are invalid because there were no occurrences in *GothLit*. What is remarkable is that the LL scores vary substantially from very low scores like 0.46 and 1.31 to staggering scores like 62.76 and 57.66, indicating large *variations* between the individual adjectives and their degrees of overuse in *PDG*. The most significantly overused colour adjectives compared to *GothLit* are 'gilt', 'scarlet', compound adjectives, 'red', 'purple', and 'yellow', words that can thus be said to be uniquely frequent in *PDG* compared to other Gothic works. Again, only one word has a Bayes factor not worth more than a mention, namely 'violet' with a factor of -2.80, indicating that the majority of the results provide valuable evidence. Since only half the log-likelihood scores are above 9.62, they neither point towards the rejection nor the acceptance of H_0 in favour of H_2. Furthermore, the Bayes factors for the individual words appear to provide equal support for both H_0 and H_2, as 11 of the 22 valid results display evidence in favour of H_0 (negative scores). However, considering once more the *totality* of the adjectives, the Bayes factor is hugely convincing in its dismissal of H_0, with a staggering result of 236.35. Such a number must be taken into consideration when evaluating whether there is more evidence for H_0 or H_2, and the hypothesis formulation considered, I would point towards H_2 as a favoured hypothesis.

Finally, when evaluating the H_3, it is found that all adjectives are relatively overused in *PDG* compared to *ContempLit*. 10 out of the 23 valid search terms have a log-likelihood score below 9.62, showing that there may be an overuse of these words, but it is not a statistically significant difference. This means that the majority of words (the remaining 13 out of 23) are statistically significantly overused in *PDG*, a stronger result compared to the H_2 test. The highest LL scores are compound adjectives with 52.00, 'gilt' with a result of 46.58, 'red' with 43.22, and 'white' with 40.55. These words are thus uniquely characteristic of *PDG* compared the works of Wilde's contemporaries. Moreover, it can now be concluded that 'gilt' and compound adjectives are significantly overused in all three corpora. Looking at the Bayes factors, there are 14 scores that show evidence in favour of the H_0, versus nine scores that favour the acceptance of H_3. The search terms 'bright' (-14.92) ,'brown' (-13.30), and 'crimson' (-11.22) show the most evidence in favour of H_0, whereas compound colour adjectives (36.98), 'gilt' (31.56), and 'red' (28.20) have a Bayes factor strongly favouring H_3. Observing once again the total scores and balancing them with the individual search terms' scores, there is very strong evidence in favour of accepting H_3 and rejecting H_0. This indicates that there is truly a difference in the number of colour adjectives used in *PDG* and the works of his contemporaries.

Considering the results in overall terms, there appears to be evidence to support the rejection of H_0 and the acceptance of H_1, H_2, and H_3 and thus the hypotheses projected have proven to be correct. However, the goal of this project is also to assess *how* this difference works from a readerly perspective using qualitative methods. In the following, I therefore analyse some of the results that warrant further inspection (a more co-text-observing analysis). First, I will treat the compound colour adjectives, as they proved to be very significantly overused in *PDG* compared to all three reference corpora. Likewise, the word 'gilt' seems to be of special importance in the novel as the word had a high log-likelihood score and Bayes factor score in all three comparisons. These special search terms will be analysed using the theoretical

framework established in sections 2.2. and 2.3. regarding cognitive stylistics and foregrounding theory and their co-text will be included to shed light on the effects created by the text.

5.1.2. Hyphenated compound colour adjectives

What may be suggested from these results is also that colour adjectives are foregrounded through frequent use, making colour adjectives an important textual feature when assessing the style of *PDG*. However, processes of foregrounding may unfold in various ways which is why it is now time to dive into the detail. This way, the effects created by any foregrounded elements are also investigated.

One of the most remarkable results concerns hyphenated compound adjectives. The results of the quantitative analysis show that these are used significantly more in *PDG* than in any of the reference corpora, making it possible to claim that this form is characteristic of the novel. From a cognitive stylistic point of view, what is special about this form is that it draws on an ICM from the reader's experience and relates specifically to the ICM's *colour*. However, it is possible that while drawing on these ICMs, not only the colour is invoked, but rather an additional range of features are transferred into the adjective. Hereby I mean that compounds such as 'apricot-coloured', 'cream-coloured', and 'straw-coloured' may well technically and semantically exist on a continuum from a warm white to a more pale yellow and finally a warm orange-like colour as illustrated below.

Figure 11. Visualization of compound adjectives.

However, it is possible that some characteristics associated with apricots, cream, and straw are transferred, lending to the descriptive language an extra dimension of *texture* and consequently

a readerly experience of it[34]. This can be supported by considering the nouns that these adjectives premodify. For example, 'silk', a texture that is soft and smooth is premodified by 'cream-coloured' – 'cream' possessing many similar sensuous characteristics, thus reinforcing the ICM of SILK by drawing on an additional and similar sensory experience. Likewise, Wilde describes a "moss-coloured velvet jerkin" (OW:76), 'moss' possessing many of the same tactile and visual characteristics as velvet. Similarly, a woman's hair is, at one point in the novel, referred to, pejoratively, as "straw-coloured hair" (OW:46), 'straw' bearing a crude, exaggerated resemblance to damaged and coarse hair, thus accentuating the ugliness of said hair by drawing on the ICM for STRAW[35]. I argue that such words as 'straw-coloured' and 'cream-coloured' are very different since the underlying knowledge of texture sensation impact the reader's *experience* of the words. It is possible that they do not just invoke the mental equivalent of said colour, but rather a complete sensory ICM for the premodifier (NACRE, PEACH, etc.). This would also explain some of Wilde's motivation for using these compounds. Consider this list of the 15 compounded colour adjectives present in *PDG*:

a) nacre-coloured
b) honey-coloured
c) peach-coloured
d) olive-coloured
e) cream-coloured
f) straw-coloured
g) rose-coloured
h) pistachio-coloured
i) ochre-coloured
j) moss-coloured
k) fiery-coloured
l) crocus-coloured
m) pearl-coloured
n) apricot-coloured
o) amber-coloured

What is noticeable is that a few of the premodifiers, like 'olive' or 'cream', could have been used without the participle 'coloured'. The reader could infer from context that "the cream silk blinds" means the same thing as "the cream-coloured silk blinds" (OW:90) or that "the nacre

[34] See sub-section 2.2.1.

[35] This is context-specific – the colour can of course also be used in a laudative way.

air" means the same as "the nacre-coloured air" (OW:90). Yet, Wilde preferred the lexical construction using noun + past participle 'coloured' in many instances. This could be due to the *precision* it adds, directing the reader's focus towards the experience of colour – essentially *foregrounding* the colour aspect of the given ICM. It could also be that it works to slow down the reading ever so slightly, creating a calmer rhythm in which the reader's attention can move slowly through the text.

Another noticeable feature in the *internal structure* of these adjectives is that the colour-defining nouns used to qualify 'coloured' are all items from the natural world, all usually considered poetic and aesthetically pleasant, holding some sensuous quality of beauty. In order to closely examine how Wilde used these adjectives, I outlined which nouns are being premodified:

Premodifier	*Noun*
nacre-coloured	air
honey-coloured	moon + blossoms
peach-coloured	divan
olive-coloured	face
cream-coloured	frieze + silk blinds
straw-coloured	hair
rose-coloured	joy + pearls
pistachio-coloured	peridot
ochre-coloured	sawdust
moss-coloured	velvet jerkin
fiery-coloured	life + life
crocus-coloured	robe
pearl-coloured	octagonal stand
apricot-coloured	light
amber-coloured	silk

Table 18. Compound adjectives and the 19 nouns they premodify.

As foreshadowed, what is noticeable from this list is that many of the premodified nouns are entities which are often thought of as beautiful – or rather their ICM is aesthetically loaded with ideas and images that form a pleasant concept. For example, the word 'divan' evokes a particularly ornamented and decorated type of lounge furniture, one that is often designed to

be considered beautiful – an exclusive and luxurious concept. Likewise, 'air', 'moon', 'blossoms', 'pearls', 'peridot', 'robe', 'light', and 'silk' are all words that, in the mind of the reader, evoke ICMs that are already connected to beauty, prettiness, and pleasure. Their ICMs are then connected with other equally beauteous and pleasant ICMs, which will be exemplified below.

As stated in sub-section 2.2.1., "there is now substantial evidence that, as we read, brain areas which would be activated for real physical and emotional responses in the real world are activated for their imaginary equivalents in stories" (Emmott et al., 2014:273). Thus, particular sensory schemata are brought into play when reading a text loaded with e.g. *visual* stimulus. This means that a textual representation of visual schemata of colours activates the reader's real-world experience of those colours. For example, this excerpt shows how ICMs of beautiful entities are juxtaposed for a poetic-aesthetic effect: "From some chimney opposite a thin wreath of smoke was rising. It curled, a violet riband, through the nacre-coloured air" (OW:90). Here, the poetically described "thin wreath of smoke" is defamiliarized in a metaphor where it becomes "a violet riband" and this poetic scenery is then situated in "the nacre-coloured air". Similarly, the beauty is maximized through the juxtaposition of romantic, poetic ICMs in this example: "(…) and through the small leaded panes of the window streamed the apricot-coloured light of a summer day in London" (OW:44). Here, "small leaded panes", "apricot-coloured light", and "a summer day in London" all invoke ICMs that are pleasant or romantic. Stylistically, this creates a reading experience that is complex and yet poetic and aesthetically powerful. The style is maximalist[36] – more is more – creating a supercharged aesthetic reading experience in which the emotional responses to colour are activated in the mind of the reader. What is surprising, perhaps paradoxical, about this style is that the narrative involves death, betrayal, moral and physical decay, and corruption of the soul – and yet *beauty* is a central characteristic of the textual form and function.

[36] Maximalist is here understood as opposed to minimalist.

Moreover, the special use of nature related colour adjectives works to introduce more nature-imagery into the reading experience, one that then becomes more sensual-aesthetic. The hyphenated compound adjectives are less well-entrenched forms as they *do* require some cognitive effort from the reader. It is reasonable to assume that loading into one's current working memory the exact colour of 'nacre' would presumably take slightly longer than loading 'red'. However, the adjectives are not equally non-entrenched[37] – e.g. 'rose-coloured' and 'cream-coloured' are probably somewhat more entrenched forms than 'crocus-coloured' and 'pistachio-coloured'. I base this argument on the relative novelty of the adjectives and their low frequency of use since that which is rare and novel or creative will be more salient *and* less entrenched. In the largest reference corpus *ContempLit*, 'cream-coloured' or 'cream-colored' occurs four times and 'rose-coloured/colored' occurs three times, whereas 'crocus-coloured' and 'pistachio-coloured' occur zero times[38]. Therefore, the latter two are more novel and creative and will also be more cognitively salient than the ones found in other corpora. It can thus be argued that since many of the hyphenated colour adjectives are salient, they also work to foreground colours in the mind of the reader, creating the sensation of a *visually stimulating* text. This preliminary conclusion leads me to the overall question of (colour-)foregrounding mechanisms in the text. Based on three text excerpts with a hyphenated compound adjective, I will analyse these textual processes in more detail.

As stated in section 2.3., foregrounding can be employed in a text in several manners. I will call attention to Emmott and Alexander again, as they point out "sound play, unusual graphic patterning, excessive lexical and pronominal repetition, unusual word choices, highly creative metaphors, parallelism, and breaches of the usual discourse structure" (2014:329) as devices that create effects of foregrounding. Repeating that foregrounding essentially exists in

[37] Non-entrenched in the sense of being opposed to well-entrenched.

[38] The latter only occur in *PDG*

the form of more-of-the-same or deviation from the norm[39], I consider the three examples below:

(a) "Where was the great crocus-coloured robe, on which the gods fought against the giants, that had been worked by brown girls for the pleasure of Athena?" (OW:140)

(b) "He had chasubles, also, of amber-coloured silk, and blue silk and gold brocade, and yellow silk damask and cloth of gold, figured with representations of the Passion and Crucifixion of Christ, and embroidered with lions and peacocks and other emblems" (OW:142)

(c) "Lord Henry Wotton could just catch the gleam of the honey-sweet and honey-coloured blossoms of a laburnum, whose tremulous branches seemed hardly able to bear the burden of a beauty so flamelike as theirs" (OW:1)

In (a), the noun phrase is very complex, having two premodifiers and two postmodifiers. The lexical 'weight' of the modifiers works to *eclipse* the noun from the attention of the reader, making for a slightly difficult reading process, but a heightened sensuous reading-experience, in which *beauty* is foregrounded. In an attempt to visualize the effect that the many pre- and postmodifiers create, I have devised the following flowchart, demonstrating the 'weight' on the head of the noun phrase (robe):

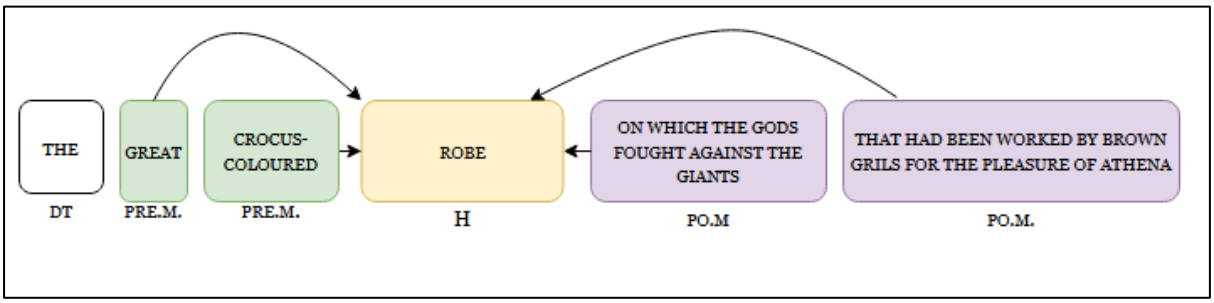


Figure 12. Visualization of the embellishment of a noun phrase.

[39] See section 2.3. – Simpson, 2014:52.

It is also interesting how the beautiful, yet mundane ICM of a ROBE is juxtaposed with an ICM of MYTHOLOGY and WARFARE – a movement between modes that resembles the modernist fragmentary style and stream-of-consciousness. In (b), the maximalism is also connected to the rhythm created by the use of polysyndeton, that is, the successive use of the coordinating conjunction 'and'. In this one sentence, Wilde used six coordinating conjunctions, creating an excessive and embellished style that works to focus the reader's attention on the adornment (here, the postmodifiers) rather than the adorned (the noun 'chasubles') and creates a hypnotic rhythm. This is a structure of parallelism, or even, as Emmott and Alexander term it, "excessive lexical repetition" (2014:329). In (c), the hyphenated adjective 'honey-coloured' is preceded by another hyphenated adjective, 'honey-sweet'. This creates an effect of parallelism, foregrounding not only the word 'honey', but rather foregrounding the poetic effect. This effect is further foregrounded by the use of alliteration ('blossoms' and 'laburnum' and 'bear the burden of a beauty'). The examples (a), (b), and (c) are similar in that they have a hyphenated colour adjective but they are also good demonstrations of the manner in which Wilde's text plays with the reader's focus, foregrounding adornment, embellishment, decoration – aspects central to the Aesthetic movement in which Wilde was a key figure (Gagnier, 1997:28). Thus, Wilde succeeds in *textually* foregrounding beauty, a declared purpose in the 1890s Aesthetic movement.

5.1.3. The word 'Gilt' – Describing materialism

Among the most convincing results of the quantitative analysis of *PDG* was the high overuse of the adjective 'gilt'. It is colour related since it denotes a golden surface, but it is also related to texture and the specific process of decorating something with gold leaf. When analysing the adjective with their co-text, it becomes apparent that it is frequently used in the description of things. Upon conducting a survey of the word 'gilt' in its co-text, I found that it was mainly

used in semi-long or long descriptions of material goods 13 out of 14 times[40]. These descriptions are often the detailed observations of a connoisseur and examples are abundant in the novel. For example, Dorian Gray loses himself in admiration of textiles for a whole year, collecting rare and valuable specimens from the world, among others "Georgian work, with its gilt coins, and Japanese-Foukousas, with their green-toned golds and their marvellously plumaged birds" (OW:141). Among other things, he turns to the collection of precious gems, ecclesiastical vestments, and perfumes. The descriptions guide the reader's attention through a myriad of details, foregrounding by means of textual abundance (more-of-the-same foregrounding) the aesthetic details and decorations rather than the things that are embellished, as in this example, where the colours and textures of clothes worn by the fictional Sir Anthony Sherard are foregrounded, leaving the knight himself in the background: "Here, in gold-embroidered red doublet, jewelled surcoat, and gilt-edged ruff and wristbands, stood Sir Anthony Sherard" (OW:145). The same process of attention diverting descriptions are found in the following example, where the descriptions are so esoteric, niche, and technical that they work to defamiliarize the described (a book that Dorian reads excessively): "It was Gautier's Emaux et Camees, Charpentier's Japanese-paper edition, with the Jacquemart etching. The binding was of citron-green leather, with a design of gilt trellis-work and dotted pomegranates" (OW:168). Surface and decoration are once more foregrounded in relief to the *substance*, the actual book, which slides into the background.

These descriptions are all rendered from Dorian's point of view[41] and they hold a central position in the characterization process. The highly specialist choice of words indicates that he is familiar with finer objects and that he has a particular appreciation for them – the minute,

[40] In one case it is used metaphorically of fuming censers, described "like great gilt flowers" (OW:135).
[41] The narrator is an omniscient heterodiegetic narrator using shifting focalization, while staying primarily with Dorian Gray.

detailed descriptions indicate that he is a true materialist and connoisseur who gains personal gratification from *possessing* (sometimes only in his mind) rare and beautiful things. Dorian becomes a collector after discovering the secret of the portrait, as collecting and possessing objects distracts him from the growing fear for his soul. It is possible to conclude that since the adjective 'gilt' is used in a notably frequent manner, and since the adjective is often connected to the descriptions of material goods, that such descriptions then hold a special position in the novel – a conclusion that is in line with several literary critics.

Gagnier has pointed out *consumerism* as a key characteristic of the late Victorian man and of Wilde himself, who she says was "tempted by the more subjective calculations of pleasure that the new psychologically based economics had introduced" and "sensitive to the revelation of personality through choice and preference" (1997:22). I would argue that the same can be said of Wilde's character Dorian Gray, who becomes a tragic comment on excessive consumption but also a man who fashions himself through possessions – essentially performing his identity. The special foregrounded position that is occupied by objects, things, trinkets, and collectables also reinforces the impression of a man trapped in his own consumerism, trying to escape the reality of his curse – a circular movement between consumption and decay. This following example can partially substantiate this impression: "He was rather sorry he had come, till Lady Narborough, looking at the great ormolu gilt clock that sprawled in gaudy curves on the mauve-draped mantelshelf, exclaimed: "How horrid of Henry Wotton to be so late!" (OW:182). What is remarkable here is that the subject 'Lady Narborough' and the verb 'exclaimed' are separated not only by a clause, but a clause with a lengthy lingering description, with a focus moving from a clock onto the mantelshelf it stands on, before suddenly returning to Lady Narborough. The narrator therefore appears distracted, almost hypnotized by the pres-

ence of these things – an effect that is transferred to the reading experience by the textual foregrounding of the objects. Similarly, this excerpt shows how focus is shifted from the agents of the sentence to the objects in the room:

> Finally his bell sounded, and Victor came in softly with a cup of tea, and a pile of letters, on a small tray of old Sevres china, and drew back the olive-satin curtains, with their shimmering blue lining, that hung in front of the three tall windows (OW:94)

Again, the agent (subject) is separated from the second verb 'drew' by a long object and a prepositional phrase, and then that verb is again followed by a long descriptive object. The descriptions are abundant in specifications of colour, texture, size, and type, foregrounding them and pushing the agent, Victor, to the background, and it is this textual focus on objects and their properties that assist in creating the impression of a mind trapped in materialistic consumerism. Again, Wilde creates the effect of a mind easily distracted by externals, objects heaping up textually and almost overshadowing the characters. The pinnacle of Victorian obsession with *objets rares* is this slightly ironic comment:

> There was a rather heavy bill, for a chased silver Louis Quinze toilet-set, that he had not yet had the courage to send on to his guardians, who were extremely old-fashioned people and did not realize that we live in an age when unnecessary things are our only necessities (OW:94)

Here, the toilet-set is foregrounded – partially because of the very long noun phrase, and partially because of the surprising and *salient* premodifiers. It is of silver and it is a special connoisseur antique, and the words 'Louis Quinze' activate the ICM of FRENCH ROYALTY and RARE COLLECTIBLES, making it a somewhat surprising predicate for a toilet-set, a rather 'mundane' object. The concluding remark regarding unnecessary things strongly (and slightly ironically) underlines the theme of superfluous and mind-numbing consumerism.

The investigation of the important keyword 'gilt' has thus led me to explore the question of materialism's textual presence in the novel. If the foregrounding of beauty and colours create the effect of an aesthetically powerful visual experience, it is possible that the foregrounding of things creates an emotional response in the mind of the reader as well. The exact nature of this response is more difficult to discern. There is possibly a level of the visual since many of these things are described in terms of their visual attributes – colour, size, material, etc. This way, the sensuous experience is amplified, making the text *sensually stimulating*. However, since the emphasis lies on commodities, products, and things you can possess, I am convinced there are grounds for claiming that the text also creates an experience of *desire* for possessing, dragging the reader, as it were, into the spiral of consumption in which Dorian finds himself trapped. These observations are in line with Gagnier's claim that "Dorian finds himself in a society that prefers form to substance" and "the narrator describes market society as society of the spectacle, style or form over substance" (1997:23). Interestingly, this claim can also be connected to the earlier observation about the foregrounding of *beauty* which also points to a form-over-substance preference on Wilde's part. Therefore, I argue that form over substance is stylistically defining of the novel – it is both a *motif* in the text (function) and a textual feature (form).

5.2. Conclusion of the corpus-assisted analysis

Summing up the results of the corpus-assisted analysis, I will first highlight the null hypothesis significance test in which it was found that H_1, H_2 and H_3 should be preferred over H_0. This yields evidence to claim that colours are uniquely foregrounded in *PDG* and play an important textual role. Colours are abundant and they are foregrounded, and this test makes it possible to claim that this is also a unique stylistic trait of the novel. What was also discovered was the stylistic importance of hyphenated colour adjectives. They are found to be used in an interesting manner, one that creates a sensuous reading experience and makes the text visually stimulating.

The compounds are also found to be used in a way that creates a strong poetic effect, introducing nature imagery and poetic defamiliarization in the text. Upon closer inspection of a number of excerpts, it is found that it is essentially the abstract entity *beauty* that is being foregrounded, resulting in a hyper-aesthetic and poetically powerful style.

Pursuing the most interesting results from the NHST, the word 'gilt' was also closely analysed. In this analysis I found that 'gilt' is often associated to objects. This pointed me in the direction of descriptions including the word 'gilt', uncovering that material possessions hold a central position, not only in the narrative but also textually. The foregrounding of objects indicates a periphrastic style and creates the effect of a text that is sensually stimulating but perhaps also stimulates the reader's own desires by transporting the reader's attention towards objects and euphoric descriptions of them. The corpus-assisted analysis has secured very trustworthy results, and also provided a quantitative argument for further analysing certain textual features (compound adjectives and 'gilt'), which has yielded very interesting results. Having concluded the second analysis, it is time to evaluate and discuss not only the results of the analyses, but also the approaches.

6. Discussion and conclusion

The present chapter will be an assessment of the results obtained in the two analyses, comparing them, and evaluating whether they are different or similar. I will also assess whether the results support the same arguments or not. In 6.2. I evaluate the two corpus methods in terms of results yielded in order to assess whether the use of said methods was successful or not, relative to Fischer-Starcke's criteria as outlined in section 2.5. Finally, in 6.3., I compare the results of the corpus-driven and the corpus-assisted analyses in order to find out whether and how the approaches lead to similar or dissimilar results and whether general recommendations can be made on the basis of this thesis, thus completing the third goal of the thesis[42].

6.1. Comparing the results of the corpus-driven and corpus-assisted analyses

First, I compare the results from the corpus-driven analysis with the results of the corpus-assisted analysis in order to assess whether there are similarities, or whether the approaches yield very different results. The former would mean that observations about the text can be arrived at in different ways and that the results obtained are very reliable. The latter would mean that the approaches can complement each other, and that the thesis design is instrumental for the type of results one wishes to obtain. This will become apparent after a complete comparison of the results.

In the corpus-driven analysis it was found that the frequently used verb 'seem' points to an *impressionist* mode of narrative, in which the characters' subjective understanding of the world was important. The subjective and inner experiences were thus found to be central. This is well matched with the finding in the corpus-assisted analysis concluding that there is a *poetic* mode in the novel, since impressionism and poetry share similarities and can be linked together.

[42] From the introduction: "I seek to compare the use of a corpus-driven versus a corpus-assisted corpus stylistic approach".

The agreement between these results is further strengthened by the observation that the verb 'seem' often initiated instances of *personifications*, a defamiliarizing device that can create poetic effect and certainly does so in the novel. Thus, the impressionist was already linked to the poetic in the corpus-driven analysis. The analysis of the uniquely overused hyphenated colour compound adjectives in the corpus-assisted approach showed that the style is *aesthetically powerful* and highly *poeticized*. Colours, nature-imagery, and a supercharged experience of beauty were at the centre of this poetic mode. Thereby, both analyses essentially pointed to the aesthetic-poetic quality of the novel, indicating its importance, both for the style of the novel, but also for the reading experience.

Moreover, it was observed, in the corpus-driven analysis, that the co-occurrence of the verbs 'murmur' and 'cry' added dramatic effect and an effect of volume. This is interesting because it points to the stimulation of the senses – an effect in the reading experience that is also observed in the corpus-assisted analysis. There, it was observed that the text creates effects of *visual stimulation*, and in much the same manner the text creates *auditory stimulation* by means of specific verbs. Furthermore, the corpus-assisted analysis rendered probable that the text was made *sensually stimulating* through its extensive descriptions of *objects*, involving all aspects of the sensory apparatus. This points to the uniquely sensuous dimension of the text, a textual characteristic that was initially suspected and which can now be confidently substantiated by the results of the analyses. This allows for the conclusion that *PDG* offers a rich sensuous reading experience in which the style is closely linked to all dimensions of sensing. This conclusion was also substantiated by the reliable NHST and ensuing analysis, in which it was found that there is indeed a statistically significant difference between the way colour adjectives are used in *PDG* and three reference corpora, and that the frequent use of colour adjectives is indeed characteristic of the novel.

The corpus-driven analysis also led to the discovery of the words ‘something’, ‘thing’, and ‘things’ as being important keywords and the subsequent discovery of an embedded motif of consumerism. In much the same line, though through a different approach, the corpus-assisted analysis led to the discovery of a foregrounding of objects and possession as a motif and a textual characteristic of *PDG*. These findings complement each other and reinforce the strength with which one can say the text creates a motif of *consumerism* and *decadence*, and that Wilde foregrounded dead objects.

The analysis of key adjectives also indicated a defamiliarization of certain noun phrases, though not with a particularly poetic effect as the result. Rather, a camp, affected, baroque, and excessive style was detected. This is in line with the corpus-assisted analysis of the word ‘gilt’, which made evident the foregrounding of objects, an aspect that makes for a maximalist, excessive, and embellished style. Though the focus is different in these two analyses, it is interesting that the results are somewhat similar, and it indicates that these are not random observations but in fact quantifiable characteristics of the novel’s style. Therefore, it can also be concluded with certainty that the style of the novel is indeed overwrought, baroque, and flamboyant.

It is surprising to find that though the results of the two approaches are different, they do point in some of the same directions, allowing for more confident statements concerning the style of the novel. It should be noted that the comparison above was not a subjective one since I purposefully looked for possible interesting overlaps and discrepancies, focusing on these results rather than the smaller discoveries that both analyses also generated. However, some overlap between the results was also to be expected since both analyses concern the same text. Nevertheless, the degree to which the results complement each other is a positive outcome, indicating that the combination of methods can be fruitful if one wishes to engage closely with a text and its style.

6.2. Fischer-Starcke's criteria – Evaluating the results

In 2.5, I introduced Fischer-Starcke's criteria for the successful merger between corpus linguistic approaches and stylistic practice. Whether or not these criteria have been met may now be assessed. The first criterion was that "literary insights or, more generally, new and additional information on the data are gained that could not or have not been generated without electronic analyses" (Fischer-Starcke, 2010:20). The second criterion was that "already known information on the data or previous interpretations of it can be supported or refuted by way of electronically generated data" (ibid.:20). The second criterion is evaluated by comparing this thesis' conclusions to a number of literary scholar's conclusions. The results yielded in both the corpus-driven and corpus-assisted analyses will be evaluated in light of these criteria.

In chapter 4, the corpus-driven analysis investigated key verbs and adjectives in their co-text. That the novel's style is *dramatic*, *exaggerated*, and *affected* is a result that one could have arrived at without the corpus method. What is new is rather the concrete evidence, the details and the knowledge concerning *how* this effect is created, and how it affects the reading experience. Yet, it is difficult to say whether these textually founded results (e.g. the specific use of verbs creating dramatic effect and effect of volume, and the use of unusual premodifications) were primarily arrived at due to the corpus method or the cognitive approach. I lean towards the latter, as the cognitive theory constitutes the interpretative framework for the corpus findings. However, the specifics of the textual level were *localized* using corpus methods, ensuring that the interpretation was founded in selected textual features that had been chosen in a replicable and rigorous manner. There is thus an important interplay between the approaches. The selection of textual features for further interpretation constitutes "new and additional information on the data" in itself, and it is unlikely that it could have been arrived at without using corpus methods as it is improbable that these patterns (particularly regarding 'thing', 'things' and something') could have been detected by the naked eye.

In chapter 5, the corpus-assisted analysis revealed that, as suspected, the senses were addressed and that colours were particularly foregrounded in the novel. This analysis provided sound evidence for claiming that colours are uniquely used in Wilde's novel (relative to his other works, other Gothic works, and his contemporaries' works). This comparative claim is one that would be unlikely to have been made without corpus methods, meaning that the approach was successful. The second analysis also led to a partial exposition of the structures of foregrounding. How are colours foregrounded? How is consumerism thematized through the foregrounding of objects? And to what effect? These questions were not answered with corpus methods but with cognitive stylistics and foregrounding theory. However, the questions would not have been posed had it not been for the preliminary corpus analysis, discerning important and interesting patterns.

The results of the corpus-driven analysis should also be evaluated in the light of Fischer-Starcke's second criterion, stating that corpus methods have been successfully employed when they can support or refute existing knowledge of the text and its interpretation. In order to evaluate whether this has been the case, the existing studies of style in *PDG* should be compared to the present thesis. However, as mentioned in the introduction, no such studies of style seem to exist, and therefore, I compare my results to those of literary critics.

Literary critic Richard Ellmann's work (1988) on Wilde's authorship is largely biographical. His interpretation of *The Picture of Dorian Gray* is founded in *historical* evidence: personal letters, correspondence between authors, letters to editors, accounts of Wilde's Oxford days, etc. Ellmann's interpretation of *PDG* is also often intertextual, drawing particularly on readings of Goethe's *Faust* (1808) and Huysmans' *À Rebours* (1884). While these are interesting frames for interpretation, Wilde's *style* remains unmentioned. Such interpretations are nearly impossible to compare to the concrete textual findings yielded by this study. Whether or not the novel was considered immoral in Victorian England, or whether Dorian is moulded

on Huysmans' character Des Esseintes is too text-distant to be comparable in a stylistic context, where the scope concerns the effect created by certain words, defamiliarizations, and foregrounding techniques. His readings involve context whereas this study primarily draws on co-text. However, the present conclusions can serve as an explication to some of the assumptions made by Ellmann. For example, he states that "the life of mere sensation is uncovered as anarchic and self-destructive" (1988:315). This statement shows that Ellmann recognizes the motif of sensuality, one that is also embedded in the textual level. The present thesis can thus be used to scaffold the impressionist claims made by the literary critic.

In *The Cambridge Companion to Oscar Wilde* (1997), edited by Wilde expert Peter Raby, the same tendency is found. However, this book is divided in three sections: *Context*, *Wilde's Works*, and *Themes and Influences*. From the Victorian context, Gagnier (1997) makes interpretations concerning consumerism that can be compared to the present results, as I have already done in 5.1.3. Calloway offers a brief evaluation of Wilde's style which he claims "is seldom less than precious; it is highly crafted, even lapidary in style, but saved from being tiresome by the fact that often, too, his phrasing can be delicately expressive and amusingly ironic in its touch" (1997:35). This claim can be somewhat substantiated by the conclusion that beauty is foregrounded, and the many mechanisms of foregrounding and defamiliarization support the claim that the style is 'highly crafted'. The delicate expressiveness can be detected in the subtle sensuous effects, the auditory stimuli, the visual as well as the tactile.

One example of a close reading of *PDG* that does not involve biographical readings is Gillespie, who also includes reader response in his study. Here, his notion of sensuality can be discerned from his scorning many critics for conflating "the issue of sensuality to encompass only eroticism – or even more reductively, homoeroticism" (1995:76). However, Gillespie goes on to review earlier critical readings of the novel, in a pursuit to pin down those homoerotic insinuations rather than the full array of sensual imagery in the novel, concluding that Dorian

is a representative of a New Hedonism. He argues that New Hedonism provides a frame of understanding through which the novel is best interpreted:

> The openness to experience that stands as the central tenet of New Hedonism serves as a useful reminder to the reader seeking the fullest possible response to Wilde's narrative. *The Picture of Dorian Gray* offers a wide range of aesthetic pleasure and leaves to the reader the task of remaining open to the aesthetic options it presents (1995:91)

Gillespie's concluding remark is in line with my observations regarding aesthetic intensity which is found to be a defining characteristic on both the corpus-driven and the corpus-assisted analyses.

In his article on the novel, Platizky (2002) argues that Dorian's collection mania can be paralleled to the empire building of the late Victorian period and that his ambiguous emotions towards his assembled treasures can be paralleled to the colonizer's paradoxical fascination and contempt for the inhabitants of colonized areas. While the results of this thesis are not strictly comparable to Platizky's conclusion, they serve to demonstrate how Plaizky (or any other reader) can arrive at the conclusion that *consumption* is a motif. As such, the results serve to offer "precise linguistic descriptions, which can substantiate otherwise impressionistic interpretations of literary texts" (Verdonk, 2004:126-127). This form of scaffolding of impressionistic interpretations is clearly in line with Fischer-Starcke's second criterion for successful corpus analyses.

In light of the comparison of the present conclusions and the conclusions of literary critics, it is now possible to re-evaluate the literary scholars' scepticism towards quantitative methods as introduced in 2.6. The analyses do concern "low level language features" (Stubbs, 2014:48) as these constitute an informed point of departure for further analysis – one that is qualitative in its approach and involves the *co-text*. However, I do refrain from involving the

context too much, as this aspect has proven to overshadow the works of Oscar Wilde in many literary scholars' works. Here, the focus often shifts from the author's works to his person, his life and scandals, the Aesthetic movement, and its members, etc. And while it may be interesting to locate the historical figures from Wilde's social circle that gave him the inspiration to create his characters Dorian Gray, Basil Hallward, and Lord Henry, it is an endeavour that distances the scholar from the specific literary aspect and moves towards the historical and cultural. The author's own attitude towards his work can surely be interesting as well and does require great involvement of the context but the goal of this thesis was sooner to assess the readerly attitude. Therefore, I estimate that the minimal involvement of historical and cultural context and the strong focus on co-text has indeed been appropriate. Overall, the results yielded in the present analyses should be considered *valuable* contributions to the study of Oscar Wilde's work and his style, both because they are very concise and because they are founded in quantitative evidence, adding strength to the argumentation.

Before moving on to a comparison of the corpus stylistic methods, I will briefly comment on the cognitive stylistic approach since it is methodologically important. What the cognitive approach has contributed is particularly tied in with the reading experience and the qualitative analysis of the discovered motifs. Since the sensory aspect was important in both analyses, the cognitive stylistic approach has proven useful, combining the textual and the human cognitive experience. In providing this bridge, cognitive stylistics was a sound choice for the investigation of textual effects specifically and style generally. Moreover, the cognitive angle has ensured a stronger argumentation for those aspects of reading that are typically thought of as purely subjective: the reading experience; the experience of textual stimuli; the effect of words and textual patterns. Therefore, this approach enabled the study of those evasive mechanisms that come into play when a reader engages with a text.

6.3. Comparing the corpus-driven and the corpus-assisted approaches

The thesis structure gives rise to a comparison and evaluation of the two approaches employed. In the following, I draw on my experience with both approaches and generalize where possible, in order to contribute to the methodological discussion and understanding. Initially, I will acknowledge that the case of *PDG* has proven to be, to some extent, ill-suited for a comparison of approaches, because existing academic literature concerning this work is scarce, and literature concerning the stylistic dimension appears to be by and large non-existent. Therefore, it is easy to claim originality in the untrodden territory of Wilde's style. However, I did encounter obstacles and advantages with both approaches that remain generalizable.

The corpus-driven analysis offers the advantage of increased originality. Hereby, I mean that the lack of imposed direction can and has led to discoveries that have not previously been 'foreshadowed' by literary critics. In the present case, the originality stems from my attempt to 'go in blind'. Conversely, the corpus-driven method also has its clear disadvantages, and a major weakness is the risk of being side-tracked by an unproductive approach as happened in 4.1, when it turned out that the analysis of tags yielded very little insight at great cost of time. The results turned out to be unfit for generalizations about the text's style and effect[43], and the track was abandoned. Another pitfall with the corpus-driven approach was that results were often more *detached from the co-text*, warranting greater effort from the stylistician (myself) to establish a link between the observation and the meaning-making of the text. Some results bore characteristics of the famous *hapax legomenon* in that they appeared very significant but did not make up any pattern characterizing the novel's style. However, once this link between observation and co-text was re-established, the results were trustworthy and possible to use in a generalizable and interpretative manner. Another slight disadvantage of the approach was the lack of inter-linkage between the results, meaning that any shift in focus (e.g. from

[43] Those results were perhaps more linguistically than stylistically interesting.

verbs to adjectives) caused quite dissimilar frames of interpretation. This issue meant that the results appeared more slightly fragmented and less like a united analysis and argumentation. However, the issue is not one that is *intrinsic* to the corpus-driven approach, but rather one that is due to my analysis design, in which I investigate several, non-linked significant results. Yet, it is difficult to imagine a design in which the analyst is guaranteed to discover patterns that warrant further analysis, as this is a factor that is highly dependent on the analysed text. In retrospect, the approach could have been more observant of Tognini-Bonelli's claim that "observation leads to hypothesis leads to generalization leads to unification in theoretical statement" (2001:85). Not having formed hypotheses on basis of the observations but rather directly analysing them in a cognitive stylistic frame was perhaps not the optimal approach, although the analysis did yield cognitive insight. It is not easy to assess whether this would have created better or simply different results. Overall, the method yielded noteworthy and unique insight regarding the text's style and effects. It is my assessment that, in stylistics, the method is probably best used in combination with the corpus-assisted method.

With the corpus-assisted analysis, the stylistician is sure to be on the hot scent from the get-go, which is one of the approach's greatest advantages and one that I experienced in the present thesis. Having a hypothesis at the outset meant that the analysis was given a clear direction, and that a greater exactitude in the argumentation was made possible. The disadvantage of using this approach is that it locks the stylistician's focus rather tightly compared to the corpus-driven method. The issue with a narrow focus is that the stylistician cannot be sensitive of some observations that are then lost. Therefore, it is once again my assessment that the approaches are probably best used in some combination. The ratio of such a combination should depend on the goal and text studied, and it is not possible to make generalizations from the present study in which the distribution was approximately 40% corpus-driven and 60% corpus-assisted.

In section 3.1. I outlined the discussion regarding the two approaches, and the impression I was left with was that the corpus-assisted approach would have a limited potential when it came to originality and innovation. However, this has proven to be less true than expected. In the present case, the results yielded from the corpus-assisted analyses proved to be innovative, but that may be due both to the fact that previous studies of the novel's style were not found and to the possibility that the original hypothesis was innovative in itself, creating a spill over effect in the actual analysis. In a summarizing conclusion, both approaches are viable and productive approaches within stylistics albeit a combination does appear to be advisable.

7. References

Anthony, L. (2014). *AntConc Readme Helpfile* (ed. 001). Tokyo: Waseda University. Available from https://www.laurenceanthony.net/software/antconc/

Anthony, L. (2015). *TagAnt* (Version 1.2.0) [Computer Software]. Tokyo: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Antony, L. (2019). *TreeTagger Tag Set (58 tags) pdf-file*. Tokyo: Waseda University. Available from https://www.laurenceanthony.net/software/tagant/

Anthony, L. (2019). *AntConc* (Version 3.5.8) [Computer Software]. Tokyo: Waseda University. Available from https://www.laurenceanthony.net/software

Archer, D. (2007). Computer-assisted Literary Stylistics: The State of the Field. In M. Lambrou, & P. Stockwell (Eds.), *Contemporary Stylistics* (pp. 244-256). London and New York: Continuum.

Balossi, G. (2014). *A Corpus Linguistic Approach to Literary Language and Characterization.* Amsterdam and Philadelphia: John Benjamins Publishing Company.

Banerjee, A., Chitnis, U., Jadhav, S., Bhawalkar, J., & Chaudhury, S. (2009). Hypothesis Testing, Type I and Type II Errors. *Industrial Psychiatry Journal, Vol 18*(2), 127-131. doi:https://doi.org/10.4103/0972-6748.62274

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing, Vol. 8*(4), 243–257. https://doi.org/10.1093/llc/8.4.243

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* Harlow: Pearson Education.

Bradford, R. (1997). *Stylistics.* New York and London: Routledge.

Brezina, V. (2018). *Statistics in Corpus Linguistics - A Practical Guide.* Cambridge: Cambridge University Press.

Burke, M. (2006). Stylistics, Cognitive. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics*, (2nd ed., pp. 218-221). Elsevier.

Burke, M. (2014a). Rhetoric and poetics: the classical heritage of stylistics. In M. Burke (Ed.), *The Routledge Handbook of Stylistics* (pp. 11-30). New York and London: Routledge.

Burke, M. (2014b). Stylistics: From classical rhetoric to cognitive neuroscience. In M. Burke (Ed.), *The Routledge Handbook of Stylistics* (pp. 1-7). New York and London: Routledge.

Burke, M. & Evers, K. (2014). Formalist Stylistics. In M. Burke (Ed.), *The Routledge Handbook of Stylistics* (pp. 31-44). New York and London: Routledge.

Calloway, S. (1997). Wilde and the Dandyism of the Senses. In P. Raby (Ed.), *The Cambridge Companion to Oscar Wilde* (pp. 34-56). Cambridge: Cambridge University Press.

Carter, R. (2010). Methodologies for stylistic analysis: Practices and pedagogies. In D. McIntyre and B. Busse (Eds.), *Language and Style*. London: Palgrave, 55–68.

Croft, W., & Cruse, D. (2004). *Cognitive Linguistics.* Cambridge: Cambridge University Press.

Dienes, Z., & McLatchie, N. M. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin and Review, Vol. 25*(1), 207-218. doi:https://doi.org/10.3758/s13423-017-1266-z

Dowling, L. (1998). Aestheticism. In *Encylopedia of Aesthetics*: Oxford University Press. Retrieved May 1st, 2020, from https://www-oxfordreference-

com.zorac.aub.aau.dk/view/10.1093/acref/9780195113075.001.0001/acref-9780195113075-e-0009

Emmott, C., & Alexander, M. (2014). Foregrounding, burying and plot construction. In P. Stockwell, & S. Whiteley (Eds.), *The Cambridge Handbook of Stylistics* (pp. 329-343). Cambridge: Cambridge University Press.

Emmott, C., Alexander, M., & Marszalek, A. (2014). Schema Theory in Stylistics. In M. Burke (Ed.), *The Routledge Handbook of Stylistics* (pp. 268-283). New York and London: Routledge.

Emmott, C., & Alexander, M. (2016). Defamiliarzation and Foregrounding: Representing Experiences of Change of State and Perception in Neurological Illness Autobiographies. In V. Sotirova (Ed.), *The Bloomsbury Companion to Stylistics* (pp. 289-307). London and New York: Bloomsbury Academic.

Fialho, O., & Zyngier, S. (2014). Quantitative methodological approaches to stylistics. In M. Burke (Ed.), *The Routledge Handbook of Stylistics* (pp. 329-345). New York and London: Routledge.

Fillmore, C. (1982). Towards a Descriptive Framework for Spatial Deixis. In R. J. Jarvella, & W. Klein (Eds.), *Speech, Place, and Action* (pp. 31-59). London: John Wiley.

Fischer-Starcke, B. (2010). *Corpus Linguistics in Literary Analysis: Jane Austen and her Contemporaries.* London and New York: Continuum.

Ganginer, R. (1997). Wilde and the Victorians. In P. Raby (Ed.), *The Cambridge Companion to Oscar Wilde* (pp. 18-33). Cambridge: Cambridge University Press.

Gavins, J. (2014). Defamiliarisation. In P. Stockwell, & S. Whiteley (Eds.), *The Cambridge Handbook of Stylistics* (pp. 196-211). Cambridge: Cambridge University Press.

Gillespie, M. (1995). *The Picture of Dorian Gray: "What the World Thinks Me"*. New York: Twayne Publishers.

Gothic fiction (2012). In Birch, D., & Hooper, K. (Eds.), *The Concise Oxford Companion to English Literature:* Oxford University Press. Retrieved March 13. 2020, from https://www-oxfordreference com.zorac.aub.aau.dk/view/10.1093/acref/9780199608218.001.0001/acref-9780199608218-e-3192.

Gries, S. T. (2009). *Statistics for Linguistics with R: A Practical Introduction*. Berlin and New York: Mouton de Gruyter.

Hall, G. (2014). Pedagogical Stylistics. In M. Burke (Ed.), *The Routledge Handbook of Stylistics* (pp. 239-252). New York and London: Routledge.

Hart, M. (2006, November, 23). *Project Gutenberg Principle of Minimum Regulation*. Retreived March 13, 2020 from Project Gutenberg: https://www.gutenberg.org/wiki/Gutenberg:Project_Gutenberg_Principle_of_Minimal_Regulation_/_Administration_by_Michael_Hart_and_Greg_Newby

Hart, M. (2010, April, 8). *The History and Philosophy of Project Gutenberg*. Retreived March 13, 2020 from Project Gutenberg: https://www.gutenberg.org/wiki/Gutenberg:The_History_and_Philosophy_of_Project_Gutenberg_by_Michael_Hart

Holland, M (1997). Biography and the Art of Lying. In P. Raby (Ed.), *The Cambridge Companion to Oscar Wilde* (pp. 3-17). Cambridge: Cambridge University Press.

Jarosz, A. F., & Wiley, J. (2014). What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors. *Journal of Problem Solving, Vol. 7*, 2-9. doi:http://dx.doi.org/10.7771/1932-6246.1167

Jeffries, L., & McIntyre, D. (2010). *Stylistics.* Cambridge: Cambridge University Press.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things.* Chicago: University of Chicago Press.

Langacker, R. W. (1991). *Foundations of Cognitive Grammar. Vol. 2. Descriptive Application.* Stanford: Stanford University Press.

Lu, X. (2014). *Computational Methods for Corpus Annotation and Analysis.* New York, Heidelberg, Dordrecht and London: Springer.

Mahlberg, M. (2007). A Corpus Stylistic Perspective on Dicken's Great Expectations. In M. Lambrou, & P. Stockwell (Eds.), *Contemporary Stylistics* (pp. 19-31). London and New York: Continuum.

Mahlberg, M. (2012). *Corpus Stylistics and Dickens' Fiction.* New York and London: Routledge.

Mahlberg, M. (2014). Corpus Stylistics. In M. Burke (Ed.), *The Routledge Handbook of Stylistics* (pp. 378-392). New York and London: Routledge.

Mahlberg, M. (2016). Corpus Stylistics. In V. Sotirova (Ed.), *The Bloomsbury Companion to Stylistics* (pp. 139-156). London and New York: Bloomsbury Academic.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies - An Advanced Resource Book.* New York and London: Routledge.

McEnery, T., & Hardie, A. (2011, October 31). *Statistics in corpus linguistics*. Retrieved March 4, 2020 from Helppage for Corpus Linguistics: Method, theory and practice: http://corpora.lancs.ac.uk/clmtp/2-stat.php

McIntyre, D. (2014). Characterisation. In P. Stockwell, & S. Whiteley (Eds.), *The Cambridge Handbook of Stylistics* (pp. 149-164). Cambridge: Cambridge University Press.

Mukařovský, J. (1964). Standard Language and Poetic Language. In P. L. Garvin (Ed.), *A Prague School Reader on Esthetics, Literary Structure, and Style* (pp. 17-30). Washington D.C.: Georgetown University Press.

Jensen, M., Lottrup, K., & Nordentoft, S. (2018). Floral foregrounding: Corpus-assisted, Cognitive Stylistic Study of the Foregrounding of Flowers in Mrs. Dalloway. *Globe, Vol. 7*, 34-56. doi:10.5278/ojs.globe.v7i0.2443

Nørgaard, N., Montoro, Rocío, R., & Busse, B. (2010). *Key Terms in Stylistics.* New York and London: Continuum.

OTA Home. (2019). Retreived March 10, 2020 from Oxford Text Archive: https://ota.bodleian.ox.ac.uk/repository/xmlui/#

Pallant, J. (2007). *SPSS Survival Manual* (3rd edition). Maidenhead: Open University Press - McGraw Hill Education.

Pallant, J. (2016). *SPSS Survival Manual* (6th edition). New York: Open Univeristy Press - McGraw Hill Education.

Penguin Classics. (n.d.). Penguin. Retreived March 10, 2020, from Read Penguin USA: https://www.penguin.com/books/browse/1125/classics?filter=

Platizky, R. (2002). Wilde's the Picture of Dorian Gray. *The Explicator, Vol. 60*(4), 202-204. https://www.tandfonline.com/doi/abs/10.1080/00144940209597714

Rayson, P. (2011). *Log-likelihood and effect size calculator*. Lancaster: UCREL - University Centre for Computer Corpus Research on Language. Retreived from http://ucrel.lancs.ac.uk/llwizard.html

Rissanen, M. (2018). Three Problems Connected with the Use of Diachronic Corpora. *ICAME Journal, Vol. 42*(1). 9-12. 10.1515/icame-2018-0002

Rumelhart, D. E., & Norman, D. (1978). Accretion, tuning and restructuring: three modes of learning. In J. W. Cotton, & R. Klatzky (Eds.), *Semantic Factors in Cognition* (pp. 37-53). Hillsdale: Lawrence Erlbaum Associates.

Rumelhart, D. E., & Ortony, A. (1977). The Representation of Knowledge in Memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 99-135). Hillsdale: Lawrence Erlbaum Associates.

Schmid, H. J. (2007). Entrenchment, Salience and Basic Levels. In D. Geeraerts, & H. Cuyckens (Eds.), *Handbook of Cognitive Linguistics* (pp. 117-138). Oxford: Oxford University Press.

Simpson, P. (2014). *Stylistics - A Resource Book for Students.* New York and London: Routledge.

Šklovsky, V. (2012). Art as Technique. In L. T. Lemon, & M. Reis (Eds.), *Russian Formalist Criticism: Four Essays* (pp. 3-24). Lincoln and London: University of Nebraska Press.

Something-Lexico. (2020). *Something, Meaning of Something*. Retreived April 2, 2020 from Lexico: https://www.lexico.com/definition/something

Stockwell, P. (2002). *Cognitive Poetics - An Introduction.* New York and London: Routledge.

Stockwell, P. (2003). Schema Poetics and Speculative Cosmology. *Language and Literature, Vol. 12*(3), 252-271. https://doi.org/10.1177/09639470030123005

Stockwell, P., & Whiteley, S. (2014). Introduction. In P. Stockwell, & S. Whiteley (Eds.), *The Cambridge Handbook of Stylistics* (pp. 1-9). Cambridge: Cambridge University Press.

Stubbs, M. (2001). *Words and Phrases. Corpus Studies in Lexical Semantics.* Oxford: Blackwell Publishers.

Stubbs, M. (2014). Quantitative Methods in Literary Linguistics. In P. Stockwell, & S. Whiteley (Eds.), *The Cambridge Handbook of Stylistics* (pp. 46-62). Cambridge: Cambridge University Press.

Säily, T., & Siirtola, H. (2011). Variation in Noun and Pronoun Frequencies in a Sociohistorical Corpus of English. *Literary and Linguistic Computing, Vol. 26*(2), 167-188. doi:10.1093/llc/fqr004

Taylor, J. R. (1989). *Linguistic Categorization. Prototypes in Linguistic Theory.* Oxford: Clarendon Press.

Thing-Lexico. (2020). *Thing, Meaning of Thing*. Retreived April 2, 2020 from Lexico: https://www.lexico.com/definition/thing

Thompson, J. (2020). *Victorian Gothic: An Introduction*. Retreived March 10, 2020 from Great Writers Inspire: http://writersinspire.podcasts.ox.ac.uk/content/victorian-gothic-introduction

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work.* Amsterdam: John Benjamins Publishing Company.

Toolan, M. (2014). The Theory and Philosophy of Stylistics. In P. Stockwell, & S. Whiteley (Eds.), *The Cambridge Handbook of Stylistics* (pp. 13-31). Cambridge: Cambridge University Press.

Wales, K. (1989). *A Dictionary of Stylistics.* Singapore: Longman Group.

Walker, I. (1. August 2007). *Null hypothesis testing and effect sizes*. Retreived March 24, 2020 from Statistics for Psychology: https://people.bath.ac.uk/pssiw/stats2/page2/page14/page14.html

West, D. (2016). Cognitive Stylistics. In V. Sotirova (Ed.), *The Bloomsbury Companion to Stylistics* (pp. 109-121). New York: Bloomsbury.

Wilde, O. (1891a). The Critic as Artist. In O. Wilde (Ed.), *Intentions* (pp. 95-224). London: Methuen & Co. Ltd.

Wilson, A. (2013). Embracing Bayes factors for key item analysis in corpus linguistics. In M. Bieswanger, & A. Koll-Stobbe (Eds.), *New Approaches to the Study of Linguistic Variability*. Language Competence and Language Awareness in Europe, (4th ed., pp. 3-11). Frankfurt: Peter Lang.

Zyngier, S. (2008). Macbeth through the Computer. In W. vanPeer (Ed.), *The Quality of Literature: Linguistic Studies in Literary Evaluation* (pp. 169-190). Amsterdam and Philadelphia: John Benjamins Publishing Company.

8. Corpora references

Abbott, E. A. (1884). *Flatland*. EBook #201. January 1995. https://www.gutenberg.org/ebooks/201 (Accessed 17 March 2020).

Ainsworth, W.H. (1842). *Windsor Castle*. EBook #2866. October 2001. https://www.gutenberg.org/ebooks/2866 (Accessed 17 March 2020).

Alcott, L.M. (1868-69). *Little Women*. EBook #514. May 1996. https://www.gutenberg.org/ebooks/514 (Accessed 17 March 2020).

Austen, J. (1817). *Northanger Abbey*. EBook #121. April 1994. https://www.gutenberg.org/ebooks/121 (Accessed 17 March 2020).

Bierce, A. (1890). *An Occurrence at Owl Creek Bridge*. EBook #375. December 1995. https://www.gutenberg.org/ebooks/375 (Accessed 17 March 2020).

Blackmore, R.D. (1869). *Lorna Doone*. EBook #17460. January 2006. https://www.gutenberg.org/ebooks/17460 (Accessed 17 March 2020).

Blackwood, A. (1908). *The Willows*. EBook #11438. March 2004. https://www.gutenberg.org/ebooks/11438 (Accessed 17 March 2020).

Brontë, A. (1847). *Agnes Grey*. EBook #767. December 1996. https://www.gutenberg.org/ebooks/767 (Accessed 17 March 2020).

Brontë, C. (1847). *Jane Eyre*. EBook #1260. March 1998. https://www.gutenberg.org/ebooks/1260 (Accessed 17 March 2020).

Brontë, C. (1849). *Shirley*. EBook #30486. November 2009. https://www.gutenberg.org/ebooks/30486 (Accessed 17 March 2020).

Brontë, E. (1847). *Wuthering Heights*. EBook #768. December 1996. https://www.gutenberg.org/ebooks/768 (Accessed 17 March 2020).

Butler, S. (1872). *Erewhon*. EBook #1906. September 1999. https://www.gutenberg.org/ebooks/1906 (Accessed 17 March 2020).

Carlyle, T. (1836). *Sartor Resartus*. EBook #1051. September 1997. https://www.gutenberg.org/ebooks/1051 (Accessed 17 March 2020).

Carroll, L. (1865). *Alice's Adventures in Wonderland*. EBook #928. May 1997. https://www.gutenberg.org/ebooks/928 (Accessed 17 March 2020).

Collins, W. (1859). *The Woman in White*. EBook #583. July 1996. https://www.gutenberg.org/ebooks/583 (Accessed 17 March 2020).

Conrad, J. (1902). *Heart of Darkness*. EBook #526. January 2006. https://www.gutenberg.org/ebooks/526 (Accessed 17 March 2020).

Dickens, C. (1837-39). *Oliver Twist*. EBook #730. November 1996. https://www.gutenberg.org/ebooks/730 (Accessed 17 March 2020).

Dickens, C. (1848). *A Christmas Carol*. EBook #30368. October 2009. https://www.gutenberg.org/ebooks/30368 (Accessed 17 March 2020).

Doyle, A.C. (1892). *The Adventures of Sherlock Holmes*. EBook #1661. March 1999. https://www.gutenberg.org/ebooks/1661 (Accessed 17 March 2020).

Eliot, G. (1871-72). *Middlemarch*. EBook #145. July 1994. https://www.gutenberg.org/ebooks/145 (Accessed 17 March 2020).

Gaskell, E.C. (1848). *Mary Barton*. OTA identifier:3105. Oxford Text Archive, http://hdl.handle.net/20.500.12024/3105 (Accessed 17 March 2020).

Gilman, C.P. (1892). *The Yellow Wallpaper*. EBook #1952. November 1999. https://www.gutenberg.org/ebooks/1952 (Accessed 17 March 2020).

Gissing, G. (1891). *New Grub Street*. EBook #1709. April 1999. https://www.gutenberg.org/ebooks/1709 (Accessed 17 March 2020).

Hardy, T. (1891). *Tess of the D'Urbervilles*. *A Pure Woman*. EBook #110. February 1994. https://www.gutenberg.org/ebooks/110 (Accessed 17 March 2020).

Hawthorne, N. (1850). *The Scarlett Letter*. EBook #25344. May 2008. https://www.gutenberg.org/ebooks/25344 (Accessed 17 March 2020).

Hughes, T. (1857). *Tom Brown's School Days*. EBook #1480. February 2006. https://www.gutenberg.org/ebooks/1480 (Accessed 17 March 2020).

James, H. (1881). *The Portrait of a Lady* (vol.1). EBook #2833. September 2001. https://www.gutenberg.org/ebooks/2833 (Accessed 17 March 2020).

James, H. (1898). *The Turn of the Screw*. EBook #209. February 1995. https://www.gutenberg.org/ebooks/209 (Accessed 17 March 2020).

Johnston, M. (1899). *To Have and to Hold*. EBook #2807. September 2001. https://www.gutenberg.org/ebooks/2807 (Accessed 17 March 2020).

Kingsley, C. (1850). *Alton Locke, Tailor and Poet*. EBook #8374. June 2005. https://www.gutenberg.org/ebooks/8374 (Accessed 17 March 2020).

Kipling, R. (1888). *The Man Who Would be King*. EBook #8147. May 2005. https://www.gutenberg.org/ebooks/8147 (Accessed 17 March 2020).

Le Fanu, J.S. (1872). *Carmilla*. EBook #10007. November 2003. https://www.gutenberg.org/ebooks/10007 (Accessed 17 March 2020).

Lewis, M. (1796). *The Monk*. EBook #601. Jult 1996. https://www.gutenberg.org/ebooks/601 (Accessed 17 March 2020).

Lytton, E.B.L. (1834). *The Last Days of Pompeii*. EBook #1565. February 2006. https://www.gutenberg.org/ebooks/1565

Marryat, F. (1836). *Mr. Midshipsman Easy*. EBook #21553. May 2007. https://www.gutenberg.org/ebooks/21553 (Accessed 17 March 2020).

Melville, H. (1851). *Moby Dick*. EBook #2701. July 2001. https://www.gutenberg.org/ebooks/2701 (Accessed 17 March 2020).

Nesbit, E. (1892). *Grim Tales*. EBook #40321. July 2012. https://www.gutenberg.org/ebooks/40321 (Accessed 17 March 2020).

Norris, F. (1899). *McTeague*. EBook #165. March 2006. https://www.gutenberg.org/ebooks/165 (Accessed 17 March 2020).

Poe, E.A. (1839). *The Fall of the House of Usher*. EBook #932. June 1997. https://www.gutenberg.org/ebooks/932 (Accessed 17 March 2020).

Radcliffe, A.W. (1794). *The Mysteries of Udolpho*. EBook #3268. June 2002. https://www.gutenberg.org/ebooks/3268 (Accessed 17 March 2020).

Reeve, C. (1777). *The Old English Baron*. EBook #5182. February 2004. https://www.gutenberg.org/ebooks/5182 (Accessed 17 March 2020).

Stevenson, R.L. (1886). *The Stange Case of Dr. Jekyll and Mr. Hyde*. EBook #42. October 1992. https://www.gutenberg.org/ebooks/42 (Accessed 17 March 2020).

Stoker, B. (1897). *Dracula*. EBook #345. October 1995. https://www.gutenberg.org/ebooks/345 (Accessed 17 March 2020).

Thackeray, W.M. (1847). *Vanity Fair*. EBook #599. July 1996. https://www.gutenberg.org/ebooks/599 (Accessed 17 March 2020).

Trollope, A. (1855). *The Warden*. EBook #619. August 1996. https://www.gutenberg.org/ebooks/619 (Accessed 17 March 2020).

Twain. M. (1884). *Adventures of Huckleberry Finn*. EBook #76. June 2004. https://www.gutenberg.org/ebooks/619 (Accessed 17 March 2020).

Walpole, H. (1764). *The Castle of Otranto*. EBook #696. October 1996. https://www.gutenberg.org/ebooks/696 (Accessed 17 March 2020).

Wells, H.G. (1898). *The Time Machine*. EBook #35. October 2004. https://www.gutenberg.org/ebooks/35 (Accessed 17 March 2020).

Wells, H.G. (1911). *The Door in the Wall and Other Stories*. EBook #456. March 1996. https://www.gutenberg.org/ebooks/456 (Accessed 17 March 2020).

Wilde, O. (1880). *Vera; Or the Nihilists*. EBook #26494. August 2008. https://www.gutenberg.org/ebooks/26494 (Accessed 17 March 2020).

Wilde, O. (1882). *Impressions of America*. EBook #41806. January 2013. https://www.gutenberg.org/ebooks/41806 (Accessed 17 March 2020).

Wilde, O. (1887). *The Canterville Ghost*. EBook #14522. December 2004. https://www.gutenberg.org/ebooks/14522 (Accessed 17 March 2020).

Wilde, O. (1888). *The Happy Prince and Other Tales*. EBook #31120. September 2009. https://www.gutenberg.org/ebooks/30120 (Accessed 17 March 2020).

Wilde, O. (1891b). *A House of Pomegranates*. EBook #873. April 1997. https://www.gutenberg.org/ebooks/873 (Accessed 17 March 2020).

Wilde, O. (1891c). *Intentions*. EBook #887. April 1997. https://www.gutenberg.org/ebooks/887 (Accessed 17 March 2020).

Wilde, O. (1891d). *The Soul of Man Under Socialism*. EBook #1017. August 1997. https://www.gutenberg.org/ebooks/1017 (Accessed 17 March 2020).

Wilde, O. (1891e). *The Picture of Dorian Gray*. London: Penguin Group. Penguin English Library. (This version published 2012).

Wilde, O. (1891f). *The Duchess of Padua*. EBook #875. April 1997. https://www.gutenberg.org/ebooks/875 (Accessed 17 March 2020).

Wilde, O. (1892a). *Salomé: A Tragedy in One Act*. EBook #42704. May 2013. https://www.gutenberg.org/ebooks/42704 (Accessed 17 March 2020).

Wilde, O. (1892b). *Lady Wintermere's Fan*. EBook #790. January 1997. https://www.gutenberg.org/ebooks/790 (Accessed 17 March 2020).

Wilde, O. (1893b). *A Woman of No Importance*. EBook #854. March 1997. https://www.gutenberg.org/ebooks/854 (Accessed 17 March 2020).

Wilde, O. (1894). *A Florentine Tragedy – La Sainte Courtisane*. EBook #1308. May 1998. https://www.gutenberg.org/ebooks/1308 (Accessed 17 March 2020).

Wilde, O. (1895). *The Importance of Being Earnest*. EBook #844. March 1997. https://www.gutenberg.org/ebooks/844 (Accessed 17 March 2020).

Wilde, O. (1898). *Poems, with the Ballad of Reading Gaol*. EBook #1057. October 1997. https://www.gutenberg.org/ebooks/1057 (Accessed 17 March 2020).

Wilde, O. (1905). *De Profundis*. Ebook #921. May 1997. https://www.gutenberg.org/ebooks/921 (Accessed 17 March 2020).

Wilde, O. (1908). *Reviews*. EBook #14240. December 2004. https://www.gutenberg.org/ebooks/14240 (Accessed 17 March 2020).

Wilde, O. (1988) *An ideal husband / Oscar Wilde*, OTA identifier:1246. Oxford Text Archive, http://hdl.handle.net/20.500.12024/1246 (Accessed 17 March 2020).

Wollenstonecraft, M.S. (1818). *Frankenstein, Or the Modern Prometheus*. EBook #84. October 1993. https://www.gutenberg.org/ebooks/84 (Accessed 17 March 2020).

Wynne, M.Y. (1895). *The Little Room and Other Stories*. EBook #58855. February 2019. https://www.gutenberg.org/ebooks/58855 (Accessed 17 March 2020).

Younge, C. (1853). *The Heir of Redclyffe*. EBook #2505. February 2001. https://www.gutenberg.org/ebooks/2505 (Accessed 17 March 2020).

9. Appendices

Appendix A: Anthony's Tree Tagger Set

TreeTagger Tag Set (58 tags)

POS Tag	Description	Example
CC	coordinating conjunction	*and, but, or, &*
CD	cardinal number	*1, three*
DT	determiner	*the*
EX	existential there	*there is*
FW	foreign word	*d'œuvre*
IN	preposition/subord. conj.	*in,of,like,after,whether*
IN/that	complementizer	*that*
JJ	adjective	*green*
JJR	adjective, comparative	*greener*
JJS	adjective, superlative	*greenest*
LS	list marker	*(1),*
MD	modal	*could, will*
NN	noun, singular or mass	*table*
NNS	noun plural	*tables*
NP	proper noun, singular	*John*
NPS	proper noun, plural	*Vikings*
PDT	predeterminer	*both the boys*
POS	possessive ending	*friend's*
PP	personal pronoun	*I, he, it*
PP$	possessive pronoun	*my, his*
RB	adverb	*however, usually, here, not*
RBR	adverb, comparative	*better*
RBS	adverb, superlative	*best*
RP	particle	*give up*
SENT	end punctuation	*?, !, .*
SYM	symbol	*@, +, *, ^, |, =*
TO	*to*	*to go, to him*
UH	interjection	*uhhuhhuhh*

POS Tag	Description	Example
VB	verb *be*, base form	*be*
VBD	verb *be*, past	*was|were*
VBG	verb *be*, gerund/participle	*being*
VBN	verb *be*, past participle	*been*
VBZ	verb *be*, pres, 3rd p. sing	*is*
VBP	verb *be*, pres non-3rd p.	*am|are*
VD	verb *do*, base form	*do*
VDD	verb *do*, past	*did*
VDG	verb *do* gerund/participle	*doing*
VDN	verb *do*, past participle	*done*
VDZ	verb *do*, pres, 3rd per.sing	*does*
VDP	verb *do*, pres, non-3rd per.	*do*
VH	verb *have*, base form	*have*
VHD	verb *have*, past	*had*
VHG	verb *have*, gerund/participle	*having*
VHN	verb *have*, past participle	*had*
VHZ	verb *have*, pres 3rd per.sing	*has*
VHP	verb *have*, pres non-3rd per.	*have*
VV	verb, base form	*take*
VVD	verb, past tense	*took*
VVG	verb, gerund/participle	*taking*
VVN	verb, past participle	*taken*
VVP	verb, present, non-3rd p.	*take*
VVZ	verb, present 3d p. sing.	*takes*
WDT	wh-determiner	*which*
WP	wh-pronoun	*who, what*
WP$	possessive wh-pronoun	*whose*
WRB	wh-abverb	*where, when*
:	general joiner	*;, -, --*
$	currency symbol	*$, £*

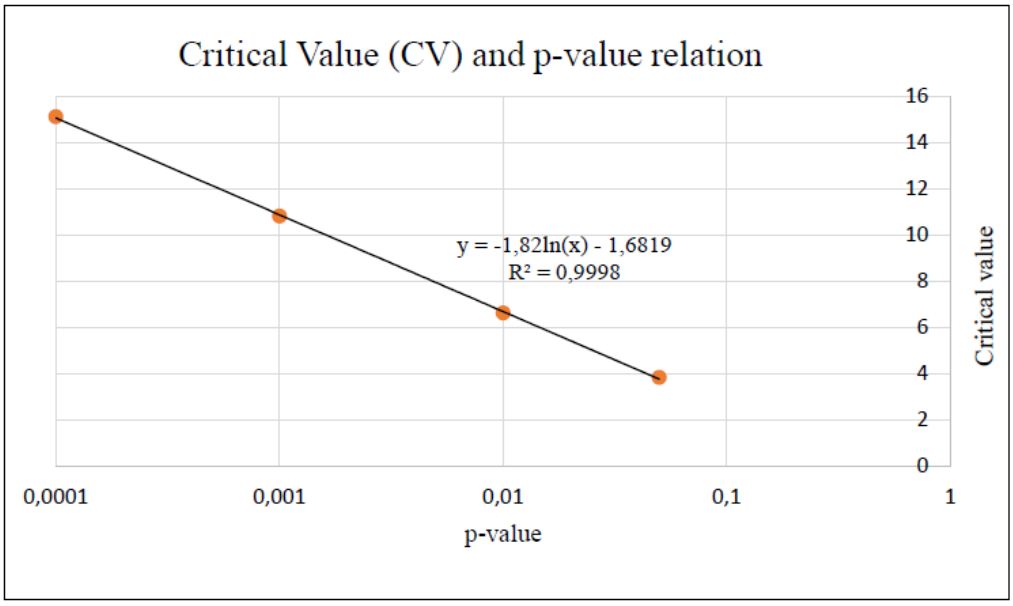
Adapted from: https://courses.washington.edu/hypertxt/csar-v02/penntable.html

Appendix B: *p* value calculation

p	CV
0,05	3,84
0,01	6,63
0,001	10,83
0,0001	15,13

CV=-1,82*ln(CV)-1,6819

0,002	9,62869



Appendix C: Keyword analysis

Keywords in *PDG* compared to *Wilde*

Rank	Freq.		Keyness (LL4)	Effect (DICE)	Keyword
1	414	+	1637.68	0.0102	dorian
2	236	+	750.53	0.0058	henry
3	833	+	738.78	0.02	had
4	199	+	731.86	0.0049	gray
5	178	+	667.43	0.0044	harry
6	158	+	616.89	0.0039	basil
7	1541	+	586.38	0.0354	he
8	1083	+	523.16	0.0255	was
9	129	+	437.49	0.0032	xa*
10	1445	+	422.59	0.0332	you
11	91	+	367.29	0.0023	hallward
12	85	+	324.59	0.0021	sibyl
13	661	+	282.54	0.0159	him
14	418	+	235.49	0.0102	t**
15	58	+	217.11	0.0014	vane
16	996	+	200.14	0.0233	his
17	255	+	196.82	0.0063	don
18	119	+	195.77	0.0029	seemed
19	1694	+	171.25	0.0376	i
20	76	+	129.41	0.0019	felt
21	1343	+	123.49	0.0304	it
22	52	+	119.92	0.0013	portrait
23	52	+	110.71	0.0013	lad
24	94	+	101.82	0.0023	looked
25	126	+	100.81	0.0031	something
26	262	+	100.5	0.0064	said
27	33	+	98.07	0.0008	alan
28	42	+	97.23	0.001	murmured
29	22	+	88.78	0.0005	narborough
30	95	+	87.27	0.0023	cried
31	28	+	86.35	0.0007	campbell
32	87	+	85.52	0.0021	answered
33	308	+	80.4	0.0075	would
34	105	+	78.07	0.0026	want
35	54	+	71.6	0.0013	turned
36	107	+	69.34	0.0026	face
37	69	+	69.14	0.0017	picture
38	223	+	67.67	0.0055	about

39	431	+	66.93	0.0104	she
40	523	+	66.84	0.0125	me
41	169	+	65.29	0.0042	over
42	48	+	65.18	0.0012	horrible
43	23	+	64.96	0.0006	senses
44	61	+	64.49	0.0015	asked
45	15	+	60.53	0.0004	jim
46	18	+	60.2	0.0004	screen
47	1361	+	59.48	0.0305	that
48	19	+	56.56	0.0005	glanced
49	23	+	53.15	0.0006	canvas
50	36	+	52.55	0.0009	mad
51	13	+	52.46	0.0003	valet
52	109	+	52.27	0.0027	eyes
53	40	+	51.15	0.001	painter
54	562	+	50.66	0.0134	at
55	18	+	49.98	0.0004	studio
56	41	+	48.49	0.001	laughed
57	28	+	48.06	0.0007	shook
58	22	+	48.05	0.0005	exclaimed
59	47	+	47.74	0.0012	suddenly
60	67	+	47.72	0.0017	door
61	399	+	46.23	0.0096	what
62	87	+	45.98	0.0021	going
63	13	+	45.54	0.0003	erskine
64	102	+	45.26	0.0025	room
65	60	+	44.67	0.0015	passed
66	59	+	44.38	0.0015	table
67	70	+	43.32	0.0017	looking
68	16	+	42.94	0.0004	upstairs
69	24	+	41.92	0.0006	chapter
70	12	+	41.66	0.0003	geoffrey
71	13	+	41.25	0.0003	sighed
72	251	+	41.22	0.0061	were
73	236	+	40.68	0.0058	been
74	232	+	40.63	0.0057	am
75	17	+	39.54	0.0004	muttered
76	40	+	39.09	0.001	met
77	32	+	38.23	0.0008	hideous
78	136	+	37.63	0.0033	could
79	24	+	37.16	0.0006	brain
80	37	+	35.5	0.0009	youth
81	31	+	35.17	0.0008	chair
82	44	+	34.63	0.0011	began

83	99	+	34.47	0.0024	came
84	18	+	34.18	0.0004	key
85	10	+	33.94	0.0002	hesitated
86	10	+	33.94	0.0002	realize
87	36	+	33.47	0.0009	yourself
88	43	+	32.81	0.0011	dreadful
89	25	+	32.74	0.0006	fingers
90	69	+	32.41	0.0017	everything
91	8	+	32.28	0.0002	hubbard
92	8	+	32.28	0.0002	singleton
93	21	+	31.43	0.0005	coat
94	77	+	30.84	0.0019	told
95	27	+	30.78	0.0007	sins
96	18	+	30.19	0.0004	started
97	10	+	30.11	0.0002	gladys
98	22	+	29.99	0.0005	broke
99	176	+	29.48	0.0043	know
100	12	+	29.25	0.0003	selby
101	18	+	29	0.0004	horribly
102	128	+	28.84	0.0031	things
103	23	+	28.83	0.0006	drew
104	26	+	28.56	0.0006	entered
105	80	+	28.36	0.002	round
106	7	+	28.25	0.0002	rejoined
107	20	+	27.87	0.0005	flung
108	22	+	27.32	0.0005	changed
109	23	+	27.18	0.0006	passions
110	30	+	26.98	0.0007	coloured
111	28	+	26.66	0.0007	moments
112	9	+	26.46	0.0002	hansom
113	17	+	26.32	0.0004	smiled
114	8	+	26.29	0.0002	kelso
115	54	+	26.27	0.0013	got
116	29	+	25.6	0.0007	opened
117	12	+	25.41	0.0003	wondered
118	106	+	25.16	0.0026	himself
119	21	+	24.99	0.0005	remembered
120	10	+	24.74	0.0002	exhibit
121	30	+	24.68	0.0007	became
122	99	+	24.62	0.0024	away
123	42	+	24.34	0.001	wish
124	6	+	24.21	0.0001	hetty
125	6	+	24.21	0.0001	hostess
126	6	+	24.21	0.0001	inquest

127	6	+	24.21	0.0001	winced
128	63	+	24.17	0.0016	went
129	147	+	24.15	0.0036	then
130	35	+	24.09	0.0009	forget
131	11	+	24.03	0.0003	gazing
132	2101	+	23.93	0.0442	to
133	31	+	23.88	0.0008	sat
134	61	+	23.67	0.0015	afraid
135	20	+	23.49	0.0005	club
136	21	+	23.44	0.0005	minutes
137	29	+	23.35	0.0007	terror
138	48	+	23.28	0.0012	curious
139	8	+	22.84	0.0002	knock
140	10	+	22.68	0.0002	cruelty
141	68	+	22.63	0.0017	anything
142	7	+	22.5	0.0002	sensations
143	13	+	22.49	0.0003	sigh
144	12	+	22.31	0.0003	boyhood
145	24	+	22.31	0.0006	walked
146	84	+	22.2	0.0021	thought
147	25	+	21.89	0.0006	monstrous
148	436	+	21.76	0.0105	one
149	176	+	21.74	0.0043	never
150	337	+	21.62	0.0081	there
151	23	+	21.54	0.0006	front
152	9	+	21.41	0.0002	frame
153	9	+	21.41	0.0002	shrugged
154	13	+	21.27	0.0003	juliet
155	61	+	21.19	0.0015	across
156	10	+	20.89	0.0002	adrian
157	10	+	20.89	0.0002	drove
158	21	+	20.63	0.0005	happened
159	8	+	20.25	0.0002	listlessly
160	8	+	20.25	0.0002	peril
161	8	+	20.25	0.0002	wrinkled
162	5	+	20.18	0.0001	dartmoor
163	5	+	20.18	0.0001	devereux
164	5	+	20.18	0.0001	divan
165	5	+	20.18	0.0001	laboratory
166	5	+	20.18	0.0001	ruxton
167	5	+	20.18	0.0001	shutting

* invalid result

** this is the abbreviated adverb 'not'

Keywords in *PDG* compared to *GothLit*

Rank	Freq.		Keyness (LL4)	Effect (DICE)	Keyword
1	414	+	2490.65	0.0102	dorian
2	178	+	1025.01	0.0044	harry
3	236	+	995.55	0.0058	henry
4	199	+	988.33	0.0049	gray
5	158	+	950.06	0.0039	basil
6	129	+	747.31	0.0032	xa*
7	91	+	547.11	0.0023	hallward
8	85	+	485.17	0.0021	sibyl
9	248	+	454.03	0.0061	lord
10	1445	+	369.43	0.0295	you
11	418	+	357.53	0.01	t**
12	58	+	338.65	0.0014	vane
13	255	+	328.25	0.0062	don
14	907	+	312.48	0.02	is
15	1541	+	273.28	0.0305	he
16	44	+	255.03	0.0011	duchess
17	231	+	223.27	0.0056	life
18	40	+	219.01	0.001	painter
19	33	+	198.38	0.0008	alan
20	52	+	171.18	0.0013	portrait
21	28	+	168.32	0.0007	campbell
22	77	+	164.37	0.0019	art
23	103	+	156.08	0.0025	really
24	52	+	155.3	0.0013	lad
25	436	+	154.93	0.0102	one
26	661	+	142.99	0.0149	him
27	1343	+	137.46	0.0268	it
28	373	+	134.11	0.0088	are
29	22	+	132.25	0.0005	narborough
30	69	+	127.82	0.0017	picture
31	128	+	127.1	0.0031	things
32	29	+	125.98	0.0007	artist
33	58	+	121.66	0.0014	wonderful
34	42	+	119.42	0.001	murmured
35	120	+	110.84	0.0029	thing
36	49	+	109.97	0.0012	gold
37	105	+	108.5	0.0026	want
38	43	+	107.41	0.0011	charming
39	69	+	105.89	0.0017	everything
40	1361	+	103.64	0.0268	that
41	27	+	103.57	0.0007	sins
42	48	+	99.73	0.0012	horrible

43	48	+	95.3	0.0012	curious
44	66	+	93.66	0.0016	women
45	20	+	92.75	0.0005	club
46	110	+	92.69	0.0027	always
47	23	+	90.6	0.0006	canvas
48	15	+	90.17	0.0004	jim
49	28	+	89.83	0.0007	sin
50	223	+	89.47	0.0054	about
51	61	+	87.29	0.0015	across
52	399	+	85.9	0.0093	what
53	126	+	85.2	0.0031	something
54	18	+	84.63	0.0004	studio
55	24	+	84.22	0.0006	romance
56	16	+	83.83	0.0004	personality
57	32	+	83.5	0.0008	hideous
58	30	+	83.05	0.0007	coloured
59	101	+	82.82	0.0025	people
60	32	+	81.89	0.0008	painted
61	213	+	79.44	0.0051	like
62	13	+	78.15	0.0003	erskine
63	13	+	78.15	0.0003	juliet
64	95	+	75.47	0.0023	cried
65	41	+	75.47	0.001	simply
66	18	+	74.75	0.0004	screen
67	996	+	73.68	0.0207	his
68	39	+	72.81	0.001	merely
69	12	+	72.14	0.0003	selby
70	15	+	71.02	0.0004	cigarette
71	17	+	69.57	0.0004	tragedy
72	25	+	69.51	0.0006	monstrous
73	26	+	69.24	0.0006	exquisite
74	19	+	69.08	0.0005	subtle
75	61	+	67.35	0.0015	afraid
76	16	+	66.44	0.0004	shallow
77	41	+	64.98	0.001	laughed
78	40	+	63.58	0.001	play
79	232	+	63.26	0.0056	am
80	18	+	62.81	0.0004	painting
81	107	+	62.75	0.0026	quite
82	112	+	61.45	0.0027	love
83	39	+	61.39	0.001	absolutely
84	18	+	60.13	0.0004	horribly
85	10	+	60.11	0.0002	adrian
86	10	+	60.11	0.0002	gladys

87	55	+	58.87	0.0014	beauty
88	43	+	58.1	0.0011	dreadful
89	29	+	57.75	0.0007	delightful
90	18	+	56.5	0.0004	scarlet
91	12	+	54.55	0.0003	geoffrey
92	87	+	54.23	0.0021	going
93	9	+	54.1	0.0002	wotton
94	12	+	52.05	0.0003	nowadays
95	68	+	51.86	0.0017	soul
96	46	+	51.44	0.0011	passion
97	12	+	49.83	0.0003	boyhood
98	12	+	49.83	0.0003	embroidered
99	87	+	49.46	0.0021	answered
100	41	+	49.18	0.001	real
101	36	+	48.54	0.0009	mad
102	8	+	48.09	0.0002	hubbard
103	8	+	48.09	0.0002	kelso
104	8	+	48.09	0.0002	misshapen
105	8	+	48.09	0.0002	singleton
106	17	+	48.04	0.0004	marvellous
107	93	+	47.68	0.0023	world
108	16	+	46.69	0.0004	fantastic
109	176	+	46.02	0.0043	never
110	20	+	45.2	0.0005	theatre
111	21	+	45.09	0.0005	coat
112	11	+	44.66	0.0003	tragic
113	14	+	44.33	0.0003	fascinating
114	21	+	43.91	0.0005	james
115	9	+	43.87	0.0002	actress
116	15	+	43.4	0.0004	vulgar
117	14	+	43.2	0.0003	gilt
118	26	+	42.81	0.0006	charm
119	833	+	42.52	0.0177	had
120	7	+	42.08	0.0002	dominated
121	7	+	42.08	0.0002	romeo
122	561	+	41.63	0.0126	have
123	68	+	41.56	0.0017	anything
124	14	+	41.1	0.0003	intellectual
125	11	+	41.07	0.0003	sailor
126	9	+	40.91	0.0002	jewelled
127	337	+	40.23	0.0079	there
128	24	+	39.61	0.0006	silver
129	57	+	39.56	0.0014	lips
130	10	+	39.55	0.0002	realize

131	11	+	39.51	0.0003	spoiled
132	9	+	38.46	0.0002	hansom
133	10	+	37.79	0.0002	elaborate
134	40	+	37.51	0.001	terrible
135	13	+	36.85	0.0003	valet
136	7	+	36.15	0.0002	artistic
137	6	+	36.07	0.0001	brandon
138	6	+	36.07	0.0001	hetty
139	224	+	35.97	0.0054	some
140	15	+	35.93	0.0004	agatha
141	23	+	35.8	0.0006	passions
142	16	+	35.64	0.0004	terribly
143	59	+	35.07	0.0015	table
144	13	+	35.05	0.0003	mode
145	13	+	35.05	0.0003	type
146	10	+	34.74	0.0002	blossoms
147	9	+	34.52	0.0002	marred
148	27	+	34.28	0.0007	makes
149	16	+	33.82	0.0004	century
150	20	+	33.65	0.0005	flung
151	15	+	33.26	0.0004	stained
152	16	+	33.24	0.0004	silly
153	8	+	33.22	0.0002	bored
154	8	+	33.22	0.0002	pearls
155	18	+	33.19	0.0004	paris
156	92	+	33.14	0.0023	why
157	7	+	32.75	0.0002	artists
158	7	+	32.75	0.0002	complex
159	15	+	32.64	0.0004	acting
160	40	+	32.57	0.001	wonder
161	29	+	32.26	0.0007	sorry
162	10	+	32.17	0.0002	worshipped
163	41	+	31.76	0.001	red
164	11	+	31.25	0.0003	tedious
165	308	+	31.21	0.0072	would
166	169	+	30.87	0.0041	over
167	49	+	30.53	0.0012	tedious
168	180	+	30.47	0.0043	man
169	5	+	30.06	0.0001	dartmoor
170	5	+	30.06	0.0001	devereux
171	5	+	30.06	0.0001	divan
172	5	+	30.06	0.0001	flowerlike
173	5	+	30.06	0.0001	gorgeous
174	5	+	30.06	0.0001	grosvenor

175	5	+	30.06	0.0001	monmouth
176	33	+	30.06	0.0008	music
177	5	+	30.06	0.0001	orchids
178	5	+	30.06	0.0001	rosalind
179	5	+	30.06	0.0001	ruxton
180	5	+	30.06	0.0001	spilled
181	5	+	30.06	0.0001	tragedies
182	10	+	29.94	0.0002	exhibit
183	10	+	29.94	0.0002	frowned
184	35	+	29.75	0.0009	forget
185	22	+	29.75	0.0005	laughing
186	37	+	29.75	0.0009	youth
187	8	+	29.58	0.0002	jew
188	17	+	29.52	0.0004	muttered
189	23	+	28.66	0.0006	senses
190	17	+	28.22	0.0004	crept
191	17	+	28.22	0.0004	stirred
192	12	+	28.05	0.0003	worship
193	19	+	27.92	0.0005	glanced
194	13	+	27.88	0.0003	alter
195	13	+	27.88	0.0003	stage
196	9	+	27.69	0.0002	shrugged
197	17	+	27.39	0.0004	huge
198	12	+	27.38	0.0003	ideal
199	6	+	27.27	0.0001	mar
200	15	+	27.26	0.0004	absurd
201	15	+	27.26	0.0004	purple
202	33	+	27.2	0.0008	joy
203	11	+	27.07	0.0003	annoyed
204	24	+	26.96	0.0006	brain
205	8	+	26.72	0.0002	listlessly
206	13	+	26.72	0.0003	opera
207	9	+	26.64	0.0002	corruption
208	9	+	26.64	0.0002	fascinated
209	176	+	26.5	0.0042	know
210	74	+	26.45	0.0018	men
211	22	+	26.11	0.0005	yellow
212	37	+	26.01	0.0009	bad
213	169	+	25.83	0.0041	has
214	8	+	25.5	0.0002	nineteenth
215	8	+	25.5	0.0002	wrinkled
216	36	+	25.48	0.0009	boy
217	10	+	25.44	0.0002	george
218	14	+	25.44	0.0003	modern

219	30	+	25.16	0.0007	pain
220	15	+	25.02	0.0004	silk
221	70	+	24.96	0.0017	looking
222	6	+	24.91	0.0001	winced
223	16	+	24.8	0.0004	horrid
224	5	+	24.75	0.0001	grandson
225	119	+	24.51	0.0029	seemed
226	11	+	24.35	0.0003	curiously
227	11	+	24.35	0.0003	shaped
228	127	+	24.06	0.0031	think
229	4	+	24.04	0.0001	australia
230	4	+	24.04	0.0001	berwick
231	4	+	24.04	0.0001	burgundy
232	4	+	24.04	0.0001	chopin
233	4	+	24.04	0.0001	crude
234	4	+	24.04	0.0001	dominate
235	4	+	24.04	0.0001	fermor
236	4	+	24.04	0.0001	ferrol
237	4	+	24.04	0.0001	gautier
238	4	+	24.04	0.0001	imogen
239	4	+	24.04	0.0001	isaacs
240	4	+	24.04	0.0001	lacquer
241	4	+	24.04	0.0001	mantelshelf
242	4	+	24.04	0.0001	masterpiece
243	4	+	24.04	0.0001	patti
244	4	+	24.04	0.0001	realism
245	4	+	24.04	0.0001	sur
246	4	+	24.04	0.0001	treadley
247	4	+	24.04	0.0001	ulster
248	45	+	23.99	0.0011	pleasure
249	20	+	23.89	0.0005	smiling
250	9	+	23.87	0.0002	francis
251	9	+	23.87	0.0002	thomas
252	133	+	23.53	0.0032	good
253	44	+	23.44	0.0011	beautiful
254	43	+	23.43	0.0011	talk
255	16	+	23.32	0.0004	killed
256	12	+	23.29	0.0003	store
257	24	+	23.27	0.0006	lives
258	10	+	23.26	0.0002	intellect
259	6	+	23.01	0.0001	americans
260	6	+	23.01	0.0001	olive
261	85	+	22.95	0.0021	yes
262	28	+	22.44	0.0007	shook

263	9	+	22.29	0.0002	shrill
264	28	+	22.25	0.0007	moments
265	7	+	22.16	0.0002	ugliness
266	5	+	21.88	0.0001	immoral
267	5	+	21.88	0.0001	petals
268	5	+	21.88	0.0001	thicket
269	25	+	21.57	0.0006	fingers
270	8	+	21.5	0.0002	mock
271	11	+	21.48	0.0003	roses
272	11	+	21.48	0.0003	scientific
273	11	+	21.48	0.0003	wrought
274	6	+	21.42	0.0001	figured
275	6	+	21.42	0.0001	pose
276	6	+	21.42	0.0001	theories
277	7	+	21.06	0.0002	shakespeare
278	11	+	20.97	0.0003	dine
279	32	+	20.92	0.0008	won
280	12	+	20.89	0.0003	sunlight
281	9	+	20.87	0.0002	fashionable
282	14	+	20.86	0.0003	romantic
283	10	+	20.77	0.0002	dreadfully
284	20	+	20.48	0.0005	bye

Keywords in *PDG* compared to *ContempLit*

Rank	Freq.		Keyness (LL4)	Effect (DICE)	Keyword
1	414	+	3066.66	0.0102	dorian
2	158	+	1175.22	0.0039	basil
3	199	+	982.37	0.0049	gray
4	236	+	856.65	0.0058	henry
5	178	+	833.93	0.0044	harry
6	91	+	676.79	0.0023	hallward
7	248	+	662.61	0.0061	lord
8	85	+	621.32	0.0021	sibyl
9	1445	+	454.32	0.0255	you
10	58	+	401.87	0.0014	vane
11	907	+	335.07	0.0181	is
12	231	+	247	0.0055	life
13	33	+	245.41	0.0008	alan
14	1541	+	233.62	0.0251	he
15	52	+	230.29	0.0013	portrait
16	28	+	208.22	0.0007	campbell

17	40	+	201.71	0.001	painter
18	69	+	184.25	0.0017	picture
19	232	+	180.7	0.0055	am
20	44	+	171.52	0.0011	duchess
21	255	+	171.24	0.006	don
22	48	+	167.66	0.0012	horrible
23	58	+	167.1	0.0014	wonderful
24	22	+	163.6	0.0005	narborough
25	103	+	146.18	0.0025	really
26	77	+	144.78	0.0019	art
27	373	+	137.61	0.0084	are
28	1343	+	125.33	0.0222	it
29	52	+	122.55	0.0013	lad
30	661	+	121.07	0.0135	him
31	1361	+	119.73	0.0223	that
32	32	+	118.91	0.0008	hideous
33	29	+	115.12	0.0007	artist
34	43	+	114.66	0.0011	charming
35	48	+	112.79	0.0012	curious
36	15	+	111.55	0.0004	agatha
37	42	+	109.64	0.001	murmured
38	39	+	104.15	0.001	absolutely
39	43	+	98.49	0.0011	dreadful
40	87	+	98.44	0.0021	answered
41	18	+	97.85	0.0004	studio
42	13	+	96.67	0.0003	erskine
43	128	+	95.64	0.0031	things
44	523	+	94.37	0.0111	me
45	41	+	93.94	0.001	simply
46	105	+	93.76	0.0026	want
47	126	+	92.14	0.003	something
48	12	+	89.24	0.0003	selby
49	436	+	89.04	0.0095	one
50	23	+	85.46	0.0006	canvas
51	30	+	85.24	0.0007	coloured
52	120	+	83.54	0.0029	thing
53	32	+	82.18	0.0008	painted
54	46	+	81.93	0.0011	passion
55	24	+	81.54	0.0006	romance
56	1694	+	81.38	0.0248	i
57	399	+	80.93	0.0088	what
58	61	+	80.09	0.0015	across
59	27	+	78.8	0.0007	sins
60	40	+	78.31	0.001	terrible

61	12	+	77.85	0.0003	geoffrey
62	25	+	77.66	0.0006	monstrous
63	55	+	77.39	0.0014	beauty
64	57	+	75.2	0.0014	lips
65	15	+	75.04	0.0004	cigarette
66	10	+	74.36	0.0002	adrian
67	10	+	74.36	0.0002	gladys
68	26	+	73.87	0.0006	exquisite
69	18	+	72.72	0.0004	screen
70	61	+	72.71	0.0015	afraid
71	68	+	69.76	0.0017	soul
72	418	+	69.65	0.0091	t
73	39	+	69.44	0.001	merely
74	69	+	69.23	0.0017	everything
75	49	+	68.3	0.0012	gold
76	66	+	67.69	0.0016	women
77	9	+	66.93	0.0002	wotton
78	18	+	64.68	0.0004	horribly
79	833	+	63.8	0.0156	had
80	23	+	63.36	0.0006	passions
81	13	+	62.65	0.0003	juliet
82	14	+	62.26	0.0003	fascinating
83	20	+	62.23	0.0005	theatre
84	561	+	61.65	0.0116	have
85	16	+	60.22	0.0004	personality
86	36	+	59.52	0.0009	mad
87	8	+	59.49	0.0002	hubbard
88	8	+	59.49	0.0002	kelso
89	8	+	59.49	0.0002	singleton
90	26	+	59.16	0.0006	charm
91	19	+	59.15	0.0005	subtle
92	110	+	57.68	0.0027	always
93	16	+	57.25	0.0004	shallow
94	112	+	56.79	0.0027	love
95	28	+	55.46	0.0007	sin
96	29	+	54.95	0.0007	delightful
97	17	+	54.07	0.0004	tragedy
98	18	+	54	0.0004	painting
99	16	+	53.94	0.0004	terribly
100	23	+	52.97	0.0006	senses
101	37	+	52.1	0.0009	youth
102	107	+	51.49	0.0026	quite
103	28	+	51.08	0.0007	moments
104	41	+	50.98	0.001	secret

105	119	+	50.75	0.0029	seemed
106	129	+	49.69	0.0031	xa
107	15	+	49.65	0.0004	stained
108	29	+	49.11	0.0007	terror
109	95	+	48.04	0.0023	cried
110	35	+	47.4	0.0009	forget
111	13	+	46.13	0.0003	valet
112	7	+	46.07	0.0002	dominated
113	33	+	45.77	0.0008	joy
114	17	+	45.5	0.0004	marvellous
115	20	+	44.58	0.0005	club
116	16	+	44.57	0.0004	fantastic
117	93	+	44.04	0.0023	world
118	41	+	43.71	0.001	real
119	10	+	42.45	0.0002	blossoms
120	9	+	42.42	0.0002	hansom
121	9	+	42.42	0.0002	jewelled
122	8	+	42.41	0.0002	misshapen
123	12	+	41.74	0.0003	boyhood
124	9	+	41.13	0.0002	corruption
125	17	+	40.99	0.0004	stirred
126	9	+	39.94	0.0002	marred
127	24	+	38.75	0.0006	brain
128	223	+	38.27	0.0051	about
129	41	+	38.26	0.001	laughed
130	9	+	37.8	0.0002	actress
131	23	+	37.39	0.0006	conscious
132	5	+	37.18	0.0001	dartmoor
133	5	+	37.18	0.0001	flowerlike
134	5	+	37.18	0.0001	orchids
135	5	+	37.18	0.0001	ruxton
136	10	+	37.16	0.0002	exhibit
137	18	+	37.14	0.0004	paris
138	101	+	36.99	0.0024	people
139	6	+	35.72	0.0001	hetty
140	24	+	35.58	0.0006	silver
141	40	+	35.32	0.001	play
142	19	+	34.86	0.0005	extraordinary
143	20	+	34.61	0.0005	flung
144	10	+	34.35	0.0002	frowned
145	15	+	34.13	0.0004	acting
146	12	+	34.09	0.0003	nowadays
147	16	+	34.01	0.0004	horrid
148	14	+	33.84	0.0003	gilt

149	13	+	33.78	0.0003	alter
150	45	+	33.73	0.0011	pleasure
151	30	+	33.68	0.0007	pain
152	11	+	33.42	0.0003	tragic
153	33	+	33.2	0.0008	music
154	14	+	32.82	0.0003	romantic
155	25	+	32.7	0.0006	fingers
156	224	+	32.63	0.0052	some
157	19	+	32.43	0.0005	prince
158	44	+	32.11	0.0011	beautiful
159	16	+	32.1	0.0004	century
160	12	+	31.95	0.0003	sunlight
161	7	+	31.72	0.0002	romeo
162	6	+	31.35	0.0001	cigarettes
163	8	+	31.34	0.0002	awfully
164	23	+	31.19	0.0006	influence
165	19	+	30.8	0.0005	glanced
166	9	+	30.66	0.0002	fascinated
167	11	+	30.58	0.0003	tedious
168	46	+	30.44	0.0011	strange
169	59	+	30.12	0.0014	table
170	213	+	30.04	0.0049	like
171	47	+	30	0.0012	suddenly
172	17	+	29.79	0.0004	crept
173	4	+	29.75	0.0001	berwick
174	4	+	29.75	0.0001	chopin
175	4	+	29.75	0.0001	dominate
176	4	+	29.75	0.0001	fermor
177	4	+	29.75	0.0001	ferrol
178	4	+	29.75	0.0001	gautier
179	4	+	29.75	0.0001	lacquer
180	4	+	29.75	0.0001	patti
181	4	+	29.75	0.0001	sitter
182	4	+	29.75	0.0001	treadley
183	996	+	29.71	0.0173	his
184	40	+	29.68	0.001	wonder
185	12	+	29.66	0.0003	embroidered
186	60	+	29.63	0.0015	passed
187	13	+	29.58	0.0003	type
188	21	+	29.41	0.0005	horror
189	10	+	29.34	0.0002	realize
190	8	+	28.98	0.0002	listlessly
191	87	+	28.96	0.0021	going
192	13	+	28.96	0.0003	opera

193	9	+	28.86	0.0002	shrugged
194	40	+	28.51	0.001	met
195	8	+	28.28	0.0002	pallid
196	13	+	28.06	0.0003	mode
197	24	+	27.78	0.0006	lives
198	22	+	27.69	0.0005	yellow
199	7	+	27.64	0.0002	odour
200	10	+	27.53	0.0002	worshipped
201	20	+	27.38	0.0005	bye
202	9	+	27.25	0.0002	francis
203	15	+	27.17	0.0004	shadows
204	5	+	26.74	0.0001	devereux
205	15	+	26.72	0.0004	absurd
206	10	+	26.69	0.0002	cruelty
207	107	+	26.6	0.0026	face
208	10	+	26.29	0.0002	elaborate
209	18	+	26.29	0.0004	scarlet
210	176	+	26.2	0.0041	never
211	18	+	25.95	0.0004	key
212	8	+	25.78	0.0002	fascination
213	8	+	25.78	0.0002	nineteenth
214	15	+	25.64	0.0004	purple
215	12	+	25.51	0.0003	ideal
216	7	+	25.27	0.0002	artists
217	7	+	25.27	0.0002	complex
218	7	+	25.27	0.0002	ugliness
219	5	+	25.01	0.0001	rosalind
220	5	+	25.01	0.0001	strangled
221	6	+	24.87	0.0001	brandon
222	6	+	24.87	0.0001	edition
223	4	+	24.79	0.0001	chapman
224	4	+	24.79	0.0001	imogen
225	4	+	24.79	0.0001	isaacs
226	4	+	24.79	0.0001	sur
227	4	+	24.79	0.0001	terrifies
228	11	+	24.72	0.0003	curiously
229	28	+	24.57	0.0007	shook
230	21	+	24.37	0.0005	colour
231	169	+	24.12	0.004	has
232	15	+	24.02	0.0004	vulgar
233	85	+	23.84	0.0021	yes
234	11	+	23.81	0.0003	annoyed
235	41	+	23.8	0.001	red
236	36	+	23.53	0.0009	certainly

237	13	+	23.22	0.0003	extremely
238	8	+	23.19	0.0002	bored
239	109	+	23.17	0.0026	eyes
240	68	+	23.16	0.0017	anything
241	176	+	23.1	0.0041	know
242	6	+	23.08	0.0001	mirrors
243	21	+	22.95	0.0005	remembered
244	11	+	22.68	0.0003	shaped
245	12	+	22.63	0.0003	mist
246	169	+	22.47	0.004	over
247	77	+	22.45	0.0019	told
248	11	+	22.41	0.0003	spoiled
249	5	+	22.32	0.0001	grosvenor
250	5	+	22.32	0.0001	monmouth
251	5	+	22.32	0.0001	tragedies
252	3	+	22.31	0.0001	ashton
253	3	+	22.31	0.0001	cupids
254	3	+	22.31	0.0001	pomegranates
255	3	+	22.31	0.0001	psychologists
256	3	+	22.31	0.0001	radley
257	3	+	22.31	0.0001	temperaments
258	3	+	22.31	0.0001	wainscoting
259	6	+	22.29	0.0001	loathsome
260	6	+	22.29	0.0001	venetian
261	20	+	22.24	0.0005	leaves
262	4	+	22.21	0.0001	tapestries
263	58	+	22.07	0.0014	white
264	10	+	21.84	0.0002	dreadfully
265	15	+	21.82	0.0004	silk
266	76	+	21.8	0.0018	felt
267	98	+	21.79	0.0024	night
268	6	+	21.56	0.0001	curves
269	6	+	21.56	0.0001	languidly
270	102	+	21.54	0.0024	room
271	13	+	21.52	0.0003	strangely
272	12	+	21.47	0.0003	store
273	5	+	21.22	0.0001	divan
274	154	+	21	0.0036	must
275	27	+	20.79	0.0007	makes
276	37	+	20.67	0.0009	past
277	16	+	20.63	0.0004	silly
278	7	+	20.59	0.0002	grotesque
279	7	+	20.59	0.0002	victor
280	29	+	20.56	0.0007	fancy

281	4	+	20.33	0.0001	burgundy
282	4	+	20.33	0.0001	carbuncles
283	4	+	20.33	0.0001	ethical
284	4	+	20.33	0.0001	mantelshelf
285	4	+	20.33	0.0001	psychology
286	8	+	20.26	0.0002	wrinkled
287	5	+	20.25	0.0001	immoral
288	5	+	20.25	0.0001	psychological
289	6	+	20.23	0.0001	perfume
290	6	+	20.23	0.0001	winced
291	9	+	20.17	0.0002	mask
292	10	+	20.16	0.0002	murdered
293	7	+	20.12	0.0002	sensations