Title A computer assisted analysis of literary text: from feature analysis to judgements of literary merit

Name Tess M. E. A. Crosbie

A COMPUTER ASSISTED ANALYSIS OF LITERARY TEXT: FROM FEATURE ANALYSIS TO JUDGEMENTS OF LITERARY MERIT

Tess M. E. A. Crosbie

A thesis submitted to the University of Bedfordshire in fulfilment of the requirements for the degree of Doctor of Philosophy

University of Bedfordshire

November 2016

Abstract

Using some of the tools developed mainly for authorship authentication, this study develops a toolbox of techniques towards enabling computers to detect aesthetic qualities in literature. The literature review suggests that the style markers that indicate a particular author may be adapted to show literary style that constitutes a "good" book. An initial experiment was carried out to see to what extent the computer can identify specific literary features both before and after undergoing a "corruption" of text by translating and re-translating the texts. Preliminary results were encouraging, with up to 90 per cent of the literary features being identified, suggesting that literary characteristics are robust and quantifiable.

An investigation is carried out into current and historic literary criticism to determine how the texts can be classified as "good literature". Focus groups, interviews and surveys are used to pinpoint the elements of literariness as experienced by human readers that identify a text as "good". Initially identified by human experts, these elements are confirmed by the reading public.

Using Classics as a genre, 100 mainly fiction texts are taken from the Gutenberg Project and ranked according to download counts from the Gutenberg website, an indicator of literary merit (Ashok et al., 2013). The texts are equally divided into five grades: four according to the download rankings and one of non-fiction texts. From these, factor analysis and mean averages determine the metrics that determine the literary quality.

The metrics are qualified by a model named CoBAALT (computer-based aesthetic analysis of literary texts). CoBAALT assesses texts by Jane Austen and D. H. Lawrence and determines the degree to which they conform to the metrics for literary quality; the results demonstrate conformity with peer-reviewed literary criticism.

Contents

List of Figures

List of Tables

Acronyms

AvSentLen average sentence length.

BOW bag of words.

CoBAALT computer-based aesthetic analysis of literary texts.

LexDiv lexical diversity.

NLP natural language processing.

NLTK natural language toolkit.

PCFG probabilistic context-free grammar.

POS parts of speech.

RelEnt relative entropy.

Penn Treebank tags

Tag	Description	Example
CC	Coordinating conjunction	and, but, either
CD	Cardinal number	5, 0.5, 1955, nineteen fifty-five
DT	Determiner	the, all, this, some
EX	Existential 'there'	there is a place…
IN	Proposition or subordinating conjunction	in, by, until
JJ	Adjective	hard, old, fifth
JJR	Comparative adjective	harder, cheaper, nicer
JJS	Superlative adjective	hardest, cheapest, nicest
MD	Modal	can, cannot, should, will
NN	Noun (singular, common or mass)	girl, computer, thing
NNP	Noun (proper, singular)	England, NFL, Crosbie
NNPS	Noun (proper, plural)	Americans, Crosbies
NNS	Noun (common, plural)	postgrads, girls, computers
PDT	Pre-determiner	all, many, this
POS	Possessive ending	's
PRP	Personal pronoun	her, us, them
PRP$	Possessive pronoun	her, ours, theirs
RB	Adverb	quickly, barely
RBR	Comparative adverb	further, louder
RBS	Superlative adverb	fastest, most
TO	'to' as preposition or infinitive marker	used to, to split
VB	Verb (base form)	go, smile
VBD	Verb (past tense)	went, swam
VBG	Verb (present participle or gerund)	going, aching
VBN	Verb (past participle)	languished, flourished
VBP	Verb (present tense, not third-person singular)	sort, tend, tease
VBZ	Verb (present tense, third-person singular)	sorts, tends, teases
WDT	Wh-determiner	what, which, that
WP	Wh-pronoun	that, which, who
WP$	Possessive wh-pronoun	whose
WRB	Wh-adverb	how, why, where

Literary terms used

Term	Description	Example or observation
Adjectival phrase	A phrase with a descriptive head word	*Easy* to please
Adjective	An attribute of a noun	A *red* rose
Adverbial phrase	A phrase that modifies a verb	He sat *in silence*
Adverb	An attribute of a verb	She walked *quickly*
Alliteration	The same letter or sound at the beginning of a series of words	Peter Piper picked a peck of pickled peppers
Anaphora	Device of repetition of the first part of a sentence	"It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair." *A Tale of Two Cities* by Charles Dickens
Article	Noun modifier that indicates whether it is definite or indefinite	It's *the* policeman (definite) It's *a* policeman (indefinite)
Assonance	Internal rhyme	H*ow* n*ow*, br*ow*n c*ow*?
Auxiliary	A verb that adds functional meaning to its clause	*Do* you take sugar?
Comparative	Used to compare and usually denoted by ending "-er"	This one is *nicer*
Conjunction	Used to connect clauses	"*and", "if", "but*"
Entropic	Highly unpredictable and therefore has high information value	"There is a…..outside my window". If the missing word was *tree*, the sentence would have low entropy, if it is *dragon* it would be high
Epistolary	Writing in the form of letters	*Dracula* by Bram Stoker
Epistrophe	Repetition of a word at the end of successive sentences	"Who is here so base that would be a bondman? If any, speak; for him have I offended. Who is here so rude that would not be a Roman? If any, speak; for him have I offended. Who is here so vile that will not love his country? If any, speak; for him have I offended…." *Julius Caesar* by William Shakespeare
Free indirect discourse	Direct insight into the mind of a character	"But Lucrezia herself could not help looking at the motor car and the tree pattern on the blinds. Was it the Queen in there – the Queen going shopping?" *Mrs Dalloway* by Virginia Woolf
Function word	Words with little lexical meaning by themselves but which	"*Most* people *with* low self-esteem *have* earned *it*."

	contribute to the structure of the sentence	(George Carlin)
Gerund	Verb form that ends with -ing	"*asking", "doing*"
Imagery	Descriptive or figurative words to describe something	"The Assyrian came down like the wolf on the fold". *The Destruction of Sannacherib* by Lord Byron
Juxtaposition	Placing concepts together to emphasise the contrast between them	"Better late than never"
Lemma	The dictionary form of a word	To *run*, you ran, s/he is running
Lexical diversity	The ratio between the total number of words and the number of different types	A high lexical diversity is indicative of a more complex text
Lexical field	Words that can be grouped together	*"father", "uncle", "daughter"*
Metaphor	A figure of speech for comparison	"All the world's a stage" *As You Like It* by William Shakespeare
Mimesis	Reflection of reality from the protagonist's perspective	A play about a war horse is a mimesis of events in WWI
Noun	A thing, whether a person, concept or place	Put the *bowl* on the *table*, *Joe*.
Onomatopoeia	A word that sounds like what it represents	*"hissing", "bang"*
Part of speech	Word classes	Nouns, adjectives, exclamations
Particle	Has grammatical function but as part of a clause	*To* run, rule it *out*
Pre-determiner	Qualifying words that modify nouns	*Such* a nice person, *Quite* a good day
Preposition	Locative or chronological words	*On* the floor, *by* winter
Pronoun	A name or substitute for the noun	*Tess, it, him*
Subordinating conjunction	Joins two clauses	I looked under the chair *where* the cat often hides
Superlative	The upper or lower ends of what is being qualified; usually ends in -est	The *strongest*, the *weakest*
Symploce	Repetition of a phrase both at the beginning and end of the sentence	"The yellow fog that rubs its back upon the window-panes, The yellow smoke that rubs its muzzle on the window-panes" *The Love Song of J. Alfred Prufrock* by T. S. Eliot
Theme	The central tenet of the text	What it is about
Token	An individual word	The girl climbed the tree = 5 tokens
Type	The number of different words	The girl climbed the tree = 4 types
Verb	A word that shows action	"I *wandered* lonely as a cloud". *Daffodils* by William Wordsworth

Publications

The following publications were produced as a result of the research in this thesis.

1. Crosbie, T., French, T. and Conrad, M. (2013b) 'Towards a model for replicating aesthetic literary appreciation' in 'Proceedings of the Fifth Workshop on Semantic Web Information Management (SWIM '13)', New York, ACM, p. 8

2. Crosbie, T., French, T. and Conrad, M. (2013a), 'Stylistic analysis using machine translation as a tool', International Journal for Infonomics (IJI) Special Issue 1(1).

3. Crosbie, T., French, T. and Conrad, M. (2012), 'How far can automatic translation engines be used as a tool for stylistic analysis?' in 'International Conference on Information Society (i-Society '12), IEEE, pp. 503-507.

Acknowledgements

My thanks to my supervisors, Tim French, Marc Conrad and Ingo Frommholz, for their help and guidance over the years and the miles and to all those at the University of Bedfordshire who supported my studies. Deep gratitude is expressed to the many who completed my surveys and allowed me to gate-crash their book group meetings. Special thanks to Andy and Evelina for helping me to unpick the mysteries of Statistics and to Helen for her guidance on all things Literary. Thanks to all my friends and family for putting up with me when things did not go so well, indulging me when I thought I was in line for the next Nobel Prize and for looking interested when I waxed lyrical about my research. Finally, I thank Andrew for his patience and understanding, his financial and emotional support and for still being the only person who can make me laugh when I am in a bad mood.

Chapter 1

Introduction

The rise of Digital Humanities is evidenced by the increase in specialist journals (*Digital Humanities Quarterly*, *Digital Humanities Now*, *Journal of Cultural Analytics* and the recently renamed *Digital Scholarship in the Humanities*, previously known as *Literary and Linguistic Computing*) and specific courses being developed by universities (UCL, Princeton, The Open University, the University of Nebraska to name but a few). According to Hammond et al. (2013), Computing and English Literature are no longer seen as incompatible areas, although they are still generally contained in separate faculties with one dealing in objective calculation and the other in subjective ambiguities. Meanwhile, advances in machine-learning have allowed a computer posing as a 13-year-old Ukrainian to pass the Turing test with 10 out of the 30 judges (Sample and Hern, 2014) and subjectivity in sentiment analysis remains an active research area (Balahur et al., 2014; Aydoğan and Akcayol, 2016; Cambria et al., 2013; Cambria, 2016). Initiatives such as PAN[1] promote authorship identification, plagiarism detection and misuse of social software detection evaluations. In short, the worlds of computing and traditional humanities are integrating, to the benefit of both disciplines (Hammond et al., 2013).

Recent researches to find computers that can write fiction have concentrated on their ability to create, with mixed results (Nield, 2016; Barrie, 2014; Hudson, 2012). Since 2013, a National Novel Generation Month (NaNoGenMo) competition has been run to examine the output of such computer-generated fiction to create a 50,000 word novel. So far, the offerings have ranged

[1] http://pan.webis.de/index.html

from copying an existing book to simply repeating the same word 50,000 times although there have been some genuine attempts to create literature (NaNoGenMo, 2016). This is just one example of moving towards computational creativity; in a recent review of Franco Moretti's *Distant Reading*, Ross (2014) observes that digital humanities are 'at a rhetorical and institutional crossroads', describing the melding of very different scholastic approaches between the quantitative and the qualitative. However, without understanding what to aim for, the computer cannot create something that appeals to human readers. The focus of this thesis is the identification of what makes literature "good" and how a computer can qualify it.

In order to achieve this, authorship attribution is investigated to see if there are tools used in this discipline that can be adapted. Authorship identification makes use of stylistic features that are used by writers, often unconsciously, that can be used to create a style "map"; using statistics or machine learning, these traits can be compared to determine the likelihood of authorship being a particular writer (Mosteller and Wallace, 1963; Forsyth and Holmes, 1996; Burrows, 2002; Stuart et al., 2013*a*; Ramezani et al., 2013; Hurtado et al., 2014). This thesis makes use of the tools used to identify the stylistic features but instead of comparing them with specific texts, uses them to identify combinations of features that characterise literary merit.

Following a literature review into authorship identification tools and current work on literary criteria analysis, the thesis investigates the features that constitute "good" literature using surveys, focus groups and interviews with experts in literature and the general reading public. A pilot study is carried out to determine how robust literary features are when subjected to computational analysis and the feasibility of the study is examined. Once key literary features are identified, experiments are carried out to extract the relevant features from freely available, out of copyright literary texts of varying quality, and non-fiction. Factor analysis is used to determine the parts of speech (POS) and other literary criteria most relevant to determining the metrics. Using this framework a model is created, named **CoBAALT** (**co**mputer-**b**ased **a**esthetic **a**nalysis of **l**iterary **t**exts), which is tested on classic works of English Literature. The results are tested on the works of two authors and compared to the findings of an expert literary panel and established, published literary criticism.

1.1 Aim and objectives

The hypothesis is that it is possible for a computer to determine the literary merit of a text using authorship attribution tools. The aim of the thesis is to explore the features that constitute "good" literature and to extract these in order to build an analytical model that can assess and calculate the degree of literary merit of a given text.

To accomplish this aim, the focus is on the following research objectives:

1. Understand the limitations of computers in interpreting text. This is achieved by a literature review and by testing and analysing the degree of robustness of literary texts (Chapters 2 and 4, respectively).
2. Develop a metric to measure aesthetics as experienced by a human reader (Chapter 5).
3. Develop a framework to identify the sub-elements and inter-relationship of literature aesthetics that address the above metric (Chapter 6).
4. Develop a model to determine the aesthetic value of a text written in English, according to the above metric (Chapter 7).

1.2 Contribution

The contributions of the thesis are as follows:

- Major - the development of a definitive model for application to a given text to qualify its degree of literary merit.
- Minor - the integration of qualitative and quantitative text-analytical metrics are a contribution to knowledge and an enrichment of existing techniques in stylistic analysis.
- Minor - the literary devices that constitute "good" literature are identified and examined.
- Minor - use of the CoBAALT model provides a way to recognise non-fiction and fiction texts and categorise them accordingly.

1.3 Summary of chapters

Figure 1.1 outlines the flow of the thesis. This introductory chapter outlines the background, contributions and aim and objectives of the research while Chapter 2 presents a comprehensive literature review of existing work on authorship analysis and related works, introducing the tools used and adapted to achieve the goals of the study. Chapter 3 outlines the methodology used in the study.

Chapter 4 details the initial experimentation with selected tools on small samples of translated texts that, through a translation process, have lost some of their literary merit. This pilot study was necessary to ensure that literary features can be retained through computational analysis without manual intervention.

A brief introduction to schools of literary criticism is given in Chapter 5 along with a discussion of the fieldwork carried out to determine how humans define "literature". This field research includes using questionnaires, surveys and interviews with literary experts as well as surveys with the general reading public. Chapters 4 and 5 comprise experimental investigations into the practicality and feasibility of the research, respectively. The work in these chapters feeds the design of the eventual CoBAALT model by providing direction and explanation; these chapters may be skipped by readers who are more interested in the actual development of specific CoBAALT feature selection.

Chapter 6 explains the investigation into the POS and other literary features that were selected as strong identifiers of literary worth. Factor analysis is used to identify the variables with the greatest impact on "good" literature and these confirm the findings from the previous chapter that a stylistic analysis is computationally feasible. The eventual model is a unified framework that combines the work from the previous chapters into a model called CoBAALT that is described in Chapter 7 where the results of testing are given. Chapter 8 provides the conclusion, limitations and suggests further work.

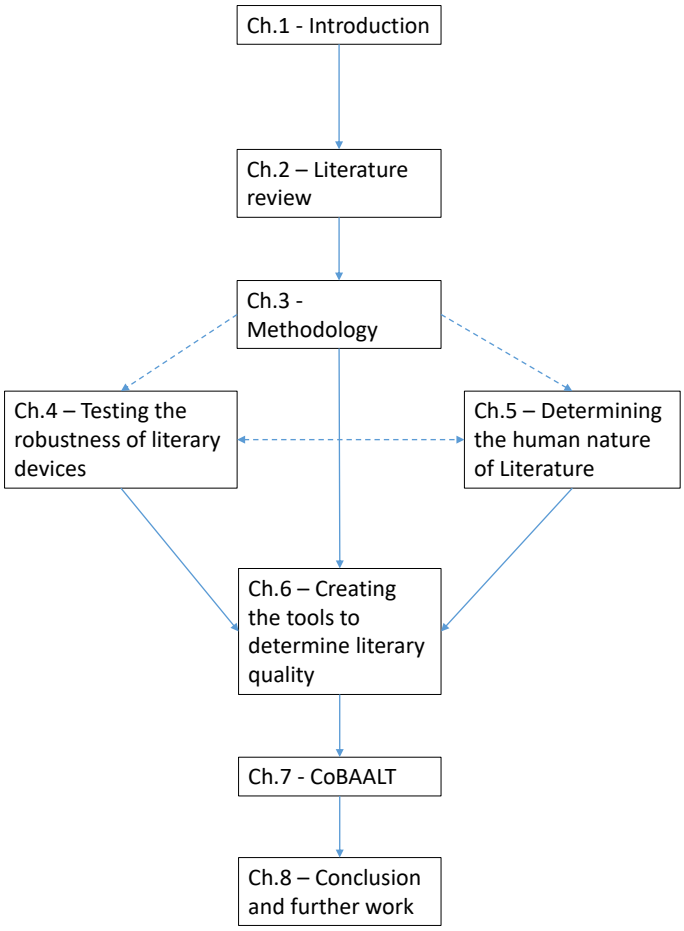


Figure 1.1: Flow diagram of chapters. The dotted line arrows indicate optional reading as the chapters indicated inform the research but do not have a direct effect on the production of the CoBAALT model.

Chapter 2

Literature Review

2.1 Analysis of text

The aim of this chapter is to give an overview of literature which is relevant to this study and which informs the tools used to achieve its objectives. Authorship attribution is the identification of a writer through their literary "fingerprint": the unconscious style they use when writing (Peng and Hengartner, 2002). Identification of that style is the first challenge and the techniques for doing so depend on the textual domain. Short texts respond differently to longer ones and attribution success often relies on the amount of training data available so a large corpus can significantly increase the chances of matching the correct author (Stamatatos, 2009). This current study does not have the advantage of multiple texts written by the same author but adapts the processes used in authorship attribution to create a style map of literary works. In this respect, function words are investigated as these are effective in creating a literary "fingerprint". Additionally, there have been some recent studies into literary style analysis and these are examined, along with studies that analyse the literary output of specific authors in greater depth. Computational tools such as lexical diversity and entropy are investigated as potential tools.

2.1.1 Authorship attribution

The first serious attempt to qualify writing analytically was made by Mosteller and Wallace (1963) in their investigation into the authorship of the Federalist Papers, a series of articles published in 1787 and 1788 concerning the ratification of the American Constitution. The three authors were known: Alexander Hamilton, John Jay (both Founding Fathers of the United States) and James Madison, a future President. What was not known was which statesman wrote which paper, a controversy which had raged since the mid-1940s and which still continues (Rudman, 2012; Savoy, 2013). Although theirs was not the first foray into the quantification of writing style (Stamatatos, 2009), Mosteller and Wallace brought a statistical approach to the debate by using Bayesian analysis on function words; words which in themselves convey little meaning but add detail to other words in a sentence. Examples of function words include articles, auxiliaries, conjunctions and pronouns. As these words are used unconsciously by a writer, they can be used to create a style map of an author and they form the identification basis for most of the researches covered here.

Sebastiani (2002) took a machine learning approach to the problem. As he quite correctly observes, the efficacy of machine learning compared to knowledge-based text categorisation is commensurate and does not require as much expert intervention, meaning that longer texts can be investigated without the expense of human labour. However, although this approach works well for simple categorisation of texts, such as for author identification, it is not suitable for the purposes of this study. Machine learning uses endogenous knowledge, restricting its information gathering solely to the texts under examination, ignoring metadata or anything else outside the confines of the text. Moreover, function words are usually removed as being superfluous to requirements whereas, in this study, they have an important role to play.

Luyckx et al. (2006) further Sebastiani's method, taking the same bag of words (BOW) approach but including more complex features, such as distributed syntactic information, and aspects related to readability in a process they define as *stylogenetics*, 'an approach to literary analysis that groups authors on the basis of its stylistic genome into family trees or closely related groups from some perspective'. The results are then clustered using a Euclidean distance-based centroid clustering technique. Included in their token-level features is a Flesch-Kincaid readability score. This is a widely used test for determining the ease of understanding a text written in En-

glish and is the 'Readability Statistics' used in Microsoft's Office packages. A high score of 90+ indicates a simple text that can be understood by a child of 11. Scores below 30 are more challenging reads, aimed at graduate-level comprehension. Luyckx et al. use the Flesch-Kincaid metric as one of several tools including POS and function word distribution to build an author profile. Their clustering results show good accuracy in gender-based and chronological predictions.

However, finding similarities in texts in order to classify them is one task. Applying qualitative judgement is another. In a study by Peng and Hengartner (2002), the authors recognised that there is 'no agreement of the unit of analysis', so it is down to the individual researcher to define how to quantify texts, whether for authorship identification or any other purpose. The search for authorship has the advantage of knowing what it is looking for; generally, there will be a set of unknown texts that can be stylistically compared to works by known authors. Forsyth and Holmes (1996) specifically tried to avoid the trap of relying on pre-existing knowledge and, more importantly, subjectivity, so that texts could be classified without recourse to huge databases. Their system also had the advantage of not being restricted to texts in English. By breaking all their testing texts into roughly 1000 byte blocks (an average of 187 words) they could provide a robust stylometric test. The system performed reasonably well, examining five different stylometric tests that gave a mean success rate of between 69.03 and 79.39 per cent. Letter frequencies were ineffective compared to other style markers and strings worked even better than word-level frequencies. These are encouraging findings for novel-length investigations.

An authorship study by Ramezani et al. (2013) investigated Persian texts and categorised their experiments, using 29 different textual features and comparing their efficacy in authorship attribution. Broadly speaking, features fall into one of three categories:

- BOW, where each token or character is taken as an element in a sequence that makes a sentence;
- syntatic and semantic which are language dependent but can reveal deeper linguistic traits;
- application-specific features which are useful in investigations into narrow applications such as online forum messages.

For their study, the authors found that specific information on the words

used was the most effective way to identify an author. This is not applicable to this thesis; however, they found that natural language processing (NLP)-based features performed well as style markers, including sentence length and verb, adjective and adverb structural information. This suggests that using combinations of POS can uncover relevant information on writing style. A similar approach used by Hurtado et al. (2014) uses a combination of 77 POS features that include punctuation as a POS.

Punctuation appears to perform well as a style marker for authorship attribution; Stuart et al. (2013*a*) found it to be the single most effective feature for identification. However, this is understandable in studies to distinguish an individual writer as any traits or quirks (such as using unusual characters, a factor the authors of the study found to be another highly useful feature) stand out as particular to that author and can consequently be used to match an unknown text to their other writings. It is unlikely to be a useful feature when creating a map of literary style as novel writers are more likely to conform to the norms of punctuation than, say, someone writing an email. The corpus used by Stuart et al. comprises academic writing so it is presumed that the texts are well-written and well-punctuated but certain identifiers, such as use of serial commas or using semi-colons where another author would put a comma, are matters of taste and cultural norms rather than indicators of literary merit.

Another work by Stuart et al. (2013*b*) extends the above paper by introducing texts written in Russian. Although this study shows that there are features common across both English and Russian, the authors specifically removed features such as function words and conjunctions. They note, however, that many of the features they combine provide diminishing returns: additional combinations do not add significantly to the accuracy. This may well be the case when identifying variables that constitute literary writing.

2.1.2 Function words

From the work done by researches into authorship attribution it is clear that some aspects may be transferable to the challenges facing this thesis. Function words consistently appear as significant markers of style. Wales (1990, p.199) defines these as 'words which have little lexical meaning, but rather grammatical meaning, and which contribute to the structure of the clause or phrase'. Because these are words that have little meaningful impact

on the text, authors use them with less attention than they do content words like nouns and adjectives, yet they are still strong indicators of style.

Furthering the authorship attribution work begun by Mosteller and Wallace, Burrows (2002) produced a method he called 'Delta' that relied heavily on relative word frequencies, seeking out differences that could indicate an author's particular style in poetry. Burrows observes that, because of their ubiquity in any piece of English, function words make up the vast majority of the 30 most frequently used. By establishing a frequency hierarchy from a pool of 25 Restoration poets, a set of norms were produced from which the degree of deviation could indicate a particular style. Longer texts were found to be easier to categorise than those under 1,500 words (Burrows, 2002) which is yet further encouragement for the use of full-length novels.

Other researchers (Mosteller and Wallace, 1963; Sarndal, 1967; Holmes, 1985) have found function words to be highly effective as style markers, mainly due to the unconscious use of them during the writing process. Peng and Hengartner (2002) used principal component analysis to investigate function word usage for a variety of authors spread across several centuries of literature, and canonical discriminant analysis was used to visualise the results. The results showed distinct clusters of style between:

- playwrights and poets (16th and 17th century);
- novelists (18th and 19th century);
- novelists (late 19th and early 20th century).

The authors observe that while function words on their own are particularly powerful as style markers, groups of indicators are even more so. This finding provides encouragement that it may be possible to find a combination of variables that form a definitive map of literary merit.

Li et al. (2006) determined some of the ways an author can be identified by their unconscious writing style, including lexical (the words they use), syntactic (punctuation and function words) and structural features (paragraph length, page layout preferences and so on). Content-specific words are also used but these are of less interest to this thesis which is more concerned with a stylistic analysis than a content analysis. From the chosen characteristics in Li et al.'s study, an accurate profile can be created to identify the writers of online messages. Specifically, the study found that it was a combination of features that contributed to the accuracy of authorship attribution.

Gamon (2004) used function words as a part of his deeper linguistic investigation to an authorship problem. By combining function word frequencies and POS with deep linguistic analysis features such as context-free grammar production frequencies and semantic graphs, authorship attribution could be improved from a "best guess" baseline of 45.8 per cent accuracy up to 97.5 per cent accuracy for context-free literature from the Brontës. Zhao and Zobel (2007) also found function words to be particularly effective as style markers.

Although authorship identification problems have helped to develop tools that can be used in stylistic analysis, it is important to appreciate that these are two very different challenges. Authorship attribution is the process of matching patterns to an author using a range of known texts. In creating a map of literary merit, however, this is not possible; effectively, there are no known texts with which to compare candidates. Another significant observation is that a writer is unlikely to be published across a wide range of genres. Context is particularly helpful in authorship attribution; in an attempt to avoid using context, Gamon's study normalised personal pronouns and names. An eighteenth century writer does not make mention of cars or computers, making attribution somewhat easier. For a stylistic problem, this contextualising is less important. This current study is more interested in the style than in simply matching likely candidates together.

2.1.3 Lexical diversity and entropy

Lexical diversity is a measurement of different words in a text formed by calculating the ratio of word types to the total number of tokens where a type is an instance of word (*the girl climbed the tree* has four types with *the* occurring twice in a sentence of five tokens). This measurement gives an indication of the richness of the text, so a high lexical diversity suggests a "better" literary experience and this measurement has been used in several studies (Savoy, 2012; Kubát and Milička, 2013; U and Thampi, 2015). Gonçalves and Gonçalves (2006) investigate Zipf's fractal power law by ascribing a lexical wealth to literary authors by calculating the ratio between the number of types (different words) and the number of tokens (total number of words) in the text. Characteristic indices can be identified for each author and for discriminating between literary and non-literary (newspaper) texts. One short-coming of this measurement is that literary writers often repeat for effect and, due to the use of function words, short texts therefore

demonstrate a higher lexical diversity than long ones due to essential type repetitions (Johansson, 2008). This may have limiting implications for its use as a tool for novel-length texts.

As an alternative approach, entropy has been identified as a potential measure of literary creativity (Kan and Gero, 2009). Low entropy indicates no unexpectedness whereas creativity is the product of the unusual and surprising, *ergo* high entropy equates to high creativity. In their study, the authors compare *The Sound of Silence* with *Twinkle, Twinkle, Little Star*, finding text entropy of 1.9 and 1.5 and relative entropy of 82 and 76, respectively, and thereby demonstrating that Simon and Garfunkel are more creative than a nursery rhyme in this example. Similar results have been achieved using texts translated from French into Chinese (Zhang et al., 2011).

The entropy approach was furthered by Haiyan and Xiaohu (2011) who quantified the novels of Scott Fitzgerald by using the power law and text entropy to determine creativity in the texts. Using the power law, the study analysed lexical measurements against text length and the authors were able to show how the types-token ratio, word repetition and word frequency entropy are effective tools to measure creativity in the novels of Scott Fitzgerald. They argue that an author's word choice determines the amount of information that can be disseminated in any given length of text, therefore the lower the correlation between word relative entropy and types-token ratio or word repetition, the more creative the work. The results were compared to the opinions of various literary critics in ordering the creative value of the four Fitzgerald novels in the study[1]. Although results were encouraging, the authors of the paper admitted that applying their rational to other novelists was not yet a viable option due to the labour-intensive nature of the task.

2.1.4 Stylistic analysis

In most of the studies cited so far, the goal has been authorship identification and finding new ways to match an unknown text to the work of a known author using a variety of machine-learning processes. Few studies have concentrated on the analysis of style alone. One exception has been the work of Keim and Oelke (2007) who created a system to visualise written work graphically. Their study observes the difficulties in analysis of literature due

[1] *This Side of Paradise*, *The Beautiful and the Damned*, *The Great Gatsby* and *Tender is the Night*

to concentration on just one aspect of the text. Therefore, their approach is to analyse text at different hierarchical levels, with at least 'one value per sentence, paragraph, chapter, or text block' and then graphically visualising the results. By using different variables for the analysis of the whole text, not only does this give a better insight into the discriminative power of each method but comparison of their effectiveness on a specific aspect of the text can indicate new methods of literary analysis. Sentence length is used as an important style indicator with more "literary" works having longer average sentence length than other works. Additionally, they stress the importance of POS and function words in analysis of quality.

Li et al. (2004) observe that although grammar and word processing have advanced in computational analytical NLP terms, semantics and, in particular, pragmatics and discourse analysis lag far behind. To rectify this, they investigated Chinese poetry for its literary language features using stylistic analysis with term connections. As an example of term connection, they use the word 'rose', breaking it down into components of semantic meaning (pronunciation and spelling), its referential semantic meaning (i.e. its dictionary meaning with genus and species) and its semantic meaning as an experience, including its literary implications and emotional impact; in this case, 'tender' and 'affection', respectively. The authors observe how Chinese poetry can be divided into eight distinct styles, according to Liu Xie; twenty-four, according to Sikong Tu; or into four with four dimensions, according to Chen Wangdao. For ease of computation, they opted to use the latter classification system and investigate the poetic styles of 'bold and unconstrained', consisting mainly of strong action words, or 'graceful and restrained' which include more gentle terms of expression.

After semantically pre-treating the poetry, the authors followed a four-step procedure of calculating the word context semantic value, the word context connotation, the poetic discourse connotation and finally classifying the poetry as either 'bold and unconstrained' or 'graceful and restrained'. Comparing their results with the opinions of 38 Chinese major seniors showed strong correlation with the computer analysis.

Poetry offers specific challenges, not least of which is that simple quantitative features fail to recognise the multi-level relationship that words can have within a poem, but other stylistic features are available to poetry analysis, including rhythm and rhyme (Kaplan and Blei, 2007). Although these are less relevant to an analysis of prose, certain aspects such as alliteration, assonance and consonance are common to both styles of writing.

Boychuck et al. (2014) use linguistic rhythm in French prose to ascertain author style. There are several proprietary and free software tools to analyse rhythm, including Alceste[2], Rhymes[3] and Tropes (cited in Boychuck et al. (2014)) for French and Russian. Their study is naturally language-specific and is based on Trope but it includes the identification of assonance, alliteration, rhyme, word repetition and coordinated words which can all contribute to a style map of the authors investigated[4].

Another avenue for stylistic analysis has been followed by Feng et al. (2012*a*) who used a probabilistic context-free grammar (PCFG) parser to identify syntactic variation among writers. A PCFG can only identify structural probabilities and it is limited by the rules that define it. As an example of this, Bird et al. (2009) quote the Groucho Marx line from the 1930 film, *Animal Crackers*, 'I shot an elephant in my pyjamas. How he got into my pyjamas, I don't know'. The joke depends on whether the prepositional phrase 'in my pyjamas' stems from the verb phrase, 'shot an elephant' or from the noun phrase, 'an elephant'. Despite the constraint, Feng et al. could successfully match syntactic patterns to specific authors and have used a similar method to detect fake hotel reviews (Feng et al., 2012*b*). This process relies, however, on having a "gold" standard with which to compare unknown texts.

2.1.5 Literary analysis and interpretation

A computational approach to literature can yield insights missed by human scholars. According to Kenny (cited in Stubbs (2005)), a stylistic interpretation must adhere to two criteria for computer-aided stylistic analysis: the computer must provide an essential component and the results must provide an 'original, scholarly contribution'. In his paper, Stubbs observes that a frequency analysis can identify the surface meaning of a novel quite easily, so *Heart of Darkness* is about a man named Kurtz and is set on a river, but underlying meanings take more unearthing. A frequency analysis of verb lemmas indicates that 'seem' and similar words that suggest uncertainty and obfuscation are common, and one of the underlying themes throughout the novel is indeed the lack of knowledge: the fog - both literal and metaphorical - and the geographic wildness that Marlow, the narrator, endures in his

[2]http://www.image-zafar.com/Logicieluk.html
[3]http://rifmovnik.ru
[4]Stendhal, Balzac, Flaubert and de Maupassant

quest. Even the structure of the novel can be interpreted by the computer by identifying those words that occur at the beginning and/or end of the story, marking a circle of narrative, or those that increase towards the end, like 'dark' and 'nightmare', features that add to the sense of heightened tension. Through collocation, 'grass' is found to be associated not with green shoots of life but with death and decay, while words that usually denote sparkle, like 'glitter' and 'gleam', are harbingers of danger. Long strings of adjectives are found throughout the novel, as are words with a negative prefix. In fact, negativity is a strong theme throughout the story, particularly in regard to things that are not as expected, a feature that marries well with the 'seem' lemma. Stubbs is aware of the limitations of computer-assisted stylistic analysis but points out that it can 'document more systematically what literary critics already know...[and] reveal otherwise invisible features of long texts'. It is these two specific areas that are of the greatest interest in this thesis.

An investigation into the correspondence of Emily Dickinson (Plaisant et al., 2006) sought suppressed eroticism, automatically classifying various letters using a multinomial naïve Bayes algorithm with a D2K data mining tool into those that were erotic and those that were not. A Dickinson expert correlated the classification both for eroticism and, in a separate exercise, for spirituality. Not only did the computer assess similarly to the literary expert, it made her 'plumb much more deeply into little four- and five-letter words, the function of which I thought I was already sure, and...enabled me to expand and deepened some critical connections I've been making for the last 20 years'. Interestingly, the expert and the computer often agreed on their classifications but apparently for different reasons. It appears that some subtle, unconscious process occurs in the mind of the human reader that the computer can only state boldly.

There are some studies that investigate specific aspects of literary quality including an ongoing experiment[5] where the authors of the study (Hammond et al., 2013) invite English majors and the general public to identify changes of voice in Eliot's *The Wasteland* and compare their opinions with those of the computer. A second experiment to identify instances of free indirect discourse (FID) in Woolf's *To The Lighthouse* is also in progress[6] although, according to the authors, the algorithm used has so far not been successful in identifying the required FIDs.

Another study (Muralidharan and Hearst, 2013) has taken a novel approach

[5]http://hedothepolice.org
[6]http://brownstocking.org

by rather than merely applying computational techniques to literary problems, approached a literary question and built WordSeer to solve it, recognising that literary studies are a progressive elaboration, a 'cycle of reading, interpretation, exploration and understanding', yet many digital humanities studies stop after the first two aspects. Using word trees, the authors are able to extract relationships between words, such as isolating incidents of 'her' as a possessive rather than a third-person pronoun.

It is this issue of progressive elaboration that causes conflict between the worlds of computation and of literary criticism. Hammond et al. (2013) observe that literature is frequently *deliberately* ambiguous (my italics) whereas a computational approach sees subjectivity as a problem to be solved. Roque (2012) approaches this challenge differently by focusing on each school of literary criticism and determining the best computational approach to analyse *Finnegans Wake* given their core beliefs. Therefore, a New Criticism approach (see Section 5.1.1) may include building an artificial intelligence computational model of culture, or a Structuralist approach (see Section 5.1.2) using intelligent agents to interpret semiotics.

Jockers and Mimno (2013) investigate the themes of over 3000 19th century novels written in English and hypothesise that anonymous texts are more likely to have controversial themes (religion, politics, etc.) than those written by a named author. Here, theme is defined as 'a type of literary content that is semantically unified and recurs with some degree of frequency or regularity throughout and across a corpus' (Jockers and Mimno, 2013). They also examine whether gender can be predicted from analysis of the theme. Function words were initially removed in this study in order to avoid influencing theme but eventually only nouns were used. Although a balance of probability suggests that their system can assess the writer's gender in 80 per cent of the anonymous texts, without being able to confirm the correct identity this remains no more than a tantalising insight. Moreover, the authors stress that a computational approach is a tool to assist interpretation of text, not to act as a replacement for human interpretation.

Ashok et al. (2013) confront head-on the perceived wisdom that there are no common stylistic qualities to successful literature. Using download rates on Project Gutenberg, literary award winners and Amazon sales figures as indicators of successful books, the authors achieve a rate of 84 per cent in predicting success. Although download and sales figures do not necessarily equate to literary qualities, the authors found that the Gutenberg figures are remarkably good indicators and, by also testing some best-sellers of ques-

tionable literary merit i.e. Dan Brown's *The Lost Symbol*, this approach was found to be effective. One of the more interesting results of this study was that, contrary to perceived wisdom, readability according to a Gunning FOG/Flesch-Kinaid index is inversely proportional to the success of the novel. Complexity appears to make for a more literary work.

2.2 Summary

The literature review suggests that a stylistic analysis of literary texts is possible but there is little current work that assesses the degree to which a text meets any specific criteria. As shown in Figure 2.1, related works which

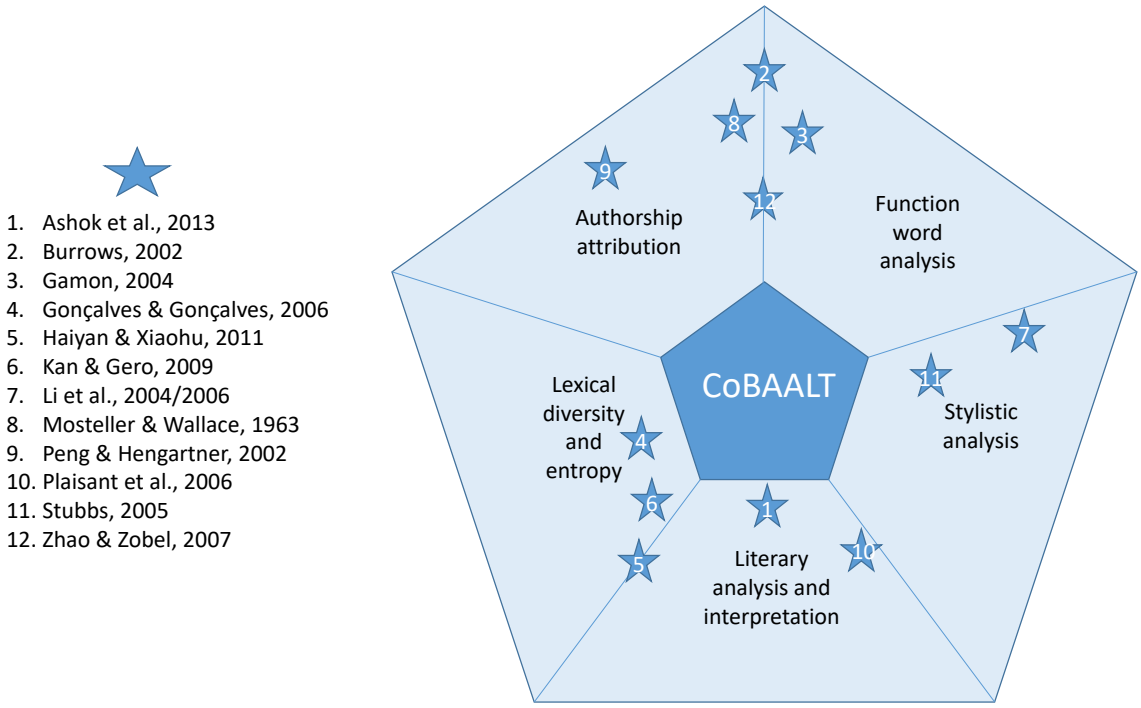


Figure 2.1: CoBAALT's origins from related work

allow the creation of a model (named CoBAALT, see Chapter 7) able to pass value judgements of literary merit are multi-disciplined. The diagram shows the background area of a dozen of the most influential works (indicated with a star) for this research. The closer the star is to the CoBAALT pentagon, the more important is the paper to the thesis. However, although these previous researches provide potential opportunities to identify literary merit,

experiments are needed to understand which tools can be used to determine these criteria along with an investigation into the features that human readers identify as important to their appreciation of a novel.

Chapter 3

Methodology

3.1 Overview

The thesis's hypothesis is that a computer can determine the degree of literary merit of a text. An inductive research approach is used that makes use of both qualitative and quantitative data. Qualitative data are obtained from human experts (people with at least a BA in English or American Literature) and the general reading public in the form of focus groups, questionnaires, surveys and face-to-face interviews and are used to uncover the way humans approach literature and form opinions on whether a work is literary. This is necessary in order to define the components that constitute "good" literature from a human viewpoint. Schools of literary criticism assume that the reader is human with all the historical, social and emotional perspective this entails but this rich background is challenged when the analysis is purely computational. Once the necessary features have been identified qualitatively, investigation can be carried out into the quantitative aspect using factor analysis to identify the features which are used to determine the components of an eventual model called CoBAALT.

3.2 Research design

Crotty (1998, pp. 2-3) observes four elements of research design (shown in Table 3.1) that need to be considered:

Table 3.1: Paradigms, methods and tools (Mackenzie and Knipe, 2006)

Paradigm	Methods (primarily)	Data collection tools (examples)
Positivist/ Post-positivist	Quantitative. "Although qualitative methods can be used within this paradigm, quantitative methods tend to be predominant . . ." (Mertens, 2005, p. 12)	Experiments, quasi-experiments, tests, scales
Interpretivist/ Constructivist	Qualitative methods predominate although quantitative methods may also be utilised.	Interviews, observations, document reviews, visual data analysis
Transformative	Qualitative methods with quantitative and mixed methods. Contextual and historical factors described, especially as they relate to oppression (Mertens, 2005, p. 9)	Diverse range of tools - particular need to avoid discrimination. E.g. sexism, racism, and homophobia.
Pragmatic	Qualitative and/or quantitative methods may be employed. Methods are matched to the specific questions and purpose of the research.	May include tools from both positivist and interpretivist paradigms. E.g. Interviews, observations and testing and experiments.

- Epistemology - the theory of knowledge adopted. Table 3.1 outlines some of the most popular paradigm options. Although a constructivist approach was initially considered, this was replaced by one of pragmatism. Constructivists literally create theory from the data collected but in this thesis a hypothesis - that a computer can be used to make judgements of literary merit - already exists.

  A pragmatic stance is taken in Chapters 4 and 5; this approach guides a practical and results-led enquiry that iteratively leads to further action and is one recommended as a way to help researchers answer their research questions (Johnson and Onwuegbuzie, 2004) by combining inductive and deductive thinking and developing new meaning through measurement and observation (Creswell, 2014). These chapters investigate the robustness of literary features and explore the way humans make qualitative decisions about their reading material, respectively. The qualitative data gathered at these stages serve as confirmation that a stylistic analysis is a suitable metric for the research's objective. However, a more positivist approach is exercised in Chapters 6

and 7 where quantitative data is used to identify the relevant variables and create the CoBAALT model. This approach can be encompassed in a pragmatic paradigm whereby the methods used, both qualitative and quantitative, are matched to the research question and include positivist tools such as observations and experiments (Mackenzie and Knipe, 2006).

- Theoretical perspective - the philosophical standpoint that guides the research. An interpretive approach is used here due to the evolving nature of the research (Section 5.2). Interpretivists understand that there is no single Truth: there are multiple interpretations of Truth and these constantly evolve. The goal is to understand rather than to predict results, producing a "hermeneutic circle" of interpretation (Hudson and Ozanne, 1988).

- Methodology - the way the methods used relate to the desired outcome. A mixed methods approach is used that aligns neatly with the pragmatic epistemology, using sequential procedures with qualitative investigation to shape the research direction followed by quantitative methods to test the theories developed (Creswell, 2014). Here, qualitative data form the basis of the research by generating categories (Sections 5.3.1, 5.3.2 and 5.3.3) that can then be investigated quantitatively.

- Methods - the way the data will be collected. Interviews, focus groups and a literature review are recommended tools (Decrop et al., 2000, p. 113) and those used in the thesis are detailed in Sections 3.2.1 and 3.2.2.

3.2.1 Qualitative data

For an investigation into computational appreciation of literature it is crucial to attempt a definition of what makes a book literary. Therefore, it was decided to hold a focus group (Section 5.3), a strategy recommended at the early stages of a study to explore preliminary findings or hypotheses (Krueger and Casey, 2009), generate new theories (Powell and Single, 1996) and guide the development of further, more detailed and specific questionnaires (Hoppe et al., 1995). Kitzinger (1995) particularly recommends the use of focus groups when the interviewer has many open-ended questions, as is clearly the case when developing a nascent hypothesis, and Goss and Leinbach (1996)

have pointed out the advantages of group discussion over individual extracted narratives. Furthermore, as observed by Morgan (1988), focus groups are a time-effective tool compared to conducting interviews. The researcher was aware that people she could recruit easily did not represent a broad spectrum of the population; however, Kitzinger (1994) recommends working with pre-existing groups as they provide a social context conducive to idea-generation, especially as interaction promulgates further discussion, and Morgan (1988) agrees that a comfortable and familiar setting can encourage participants to speak out.

The categorised results from the focus groups are then used as quantitative data for the online survey in Section 5.4.

Finally, interviews with English Literature teachers (Section 5.5) are used to understand how literary criticism is commonly taught to children in order to give a greater in-depth understanding of the tools available (Anyan, 2013). The qualitative data allows the collection of coded features which can then be used for quantitative analysis (Richards, 2009).

3.2.2 Quantitative data

The responses from the focus groups are categorised and form the basis of the questions for an online survey (Section 5.4) which is open to general readers rather than expert literary critics. The advantages of using a survey in conjunction with a focus group include the ability to amass a large quantity of empirical data at minimum cost while avoiding the potential pitfall of producing data that lack detail (Kelley et al., 2003); the result is that the data become structured (Sofaer, 1999).

The structured data form what Neuman (2013) calls a 'conceptual definition' that measures what constitutes "good" literature before forming an 'operational definition' that encapsulates the scope of the research. Quantitative measurement then converts the abstract ideas obtained through qualitative research into a single medium (i.e. numbers) that can be measured to see whether the hypothesis is supported. Here, these are shown in Tables 6.4 and 6.5.

Once coded, the features that comprise a literary text were to be further clustered using principal component analysis. However, this approach did not produce strong correlations and it was not possible to reduce the large

number of factors. Instead, Minitab's factor analysis was used to identify the variables that can be combined to create a framework to identify "good" literature (Section 6.3).

The elements of the framework are collated into a system called CoBAALT (Section 7.1) that identifies the relevant literary features and POS and determines to what degree the text can be called "literary".

3.3 Data collection

3.3.1 Pilot study

Although the literature review implied that the hypothesis was feasible, a pilot study was run (Chapter 4) to see whether individual stylistic features are sufficiently robust to identify patterns of literary merit. Through translating sections of literary text (prose and poetry) into various languages and then back into English, it was possible to compare the results and determine the extent to which the stylistic features were retained. In fact, the results showed that up to 90 per cent of the literary devices remained.

3.3.2 Focus groups

Two focus groups were held (Section 5.3), the first in December 2013, the second in June 2015. The guidelines given in the paper by Krueger (2002) were used for both groups. The researcher was conscious of possible bias in the first group as the members were all well-known to her as she was a member of this particular book group, therefore it was felt that the second event with more unfamiliar people was necessary. In both cases the groups were not recorded at the groups' request but the researcher took notes. Apart from one or two questions for clarification of a point made or to return the conversation back to the point under discussion, the researcher remained an observer. Three key areas were coded and identified: plot (see Section 5.3.1), descriptions (Section 5.3.2) and theme (Section 5.3.3).

3.3.3 Human panel of experts

A human panel of experts with at minimum a first degree in English or American Literature was recruited for Sections 6.1 and 6.2.1 and as part of the results triangulation (Sections 7.3.2 and 7.3.3). In the first two cases a non-systematic approach was used: unlike Delphi or RAND methods, there was no feedback between the participants to obtain a consensus of opinion (Campbell et al., 2002). The triangulation sections were fed back once to obtain a consensus but the results were already very similar between them.

3.3.4 Surveys

An online survey was carried out that was open to members of the general reading public (Section 5.4). The guidelines given by Kelley et al. (2003) were followed although these do not specifically cover online as a research method; the principles are the same as for a postal questionnaire. The questions were piloted by four volunteers and then made available online and the survey advertised through social media.

After generalised questions about reading habits and preferences, question 3 asks 'What do you look for in a good book?' and asks the respondents to score the features found as a result of the focus groups (Section 5.3) on a 1 to 5 Likert scale with options of 'not important', 'somewhat unimportant', 'neutral', 'somewhat important' and 'important'. Following findings from the pilot study, 'Theme' was changed to 'Learning something new' as it was felt to be a more widely understood term.

Respondents were also asked for the reasons behind their choices of their three favourite books in case there were other factors not brought up by the focus groups that should be considered. 'Gripping' and 'characters' were the most common responses but are outside the scope of this thesis.

A second survey that was open to the public was used to establish the POS of most significance to literary merit and to ensure that these could be readily identified (Section 6.1.1). A pilot study suggested that respondents would be unwilling to read two entire novels purely for the sake of a questionnaire without some financial incentive, a factor not budgeted in the work. Therefore, the human panel agreed to identify short passages within the two books that they found to be of particular literary merit and a consensus of 10 passages

was made available for the open survey.

3.3.5 Interviews

Two semi-structured interviews (Section 5.5) were carried out with English teachers from Bedfordshire schools (age range of children from seven to eighteen). The purpose of these was to identify how children are taught to appreciate literature to see whether a similar approach could be used to teach a computer. Interviews followed the guidelines set out in the article by DiCicco-Bloom and Crabtree (2006) and were conducted face-to-face.

The interviews helped to establish that of the three main teaching focus areas, structure was computationally the most feasible.

3.3.6 Feature selection

Factor analysis identified the most relevant POS and features (Section 6.3) and these were scored by determining the average score across four grades of fiction and one of non-fiction (Section 6.3.1). The further the feature is from the average, the higher or lower the score for that feature.

3.4 Summary

A pragmatic and inductively interpretive approach is used, employing mixed methods with the qualitative aspects shaping the research direction for the quantitative analysis. Qualitative data are collected through focus groups, interviews, questionnaires and surveys. These data inform the direction of the search for quantitative data but do not directly affect them. Quantitative data are collected through factor analysis, questionnaires and surveys. Results are triangulated for validation through correlation with the results of the human panel, the Gutenberg Project download counts and published literary criticism.

Chapter 4

Testing the Robustness of Literary Devices

This chapter outlines the preliminary work done to ensure that the concept of analysing literature from a stylistic aspect was feasible and serves to demonstrate the exploratory experiments carried out to determine whether this computational approach was appropriate to the research's aims. From the literature reviewed it seemed likely that a stylistic analysis was possible using a computational approach. However, with a steep learning curve in NLP ahead, it was decided to run a pilot study that would not only indicate which literary devices might be identified but would serve to produce a peer-reviewed paper to give an initiation into presenting at conferences. The literature suggests that POS were among the strongest indicators of literary quality but it was not clear how robust they are when subjected to a computational analysis. In short, to what extent would errors in automatic tagging affect the result? If the literary devices that indicate quality are easily mistaken or lost due to the ineffectiveness of the computational parser, the study would be heavily reliant on manual classification with associated time and labour costs. To determine the robustness of literary features, a pilot study was carried out using a translation tool to examine the extent to which texts could be corrupted and yet still retain specific stylistic features, an approach used by Banea et al. (2008) which had revealed interesting insights, capturing subjective text semantics effectively.

The purpose of this chapter is to present the pilot work that had two main goals:

1. to perform a preliminary exploration of language features in an accessible environment;

2. to determine whether the computational parsing would have to be manually reviewed and to what extent parsing errors would affect the literary quality of texts. Severe impacts would suggest that stylistic analysis would not be possible without manual intervention or a different machine-learning approach.

For those readers more interested in the direct development of the CoBAALT model, this preliminary work chapter may be skipped. The work that follows has been communicated in the paper presented by Crosbie et al. (2013*a*).

4.1 Translated and re-translated texts

Two texts were used as samples: one a piece of prose, one a sonnet. Due to the limitations of the online tools used, these were necessarily short texts (under 100 words). Both texts underwent a fine-grained analysis by volunteer literature graduates to identify the literary features. Each text was then subjected to a dual machine translation process, from English into 62 different languages, and then the results were translated back into English. Several free tools were considered for this task, including Yahoo's Babelfish, Bing Translator and the online version of Babylon, but Google Translate was chosen as it provided the most consistent and accurate results.

Both texts now had 63 versions: the original and 62 texts that had been translated from, and back into, English. Each translated text was compared to the original using comparison software. Several comparison tools were tested, including KDiff, WinMerge, WordCompare and the free online version of Compare Suite. The latter was chosen as it gives a graphical representation of the textual differences (Figures 4.1 and 4.2), making comparisons quick and easy; however, the free version does restrict the text length to a single paragraph. The results were ranked according to the degree of similarity with the original text, as shown in Figure 4.1 which shows the result of the Catalan re-translation, the worst-performing language in terms of similarity with the original.

Two text were compared.

Text 1: 565 byte(s), 101 word(s), 30 unique word(s)
Text 2: 506 byte(s), 88 word(s), 21 unique word(s)
Common words number: 25
Similarity (by keywords): 32,9%

Legend:

- common keyword, appears in both texts
- unique keyword, appears only in one of the texts

Text 1: 565 byte(s), 101 word(s), 30 unique word(s)

In an instant the atmosphere was transformed to Bathsheba's eyes . Beams of light caught from the low sun's rays , above, around , in front of her, well -nigh shut out earth and heaven --all emitted in the marvellous evolutions of Troy's reflecting blade , which seemed everywhere at once , and yet nowhere specially . These circling gleams were accompanied by a keen rush that was almost a whistling --also springing from all sides of her at once . In short , she was enclosed in a firmament of light , and of sharp hisses , resembling a sky -full of meteors close at hand .

Text 2: 506 byte(s), 88 word(s), 21 unique word(s)

In an instant the atmosphere was transformed in the eyes of Bathsheba . The light rays trapped rays of the sun down , over, around , in front of her, almost excluding land and sky - all the wonderful changes in the cast sheet reflecting Troy , which seemed everywhere , and none in particular . These flashes were sometimes accompanied by acute fever was almost a whistle - are flowing around it immediately . In short , he was locked in a vault of light and sharp whistles , like a sky full of meteors in hand ..

Legend:

- common keyword, appears in both texts
- unique keyword, appears only in one of the texts

Common Keywords:

Keyword	Frequency in text 1	Frequency in text 2
light	2	2
all	2	1
around	1	2
rays	1	2
sky	1	2
almost	1	2
these	1	1
accompanied	1	1
short	1	1
sharp	1	1
hand	1	1
meteors	1	1
full	1	1
everywhere	1	1
reflecting	1	1
eyes	1	1
bathsheba	1	1

Figure 4.1: Similarity between Text A original and the version translated back into English from Catalan

4.1.1 Prose: Text A

The fine-grained literary analysis of Text A is as follows.

Original text

> In an instant the atmosphere was transformed to Bathsheba's eyes. Beams of light caught from the low sun's rays, above, around, in front of her, well-nigh shut out earth and heaven–all emitted in the marvellous evolutions of Troy's reflecting blade, which seemed everywhere at once, and yet nowhere specially.

> These circling gleams were accompanied by a keen rush that was almost a whistling–also springing from all sides of her at once. In short, she was enclosed in a firmament of light, and of sharp hisses, resembling a sky-full of meteors close at hand.

Hardy, *Far From the Madding Crowd*

Features that emphasise speed and movement

The double alliteration of 'In an instant' places great emphasis on the word 'instant' so the reader is made aware of the speed of the change. The juxtaposition of 'at once. In short' reinforces the suddenness of the transformation, the more so because 'at once' is repeated in this short passage. Movement is suggested by the asyndeton of 'above, around, in front' and by the paradox of 'everywhere at once, and yet nowhere'. The word 'springing' also emphasises movement.

Features that emphasise light and sound

'Beams of light' literally and metaphorically mirrors 'sun's rays' and in their respective positions either side of the caesura, emphasises the image of light while the reference to 'earth and heaven' demonstrates the all-encompassing quality of the light. Nouns referring to the light are compounded by adjectives; 'reflecting blade', 'circling gleams' and these are added to as the sound is introduced; 'keen rush', 'sharp hisses'. The author's intention is to make this a sensuous description. The consonance of the 'st' in 'almost a whistling' coupled with the alliterative 's' of 'sharp hisses' creates an onomatopoeic effect, creating a sound for the 'meteors'.

Features that emphasise sex

Using 'the atmosphere' as the subject of the sentence accentuates its importance in the subsequent passage and reiterates Bathsheba's position as bystander. The imagery of Troy's blade is phallic, particularly as Bathsheba is behaving improperly by being alone with Troy, and the chapter title (although not included in this extract) is called 'The Hollow Amid the Fens',

suggesting a secret hiding place where a seduction could take place as well as echoing Bathsheba's sexual 'hollow'. Being 'enclosed in a firmament of light' is a reference to Genesis and the Creation, again trapping Bathsheba in a metaphor of sexuality and fertility. The simile is left to the end, 'resembling a sky-full of meteors' to leave the reader with that vivid image of sexual fireworks in his/her mind.

Re-translated texts

The texts vary in degree of similarity, although even the least similar, Catalan, is coherent. For simplicity, only the two most and two least similar texts are examined here.

Norwegian, with 83.1 per cent similarity to the original

> In an instant the atmosphere was transformed to Bathsheba's eyes. Rays of light trapped by the low sun's rays, above, around, in front of her, well almost shut out earth and heaven - all emitted in the great evolutions of Troy's reflecting blade, which seemed everywhere at once, and yet nothing in particular. These circling gleams were accompanied by a keen rush that was almost a whistling - also springing from all sides of her once. In short, she was enclosed in a firmament of light, and sharp hisses, resembling a sky full of meteors close at hand.

Latvian, with 77 per cent similarity to the original

> In the immediate atmosphere was transformed into the eyes of Bathsheba. Beams of light caught from the low sun's rays, above around in front of her, well almost removed from the earth and heaven - all emitted in the wonderful events of Troy reflecting blade, which seemed to be everywhere at once, and yet nowhere specially. These circling gleams followed by a keen rush that was almost whistling - also springing from her hand all at once. In short, she was enclosed in a firmament of light, and sharp hisses, resembling a sky full of meteors close at hand.

Latin, with 37.2 per cent similarity to the original

> In a moment the air was transformed to Bathsheba the eyes. Rays of light in front of him almost to the exclusion of taking a low rays of the sun between the earth and the sky above - all reflecting the emission of miracles in the course of Troy, the grass, which seemed everywhere at once, and yet never properly. These are the embrace he rushed shine with the keen hissing was near - and at the same time from all sides thereof. Finally that closed the firmament of light, and hisses like a sharp, the air full of meteors close at hand.

Catalan, with 32.9 per cent similarity to the original

> In an instant the atmosphere was transformed in the eyes of Bathsheba. The light rays trapped rays of the sun down, over, around, in front of her, almost excluding land and sky - all the wonderful changes in the cast sheet reflecting Troy, which seemed everywhere, and none in particular. These flashes were sometimes accompanied by acute fever was almost a whistle - are flowing around it immediately. In short, he was locked in a vault of light and sharp whistles, like a sky full of meteors in hand.

Surviving literary features

Table 4.1 shows the degree to which the literary features survive the re-translation process. It is clear that there is considerable difference between the texts, with the more similar texts retaining a high proportion of literary features. However, many features do survive, even if only in a modified form (marked as 'partial' or 'implied').

Table 4.1: Feature analysis of re-translated versions of Text A

Features	Present in Norwegian version 83.1%	Present in Latvian version 77%	Present in Latin version 37.2%	Present in Catalan version 32.9%
Alliteration 'In an instant'	Yes	No	No	Yes
Juxtaposition of 'at once. In short'	Yes	Yes	No	Partial
Repetition of 'at once'	Yes	Yes	No	No

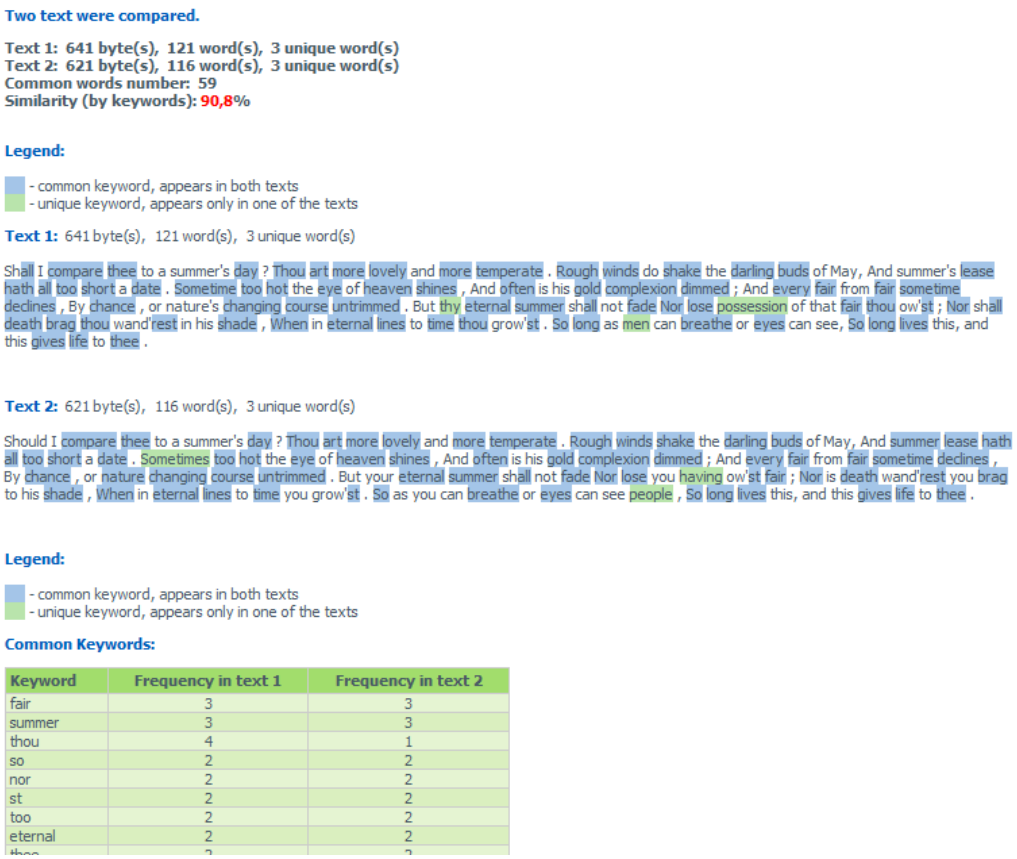
Feature	Present in Norwegian version 83.1%	Present in Latvian version 77%	Present in Latin version 37.2%	Present in Catalan version 32.9%
Asyndeton of 'above, around, in front'	Yes	Yes	No	Partial
Paradox of 'everywhere at once, and yet nowhere'	No	Yes	No	Partial
'springing'	Yes	Yes	No	No
'Beams of light' mirroring 'sun's rays'	No	Yes	Partial	Partial
'earth and heaven' expression	Yes	Yes	Partial	No
Adjective of 'reflecting blade'	Yes	Yes	No	No
Adjective of 'circling gleams'	Yes	Yes	No	No
Adjective of 'keen rush'	Yes	Yes	No	No
Adjective of 'sharp hisses'	Yes	Yes	No	No
Alliteration of 'sharp hisses'	Yes	Yes	No	No
Assonance of 'almost a whistling'	Yes	Yes	No	Partial
Onomatopoeic 'st' and 's'	Yes	Yes	No	No
Subject of sentence 'the atmosphere'	Yes	Implied	Yes	Yes
Phallic 'blade'	Yes	Yes	No	No
Trapping of Bathsheba by 'enclosed'	Yes	Yes	No	No
Expression from Genesis, 'firmament of light'	Yes	Yes	Yes	No
Simile of 'sky-full of meteors'	Yes	Yes	Yes	No

4.1.2 Poetry: Text B

The following literary analysis of the sonnet is published at https://letterpile.com/poetry/A-Literary-Criticism-of-Shakespeares-Sonnet-18. A Shakespearean sonnet was used as the poetry text and produced the re-translation with the highest degree of similarity in Filipino (Figure 4.2). A poetic analysis of Text B is as follows.

Original text

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate.



Two text were compared.

Text 1: 641 byte(s), 121 word(s), 3 unique word(s)
Text 2: 621 byte(s), 116 word(s), 3 unique word(s)
Common words number: 59
Similarity (by keywords): 90,8%

Legend:

- common keyword, appears in both texts
- unique keyword, appears only in one of the texts

Text 1: 641 byte(s), 121 word(s), 3 unique word(s)

Shall I compare thee to a summer's day ? Thou art more lovely and more temperate . Rough winds do shake the darling buds of May, And summer's lease hath all too short a date . Sometime too hot the eye of heaven shines , And often is his gold complexion dimmed ; And every fair from fair sometime declines , By chance , or nature's changing course untrimmed . But thy eternal summer shall not fade Nor lose possession of that fair thou ow'st ; Nor shall death brag thou wand'rest in his shade , When in eternal lines to time thou grow'st . So long as men can breathe or eyes can see, So long lives this, and this gives life to thee .

Text 2: 621 byte(s), 116 word(s), 3 unique word(s)

Should I compare thee to a summer's day ? Thou art more lovely and more temperate . Rough winds shake the darling buds of May, And summer lease hath all too short a date . Sometimes too hot the eye of heaven shines , And often is his gold complexion dimmed ; And every fair from fair sometime declines , By chance , or nature changing course untrimmed . But your eternal summer shall not fade Nor lose you having ow'st fair ; Nor is death wand'rest you brag to his shade , When in eternal lines to time you grow'st . So as you can breathe or eyes can see people , So long lives this, and this gives life to thee .

Legend:

- common keyword, appears in both texts
- unique keyword, appears only in one of the texts

Common Keywords:

Keyword	Frequency in text 1	Frequency in text 2
fair	3	3
summer	3	3
thou	4	1
so	2	2
nor	2	2
st	2	2
too	2	2
eternal	2	2

Figure 4.2: Similarity between Text B original and the version translated back into English from Filipino

Rough winds do shake the darling buds of May,
And summer's lease hath all too short a date.

Sometime too hot the eye of heaven shines,
And often is his gold complexion dimmed;
And every fair from fair sometime declines,
By chance, or nature's changing course untrimmed.

But thy eternal summer shall not fade
Nor lose possession of that fair thou ow'st;
Nor shall death brag thou wand'rest in his shade,
When in eternal lines to time thou grow'st.

So long as men can breathe or eyes can see,
So long lives this, and this gives life to thee.

Shakespeare, *Sonnet 18*

Structure

This sonnet is an example of typical Shakespearean style, comprising three quatrains in iambic pentameter ending in a heroic couplet, following a rhyming scheme of abab cdcd efef gg. It follows the tradition of dividing the sonnet into two parts. In the octave, Time is shown as the enemy of the transitory nature of beauty and there are references to different passages of time, 'day', 'May', 'date', 'summer'. After the volta, highlighted by 'But', the sestet introduces Time as the solution: the youth's beauty will be everlasting as long as the sonnet exists and the references are to the 'eternal' and 'So long as'. The final couplet, although part of the sestet, could stand alone and provides a strong closing point.

Technical devices

It is significant that there is only one enjambment; every line except line 9 finishes with punctuation. This is a poem of stated facts rather than rambling musings.

Repetition ('more lovely and more temperate', 'every fair from fair') and anaphora (lines 6 and 7, lines 10 and 11, lines 13 and 14) are used heavily throughout the sonnet. These techniques are used for emphasis, to accentuate the point being made. Contrasts are emphasised by antithesis, 'more temperate./Rough winds' and the last word of lines 5 and 6, opposing 'shines' with 'dimmed'.

Alliteration, a linking device, is lightly used which makes it more effective when it does appear, 'chance, or nature's changing course', used at the end of the octave. The next use is in the final line, 'long lives this, and this gives life to thee' where the double alliteration of the 'l' and 't' force the line into prominence.

The object of the sonnet

The poem begins with a rhetorical question to 'thee' (commonly assumed to be a youth (Drabble, 1996)) so it seems as though the poem is going to be about the young man. However, the stressed 'I' of the first line contrasts with the unstressed 'Thou' of the second, foreshadowing the theme of the poem; it is less a tribute to the youth's beauty than a proclamation of the writer's skill and his assurance that his poem will be a future classic. This suggestion is furthered in the 12th line, 'in eternal lines', referring to the lines of the poem. Shakespeare has broken the fourth wall by acknowledging the poem and the existence of readers.

Use of metaphor

Personification occurs throughout the poem in the form of Summer ('summer's lease'), the Sun ('his gold complexion'), Nature ('nature's changing course') and Death ('shall death brag'). Summer and Death are personified to suggest a human relationship: Death is a rival for the poet's love. The 'summer's lease' is echoed in line 8 at 'thou ow'st', extending the metaphor further. From the first line, Shakespeare invites a comparison with summer and this continues through to the final couplet. Summer, generally presented as the perfect season, falls short of the youth's perfection and is unworthy to be compared to him. Summer has 'Rough winds', and 'too short a lease' while the youth's 'eternal summer' is reinforced at the beginning of the sestet.

The sun is represented as 'the eye of heaven'. The 'gold complexion dimmed' can be interpreted both as the sun's strength and beauty tarnished by clouds, just as the youth's beauty will be tarnished by time, but also 'complexion' can be read as 'temperament' (i.e. a combination of the four humours). This latter interpretation echoes 'temperate' of line 2 effectively (Ray, 1994).

Nature's 'untrimmed' has a double meaning. It can mean either unadorned, indicating that Nature will strip the youth of his 'fair' beauty but can also refer to the sails of a ship, suggesting that Nature's course is unadjusted. However, Jungman (2003) has suggested that the 'untrimmed' may actually mean 'unadjusted' and therefore Shakespeare is saying that the thing that remains unchanged is Nature's changing, 'mutability is eternal'. This interpretation strengthens the structure of the sonnet with the octave representing change and the sestet reinforcing the endurance of the written words.

Re-translated texts

Once again, for simplicity, only the two most and two least similar texts are included.

Filipino, with 90.8 per cent similarity to the original

Should I compare thee to a summer's day?
Thou art more lovely and more temperate.
Rough winds shake the darling buds of May,
And summer lease hath all too short a date.

Sometimes too hot the eye of heaven shines,
And often is his gold complexion dimmed;
And every fair from fair sometime declines,
By chance, or nature changing course untrimmed.

But your eternal summer shall not fade
Nor lose you having ow'st fair;
Nor is death wand'rest you brag to his shade,
When in eternal lines to time you grow'st.

So as you can breathe or eyes can see people,
So long lives this, and this gives life to thee.

Hindi, with 89.2 per cent similarity to the original

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate.
Rough winds do shake the buds of May, Sir,
And summer's lease up all too short a date.

Sometime too hot the eye of heaven shines,
And often is his gold look dim;
And every fair from fair sometime declines,
By chance, or nature's changing course untrimmed.

But thy eternal summer shall not fade

Nor lose possession of that fair thou ow'st;
Nor shall Death brag thou wand'rest in the shade,
When in eternal lines to time thou grow'st.

So long as men can breathe or eyes can see, can
So long lives this, and this gives life to thee.

Korean, 53.8 per cent similarity to the original

I compare the summer day you do?
You art more lovely and more temperate.
Rough winds do shake the bud a lovely February
Summer's lease is too short a date in the dark.

Sometimes it's too hot in the sky shining eyes,
And often his gold complexion dimmed;
When all the processes in the process, rejection
By chance, or nature's changing course untrimmed.

But thy eternal summer will not fade
Ow'st you have also lost equity;
In addition, death, to thee in the shade boasts wand'rest
When you time a permanent line in. grow'st

This person can rest or one eye to see
Too long for this life, it gives life to thee.

Latin, 50 per cent similarity to the original

If your summer compare this day
You are more handsome and more temperate.
Changes darling buds of May rough winds
And he has a short summer course too friendly.

Once too hot the eyes of heaven shines,
And often, the gold complexion dimmed;
Every fair is now on equal terms, retired,

It may be promised and changing the course or nature.

Not disease but eternal
Nor lose possession of that fair thou ow'st;
Neither the death of you in the shadow of his more wand'rest,
Then into the field with grow'st everlasting.

As long as they can breathe or eyes can see,
While life, and gives life to thee.

Surviving literary features

Table 4.2 shows the degree to which literary features survive the re-translation process. In common with the prose, the poetry retains a higher proportion of literary devices in the texts with greater similarity to the original, but again, more features were at least partially retained.

Table 4.2: Feature analysis of re-translated versions of Text B

Feature	Present in Filipino version 90.8%	Present in Hindi version 89.2%	Present in Korean version 53.8%	Present in Latin version 50%
Iambic pentameter	Yes	Yes	No	No
Rhyming scheme abab cdcd efef gg	abab cdcd efeg hi	abcb dedf ghgh ij	abcd efgf hijk ll	abcd efgh ijjk ll
Clear difference between octave and sestet	Yes	Yes	Yes	Partial
Time references in octave, 'day', 'May', 'date', 'summer'	Yes	Yes	Partial	Yes
'But' at volta	Yes	Yes	Yes	No
Expressions of endurance, 'eternal', 'So long as'	Partial	Yes	Partial	Yes
Strong final couplet	No	Partial	No	Partial
Little enjambment	Yes	Yes	Partial	Partial
Repetition in 'more lovely and more temperate'	Yes	Yes	Yes	Yes
Repetition in 'every fair from fair'	Yes	Yes	No	No
Anaphora of 'And often ... And every'	Yes	Yes	No	No
Anaphora of 'Nor lose ... Nor shall'	Yes	Yes	No	No

Feature	Present in Filipino version 90.8%	Present in Hindi version 89.2%	Present in Korean version 53.8%	Present in Latin version 50%
Anaphora of 'So long... So long'	No	Yes	No	No
Antithesis of 'more temperate. Rough winds'	Yes	Yes	Yes	No
Antithesis of 'shines' and 'dimmed' at end of lines	Yes	Yes	No	No
Alliteration of 'chance, or nature's changing course'	Yes	Yes	Yes	No
Alliteration of 'long lives this, and this gives life to thee'	Yes	Yes	Partial	Partial
Stressed on 'I', unstressed on 'Thou'	Yes	Yes	No	No
Broken fourth wall	Yes	Yes	No	No
Personification of summer	Yes	Yes	Yes	Partial
Personification of the sun	Yes	Yes	Yes	Partial
Personification of nature	Yes	Yes	Yes	No
Personification of death	Yes	Yes	Yes	Partial
Echoing of 'summer's lease' and 'thou ow'st'	No	Yes	Yes	No
Comparison with summer	Yes	Yes	Yes	No
Double meaning of 'complexion'	Yes	No	Yes	Yes
Double meaning of 'untrimmed'	Yes	Yes	Yes	No

4.2 Implications of using translation tools

This study was necessarily limited in scope and self-limited by the exemplar textual fragments chosen and by tools selected with consideration of fine-grained category and feature stylistic analysis rather than, for example, hermeneutics, narrative patterns and holistic deconstruction. A structural analysis is only one way to approach literature and does not include the rich, deep analysis provided by taking a post-structuralist, post-modern approach or by using a feminist/Marxist/psychoanalytic critical view of the texts.

A high degree of similarity reveals that fine-grained style (such as alliteration, use of adjectives, anaphora) is quite well preserved with viability induced by the efficacy of black-box engines and pre-stored corpora. Particularly with

regard to Text A, the linguistic family etymology demonstrates greater similarities between Germanic than Romance languages which produced garbled sentences such as, 'These are the embrace he rushed shine with the keen hissing' (Latin) and, 'These flashes were sometimes accompanied by acute fever was almost a whistle' (Catalan), although there were some notable exceptions. Some errors were understandable, such as the Bulgarian translation of 'close at hand' into 'at your fingertips', while the German translation of 'These circling gleams' into 'Even mushrooms' was baffling. Although a closely related language to English, German actually scored the same similarity (51.4 per cent) as Basque, Chinese and Vietnamese, suggesting a specific issue with the German translation, particularly as the same issue occurred using different texts. Some translations became gibberish. In Persian, 'the atmosphere' became 'Joe'. Occasionally, all meaning broke down, as in the Korean which adds, 'a sharp gyeuidoeeotseupnida Ballroom' at the end of the text.

Not surprisingly, there was no "infinite monkey" effect[1] and none of the re-translated texts were an improvement on the originals, nor did any of the texts produce any novel literary device.

4.3 Summary

Prose and poetry text were translated from English into 62 different languages, then re-translated back into English. This meant that some literary qualities were lost through the translation process. The texts were then examined to see which stylistic features survived the transformation. More subtle aspects such as the use of similes and appropriate adjectives were not always well preserved. A maximum of 90 per cent similarity between texts suggests literary excellence is reliant on implicit stylistic norms and cultural semantic contexts which operate at aesthetic levels. The missing 10 per cent is significant and the human experts regarded the missing components as being aesthetically detrimental compared with the original literary quality. However, the texts were still recognisable as literature; they would not be mistaken for a news article or a non-fiction work, for example, suggesting that the literary devices are reasonably robust and therefore likely to overcome

[1] Attributed to Émile Borel, the theory that an infinite number of monkeys randomly hitting keys on a typewriter for an infinite amount of time will eventually produce the complete works of Shakespeare.

any minor parsing errors.

As a result, it was decided that the translation experiment had served its purpose and the next stage of the study should begin by determining the nature of Literature.

Chapter 5

Determining the Human Perspective of Literature

The previous chapter focused on the feasibility of the research project and concluded that fiction texts are sufficiently robust to maintain a high degree of literary features even when subjected to computational analysis. This chapter focuses instead on the factors that humans use to judge how well-written is a particular book. As Figure 1.1 shows, Chapters 4 and 5 are independent of each other but serve to provide a background to the practicality and feasibility of the research.

Without a clear understanding of what constitutes "good" literature, it is not possible to attempt judgements of literary merit, therefore the focus of this chapter is to provide an overview of the human reaction to literary texts by providing a brief introduction to theories of literary criticism that will be used in assessing the efficacy of the CoBAALT model in Sections 7.3.2 and 7.3.3 and to examine how far stylistic analysis is a human measurement of literary merit. Reading provides a unique, rich, emotional experience for the reader, 'Reading is to the mind what exercise is to the body' (Steele, 1709), and their reaction to a novel can even change over time as observed by Thomasson (2004, p. 152). If the computer is to make an aesthetic judgement based on the style of writing, it is important that humans make similar decisions from the same information. This chapter examines the human perspective to guide the development of the CoBAALT model's focus, finding that stylistic analysis is a key component in determining literary merit. Had this not been an important aspect in human responses, CoBAALT would have had to

move towards different metrics that reflect other qualities of human aesthetic judgement.

5.1 A brief history of modern literary criticism

The roots of literary criticism date back to Plato (circa. 428 - 347 B.C.) and Aristotle (384 - 322 B.C.) who are acknowledged as the first to express Art as something that can be interpreted and evaluated (Habib, 2005); their influence held sway until the beginning of the twentieth century. The definition of Literature is notoriously difficult to pinpoint, with schools of thought from Formalism to Structuralism, Post-Structuralism, Post-Modernism, Feminism, Marxism, and still there is no scientific consensus of agreement. Literature has been neatly classified as something which produces 'a sense of universal value' where 'a rare glimpse of transcendence can still be attained' (Eagleton, 2008), while Eco has called it 'a universe in which it is possible to establish whether a reader has a sense of reality or is the victim of his own hallucinations' (Eco, 2012). These explanations, however, describe the effect literature has, rather than its essence. In crudest terms, literature can be thought of as fiction, but this unfortunately excludes literary non-fiction, such as *Testament of Youth*, and yet encompasses much that is not considered literary, an example being the currently popular, yet poorly written, *Shades of Grey*. As readers, we have an intrinsic understanding of what is and what is not literature, but firmly categorising certain works highlights the ambiguity of established definitions. Much like Justice Stewart's definition of pornography[1], we just know it when we see it. However, specific schools of thought have been established which can be used to formulate a method for the identification of what, in this thesis, comprises good literature.

Figure 5.1 illustrates the flow of literary theories from Aristotle onwards. The schools born of the New Criticism are reliant on the reader's responses and world knowledge for their assessment of literature; those theories that evolved from Structuralism, however, are computationally more easily quantifiable.

[1] http://corporate.findlaw.com/litigation-disputes/movie-day-at-the-supreme-court-or-i-know-it-when-i-see-it-a.html

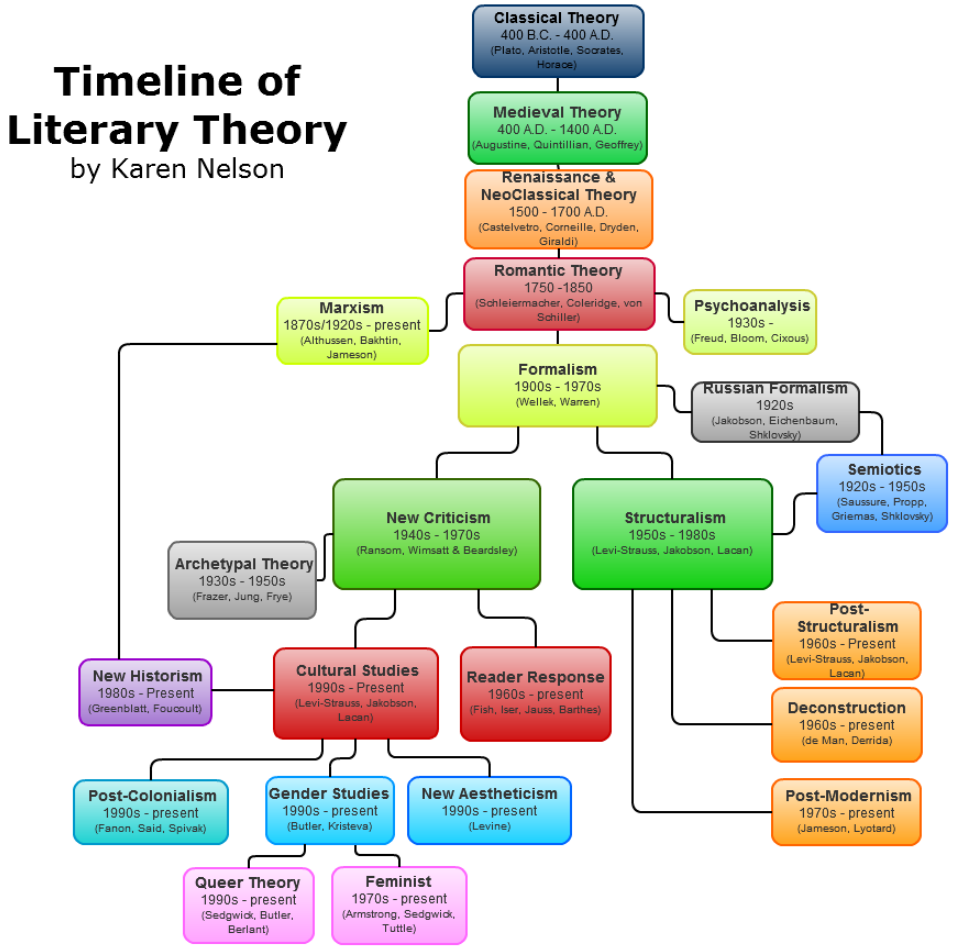


Figure 5.1: Literary theory timeline with key players (Nelson, n.d.)

5.1.1 Formalism and New Criticism

Formalism is a concentration on literary form and the structure of language that greatly influenced literary criticism throughout the first three-quarters of the twentieth century (Wales, 1990, pp. 184-185). The Formalists were a movement dedicated to emphasising the separation of literature from reality rather than acting as a mirror to it (Barry, 2009, p. 155) and sought to understand how literature worked - what made it literature? (Habib, 2005, p. 603).

For the Formalists, everything needed to know about a piece of literature is found in the text itself. What makes literature literary, they postulate,

is the use of language in a way that is not commonplace (Barry, 2009, p. 155). Eagleton gives an excellent example of this: 'If you approach me at a bus stop and murmur "Thou still unravished bride of quietness", then I am instantly aware that I am in the presence of literature' (Eagleton, 2008, p. 2). The language used is different, intensified and excessive. 'There is disproportion,' as Eagleton puts it, 'between the signifiers and the signified' which are defined as the means of identifying something and the concept, respectively; as an example, the letters *c,a,t* in that specific order are the English language signifiers for the concept of a feline mammal.

This Formalist view of literature as an amalgam of literary devices that are utilised in unusual ways is echoed in the argument made by Shklovsky, that the purpose of art is to make things strange, or *ostranie* (Hawkes, 1977). This has the effect of surprising the reader (or viewer, or watcher, or listener), of making him or her look again at something commonplace. Such technical literary devices are accessible and quantifiable by the computer.

Out of the Formalist school arose New Criticism, a movement that further focussed literary attention on the text. Ransom, for example, specifically excludes the relevance of analysis of personal impressions, synopsis, historical background, linguistics, morality or anything outside the work itself (Ransom, 1937), all of which are factors that are difficult to reproduce with accuracy in computational analysis. Any investigation into the writer's motivation or background was discouraged by the New Critics as being irrelevant (Drabble, 1996, p. 704), as was the reader's response to the writing (Eagleton, 2008, p. 42). As a theory, however, New Criticism was more concerned with poetry than prose and the movement had reached the height of its popularity by the 1950s. Academic attention, fuelled by the rise of influences like Chomsky, was beginning to focus on a linguistic approach (Barry, 2009, p. 264).

5.1.2 Structuralism and Semiotics

Structuralism is what Golban and Ciobanu (2008) call 'a human *science*' (their emphasis) that sees literature in terms of its relation to linguistics and it was a movement that was profoundly influenced by the founder of modern linguistic theory, Ferdinand de Saussure (Eagleton, 2008, p. 84).

Semiotics is the theory of signs (Drabble, 1996, p. 880). From a semiotic viewpoint, literature is merely the medium for a sign or concept. Saussure

identified meaning to consist of the signifier and the signified, with the text acting as a 'two-sided psychological entity' (de Saussure, 1983). In terms of literature, the signifier and the signified equate to the written word and the concept (i.e. the message the author wishes to convey), respectively. The concept is where literary language comes into play. Whether it is called 'a rose' or 'a flower with soft pink petals like the cheeks of a child', the reader understands what is meant.

Another founding father of semiotics, C. S. Peirce takes this structure further and introduces a triadic model consisting of the representamen or signifier (the symbol), the interpretant or signified (the sense made of the sign) and the object or referent (what the sign represents) (Wales, 1990, p. 420). In literature, these become the written words, the reader's reaction or understanding of those words, and the concept, respectively. For Structuralists, the sign (the concept) is understood in different ways by different people depending on their individual interpretation so there can be no concrete meaning. Barry (2009, p. 42) illustrates this notion of individual preconception by recalling an event when asking a ticket collector at the train station for directions to the Brighton train. It being a Sunday and with engineering works under way on the tracks, the train had been replaced by a bus service. When the ticket collector pointed to the bus, there was instant understanding that this bus service was, temporarily, the train to Brighton. This poses a conundrum for automatic analysis as the computer does not bring interpretation with it; it has no preconceived understanding of the world unless it has been programmed to do so.

5.1.3 Post-modernism

Postmodernism dissolves the boundary between the real and the simulated (Barry, 2009, p. 86). Writing, says Wales (1990, p. 366) is 'highly self-conscious, aware of itself and of the reader reading it'. For Postmodernists, there is greater emphasis on the role the reader plays in the appreciation of literature. There are no absolutes; everything is an inference, formed by the reader's cultural and historical experiences.

Tyson (1999, Ch. 8) gives an example of this using the expression 'Time flies like an arrow'. Our usual interpretation of this phrase is that time passes quickly, where

(noun)	(verb)	(adverbial clause)	(meaning)
Time	flies	like an arrow	Time passes quickly

However, there are other ways to understand the same line, such as

(verb)	(object)	(adverbial clause)	(meaning)
Time	flies	like an arrow	Take out your stopwatch and time the speed of flies in the same way as you would time an arrow

or even

(noun)	(verb)	(object)	(meaning)
Time flies	like	an arrow	Time flies (probably little insects resembling fruit flies) are fond of at least one arrow

In Hall's model (Hall, 1973), one of three 'reading positions' is adopted by the reader: hegemonic, negotiated or oppositional, depending on the degree to which the reader agrees with the intended interpretation of the text. Note that this interpretation does not necessarily mean the author's intention but is the accepted reading, that position which fits in with the world view of the majority. If Hall's model of reading position is applied, this raises the question of what happens when the reader is not human and therefore unable to take a reading position. Where does this leave the interpretation and the prognosis for machine analysis?

Culler calls the reader 'a virtual site for the location of codes of literary interpretation' (Culler, 1992). His argument is that each reader interprets text as they understand it, so two readers with different cultural and historical backgrounds will interpret differently, neither being "correct" or "wrong". If this is so, then there is little difficulty in substituting a machine for a human. The computer is just another receptacle for the written word, albeit one that takes a distinctly conformist stance unless instructed (programmed) to do otherwise. However, the amount of programming needed to bring a computer to even the most basic levels of human aesthetic responses is considerable and the risk is that the response would be a mere replication of the programmer's own. For a more independent appreciation, we need to look at stylistics.

5.1.4 Stylistics

Stylistics evolved from the study of classical rhetoric to become a text-centred literary critical theory in its own right, marrying literary effects to their linguistic origins (Wales, 1990, pp. 437-8). Analytical tools used by linguistics scholars are adapted to identify features in literature, bringing a scientific approach to what had previously been an impressionistic and intuitive art. This new approach was not welcomed by the traditionalists and a schism between the linguistic parsers on one side and the literary academics on the other that resulted in vicious verbal pugilism between Roger Fowler, the editor of *Essays on Style and Language: Linguistic and Critical Approaches to Literary Studies* (1966) and F. W. Bateson, the editor of *Essays in Criticism*, in which a reviewer suggested that linguists were inadequate to the task (Simpson, 2004). This antagonism between the approaches continued well into the 1980s although Barry (2009, p. 201) suggests that there is still deep suspicion by academic critics about stylistic analysis.

Wales defines stylistics as a method of showing how the functional significance of formal textual features impact the interpretation of the literature, adding that stylometry is a sub-discipline that takes a statistical analysis approach in order to determine stylistic patterns (Wales, 1990, pp. 438-9). Background features such as sentence length and function words are used unconsciously by authors and can be used to determine a particular writing style. These can then be analysed to determine literary merit as defined by a chosen set of metrics.

Barry (2009, pp. 203-5) outlines three main objectives for a stylistic approach to literature:

1. to provide hard data to support intuitions;
2. to bring new interpretations based on linguistic use;
3. to determine how literary meaning is created.

The focus of this thesis is on the first factor: providing computationally derived evidence to support a definition of literature.

Without a human reader to bring their wealth of life experiences to the text, a post-New Criticism computer analysis is not possible unless specifically programmed by a human who will bring their own insights and prejudices

to the computer, thereby excluding any psychoanalytical, feminist, Marxist or eco-critical school of true interpretation, and restricting the analysis to one of stylistics. A stylistic analysis, however, lends itself very well to the statistical techniques used to determine particular writing traits. In order to accomplish this, the appropriate features must be identified.

5.2 The human perspective

For an investigation into computational appreciation of literature, it is crucial to attempt a definition of what makes a book literary. Human interpretation of text and the reading experience would therefore need to be investigated. A focus group was the most appropriate choice for the development of a new hypothesis (Powell and Single, 1996; Krueger and Casey, 2009). The findings of the focus group could then be used as a basis for more detailed and specific investigation through questionnaires and surveys (Hoppe et al., 1995).

To this end, two focus groups were held with different participants, all keen readers, to discuss what makes a book literary. Both groups were given minimal direction so that the opinions of the moderator did not prejudice the results of the discussion. Their opinions were broadly classified into three main areas of interest: plot, descriptions and theme. Once the main areas of literary influence were identified, an online survey was produced which was open to members of the general public to see if they broadly agreed with the assessment of literary device influence on their reading experience. Finally, face-to-face interviews were carried out with English Literature teachers to see what areas of critical analysis are used by humans that can be identified and qualified automatically.

5.3 Focus groups

Two separate focus group were gathered. The first (FG1) comprised six people: five were female, with an age range of between forty and fifty-one, and the male declined to give his exact age but is somewhere between fifty and sixty. The second group (FG2) comprised eight people: four male and four female, aged between 50 and 80. All participants are from the same socio-economic demographic and live in Bedfordshire, and they are all regular

members of a book group. The groups were asked to discuss what they feel makes a book literary.

The sessions were not recorded at the request of the groups. Instead, the researcher (moderator) took notes (Appendix A), a process that was facilitated by the relaxed and congenial nature of the discussions. Observations and opinions were jotted down verbatim where possible along with the participant's initials. Once the discussions were over and the focus groups disbanded, the responses were coded into common topics such as 'plot', 'description' and 'theme'. Characterisation did not feature as strongly as expected but was also coded. Where several people had made the same observation, the most relevant or articulate quote was chosen for inclusion.

The following sections present some of the participants' responses grouped into the important literary aspects identified (see Appendix A for focus groups' instructions).

5.3.1 Plot

It was quickly agreed that plot is important but that it is not the most crucial aspect of a book's literary credentials; in fact, it was observed that it is rare to find a plot-driven novel that is also well-written. There were several disparaging remarks about Dan Brown's novels which are generally acknowledged to be page-turners with complex, fast-paced plots without pretension to being literary. This opinion of plot as less important for literary quality is consistent with the Formalist movement. However, it was also pointed out that there must be some plot for the writing to retain the characteristic of a book.

> The plot is what holds the story together. If you don't have a plot, you haven't really got anything worth reading. J.

> The flow is important and it is the plot that controls that. I want there to be some mystery right to the end. L.

The inaugural book choice of FG1 was *Gadsby*, a 50,000 word lipogram without the letter 'e'. Although this was an interesting choice from the point of view of a writing challenge, the restriction meant that there was little plot to

the story and five of the six participants admitted that they would not have read it completely if it had not been for the requirements of the book group.

> A book with no letter 'e'? What tosser thought that would be a good idea? J.

> Even poetry has a plot of sorts. H2.

> Would you call it a plot? Maybe a narrative thread. H1.

> It might be an interesting exercise but I wouldn't enjoy it as a good read. T.

The groups agreed with this latter point.

It was pointed out by one of the participants that detective fiction is often an enjoyable read but difficult to discuss at a book group meeting because the genre tends to be wholly plot-driven, leaving little else to debate apart from the "whodunnit" aspect. However, it was agreed that absence of *any* plot would definitely detract from literariness.

5.3.2 Description

Descriptive passages were suggested as a guide to literariness, but this led to considerable debate. On one hand, descriptions can be used to evoke a strong sense of place and time which is crucial to the enjoyment of a novel but on the other hand, clumsy descriptions detract from it. An example was given of novels that involve the character looking in a mirror, purely for the author to have a lazy way of describing what the character looks like.

> I like books that paint a word picture. R.

> I agree. A book should draw you in with the description so you feel you are really there. H1.

> Where does description end and purple prose begin? How much is too much? Moderator

> It depends on the book. A.

> When it starts to impose on your reading. H1.

> I don't like too much description if it gets in the way. Just get on with the damn story. C.

FG2 was unanimous that description is important in marking a book out as literary.

> Descriptions have to be real, to paint a visual picture so you are drawn right in to the story. T.

> They give you a sense of place and person straight away. L.

> A large vocabulary is an asset in a book and that becomes more obvious in description. It gives the novel an artistic element and that, surely, is what we mean by *literary qualities*. F.

5.3.3 Theme

The theme of a novel is what it is about or the underlying message the writer wishes to convey. The focus groups had very different opinions on their favourite themes but some common strands did arise.

> I want to feel I have learnt something new. H2.

There was a general consensus of opinion on this point.

> Context is important. Like Dickens in his time. I like that sort of social comment on a period in time. R.

> Allusion to other things and other works is important...intertextuality. References to other literature can show me links I had not seen myself or confirm what I have already know. H1.
>
> Isn't that a bit pretentious? It's just showing off how much the writer knows. A.
>
> Good. I want a clever writer. Layers of metaphor, too, so you get a story within a story. H1.

'Can you give us an example?' Moderator

> There are loads. Like, the Harry Potter series is about a boy wizard fighting evil, but it is also about the class system and oppression of minorities. *La Peste* is about a real and, at the same time, a metaphorical plague... I also like to see some foreshadowing or misdirection in true tragic style...an example would be something like *A Prayer for Owen Meaney* where we know something awful is going to happen but don't realise how it all fits together until the end. H1.

FG2 was less concerned with the theme of a novel, although there were a few suggestions such as 'historical accuracy', 'interesting characters', 'humour' and 'active voice'. It was observed that for this last point, modern novels are almost always in the active voice whereas classics use a more passive voice.

The groups were asked if they thought a computer could be taught to appreciate literature if it knew what to look for. Three (A., J. and C.) immediately said 'No'. However, it was pointed out that students are taught to analyse literature in terms of authors' use of form, structure and the language used, therefore there are features that can be quantified. Other aspects such as alternative interpretations and understanding inter-textuality or references to other cultural identifiers would be more problematic.

5.4 Online survey

Using the results of the focus groups, an online open survey was conducted (Appendix B). Readers with a spread of preferred genres were asked what

they looked for in a good book. A Likert scale was provided for the features identified by the focus group and a box provided for other suggestions but many were easily incorporated into the existing choices, such as 'Pace' being part of 'Plot' and 'Witty' or clever dialogue coming under 'Use of language'. Thirty-eight respondents completed this section and the results confirmed that the degree to which the features identified are important are similar, although 'Learning something' was the least important factor to the respondents.

Respondents were also asked 'What makes a good book stand out?' Thirty-one answered this question. Among the more common answers were 'A good book is one I don't want to put down', 'A good plot and beautiful writing' and 'Credible, interesting characters'. Respondents were then asked for their three favourite books with a reason for their preferences and twenty-nine answered. The most common reasons are shown in Table 5.1.

Table 5.1: Respondents gave reasons for their choice of favourite book

Feature	Number of respondents who identified this feature
Gripping or cannot put it down	8
Characters	7
Plot	6
Use of language	4
Unpredictable	4

Of these responses, 'Gripping or cannot put it down' is a description of the reader's emotional reaction to the story, a factor that is too subjective to the individual to quantify computationally. Character and plot are specific to each novel so although a series could be investigated, such as Trollope's Barchester novels which follow characters from book to book or a crime series featuring the same detective, there is too much variation to compare these factors across different genres. Even comparing character types such as villains would be difficult when considering the difference between Satan in Milton's *Paradise Lost* (often considered to be the true hero of the poem (Steadman, 1976)) and C. S. Lewis's White Witch from the Narnia stories who is 'evil itself' (McSporran, 2005). Fascinating as this line of enquiry would be, it is outside the scope of this thesis.

The final question of the survey was answered by thirty-three respondents and

asked whether they thought a computer was capable of telling the difference between a good book and a poor one. Although this was not analysed, it was an interesting question because of the different reactions the researcher has received when discussing this thesis. Thirteen said it was not possible, seven thought it might be possible one day and only three said it was feasible (the other answers were not classifiable). Some of the verbatim answers are as follows:

> No. This is a philistine idea. The soul exists. A great book taps into it, and a computer cannot.

> ...while they can recognise use of language, probably evaluate development of a plot, I rather doubt they can have that "Aaahh!" of charm and later remember it in a thinking way.

> No. I consider that whether we find a book good is driven by its quality but also the reader's emotional context and desires at the time which a computer cannot anticipate or emulate.

> A "good book" is one that connects with you personally as a reader, not the one that is technically and grammatically correct, or the one that has the correct elements to make the formula of a "good" book.

5.5 Feature extraction for humans

As the comments in the previous section show, humans relate to literature for personal and different reasons. However, the fact that English Literature exists as an academic subject demonstrates that there are aspects that can be qualified and quantified by competent reading and that these are features that can be taught. Interviews with teachers were carried out to determine how humans are taught to differentiate standards of literature (see Appendix C for interview instructions).

There are three main areas that are identified: form, language and structure. Form studies the literature within its genre, determining whether the text

conforms to the expected norms of the genre and identifying its type (epistolary, narrative voices and so on). Language investigates features like imagery, metaphor and lexical fields (words that belong together). Structure is the investigation of patterns within the text such as looking for repetitions (assonance and alliteration), mimesis (characters or situations reflecting real life), juxtapositions and lengths of paragraphs, sentences and words. Structure can be further split into micro and macro structures, investigating details like punctuation (micro) and chapter structure (macro).

As form relates to a particular genre, it is book-specific and so not relevant to the investigation of general literature in this thesis. Language, with its focus on metaphor and imagery, demands a real-world knowledge that is less available to the computer on a literature-wide scale. Structure, however, is an area of interest.

> You can teach children as young as eight to look for patterns in stories and they quickly pick up how to identify features such as alliteration. It is more difficult to teach them *why* a particular feature is important. This is one of the problems teachers face with the current SAT demands for 11 year-olds. For example, children must include a fronted adverbial in their composition to gain a mark but there is no reason that "Happily, she skipped across the road" is a better sentence construction or more literary in any sense than "She skipped happily across the road", yet marks would be allocated for the first example but not for the second one. It is feature identification without any understanding of its implications.

'What is the implication of alliteration or assonance once it has been identified?' Interviewer

> Both are often used to slow down the pace where an author might want to place some particular emphasis. Alternatively, the use may be mimetic. Alliteration is often used for onomatopoeic effect, like the sibilance of an 's' echoing the hissing of a snake. It can be a linking device, too, bringing parts of a sentence together.

'Can you give examples of literature that includes varying lengths of chapter/paragraphs/sentences? Why is this effective?' Interviewer

> Toni Morrison uses brevity to great effect in *Beloved*. The first sentence of the book is only three words, the next only five. These sentence lengths gradually get longer, reflecting the initial reluctance of the storyteller to reveal what has happened. Or you have someone like Kurt Vonnegut who wrote a short story called *Cat's Cradle* that comprises 127 chapters. He himself called his books 'mosaics'. That's an effective writing technique right there. *Tristram Shandy* is another one that plays with form, having black pages after the death of a character.

5.6 Summary

This chapter has outlined the nature of what constitutes "good" literature by understanding how human readers interpret text. It is evident from the historical perspective that literary criticism is subjective. Human interpretation of literature is multi-faceted and there is no single aspect that separates good writing from bad. Some features are entirely subjective and dependent upon the individual, such as reading to learn something new. However, sufficient features are identifiable to determine literary worth, as suggested by the fact that children can be taught how to recognise those that add meaning or enrich the text.

By using focus groups, surveys/questionnaires and interviews, it has been possible to identify standards for literary merit. However, these are only pointers towards the overall reading experience. Only by breaking down texts into their components can an analytical model be created, something that the average reader does not consciously do. The identification of these components will be achieved computationally and the results compared to the human experience (Section 6.3.2). Feature identification is one aspect but qualification appears to be equally important. It is not sufficient to observe a textual feature and call it literary; it has to serve a purpose. For this reason it is unlikely that POS alone will be adequate as badges of merit, although they may be good indicators of a specific style, and alternative features need to be investigated. Those eventually selected are given in the following chapter in Table 6.6.

Chapter 6

Creating the Tools to Determine Literary Quality

Chapter 5 demonstrates that there are specific literary features that constitute what we can identify as "good" literature through stylistic analysis. However, a human rarely breaks down their emotional reaction to a written work of fiction by parsing and analysing the text. For this reason, the human factor is used to guide the observations (Section 6.3.2) but not the chosen variables. This current chapter concentrates on computationally identifying the most important features to build a framework that can be used to create the eventual model of literary judgement, CoBAALT. Factor analysis is carried out to determine which variables are the most effective features for identifying "good" literature.

Categorisation and identification of POS is achieved using the natural language toolkit (NLTK). This is an open-source platform that allows users to build Python programs for NLP problems. The version used in this thesis is NLTK2.0 using Python 2.7; this version is still available but has now been superseded by NLTK3.0 which utilises Python 3. The platform was chosen for several reasons:

- it is open-source and so no purchase is necessary;
- it is well-documented with an online instruction manual with both examples and set problems (Bird et al., 2009);
- there is an active online community of users.

In addition to using the NLTK for identifying POS and lexical diversity, relative entropy is calculated using a program adapted from the paper by Torres (2002) and used in the study by Kan and Gero (2009)(see Appendix D).

6.1 Towards a POS framework

To assess the aesthetic quality of literary texts, a panel of four human experts with at least a BA in English or American Literature was recruited and asked to read two literary novels: *Heart of Darkness* (Conrad, 1899) and *Three Men in a Boat* (Jerome, 1889). Both books were written at the close of the nineteenth century and are stories set on a river, so the style and subject matter were similar although the genres were not. Each expert could select up to twenty segments from each book that they felt were particularly literary. Ten segments were chosen by more than one of the panel and these were selected for inclusion in a literary survey that was open to the general public, as it was deemed unlikely that sufficient numbers of responses would be returned if people were asked to read the entire books. Survey participants were invited to rate each segment on a Likert scale, according to how literary they found each to be. Results were scored as 5 points for 'Very literary' to 1 point for 'Not at all literary'. See Appendix E for questionnaire.

Each segment was then subjected to a series of tests including lexical diversity analysis, sentence length and POS tagging. POS tags correspond to those used in the Penn Treebank Project (Table 6.1).

Table 6.1: Penn Treebank tags

Tag	Description	Example
CC	Coordinating conjunction	and, but, either
CD	Cardinal number	5, 0.5, 1955, nineteen fifty-five
DT	Determiner	the, all, this, some
EX	Existential there	There is a place...
IN	Proposition or subordinating conjunction	in, by, until
JJ	Adjective	hard, old, fifth
JJR	Comparative adjective	harder, cheaper, nicer
JJS	Superlative adjective	hardest, cheapest, nicest
MD	Modal	can, cannot, should, will
NN	Noun (singular, common or mass)	girl, computer, thing
NNP	Noun (proper, singular)	England, NFL, Crosbie
NNPS	Noun (proper, plural)	Americans, Crosbies
NNS	Noun (common, plural)	postgrads, girls, computers

Tag	Description	Example
PDT	Pre-determiner	all, many, this
POS	Possessive ending	's
PRP	Personal pronoun	her, us, them
PRP$	Possessive pronoun	her, ours, theirs
RB	Adverb	quickly, barely
RBR	Comparative adverb	further, louder
RBS	Superlative adverb	fastest, most
TO	"to" as preposition or infinitive marker	used to, to split
VB	Verb (base form)	go, smile
VBD	Verb (past tense)	went, swam
VBG	Verb (present participle or gerund)	going, aching
VBN	Verb (past participle)	languished, flourished
VBP	Verb (present tense, not third-person singular)	sort, tend, tease
VBZ	Verb (present tense, third-person singular)	sorts, tends, teases
WDT	Wh-determiner	what, which, that
WP	Wh-pronoun	that, which, who
WP$	Possessive wh-pronoun	whose
WRB	Wh-adverb	how, why, where

Results were expressed as a percentage of the total word count for each segment to allow for discrepancies in length of text. The average segment word count was 683 words: the longest segment contained 850 words, the shortest 214. The results were then mapped to the survey results to compare.

The work in the following Section 6.1.1 has also been reported by Crosbie et al. (2013*b*).

6.1.1 Literary segment results

From the Likert questionnaire, the responses were totalled by giving one point for each step of the scale so that a segment that was perceived by all respondents to be 'Very literary' would score a maximum of 265. In practice this did not occur, but it is clear from Figure 6.1 that segments 4, 5, 8 and 9 were perceived as the most literary by the respondents. Using the criteria of literariness proposed by Gonçalves and Gonçalves (2006), the lexical diversity of each segment was also calculated. This is a simple calculation of the ratio of the total number of words in the text to the number of tokens. A type here is defined as an instance of a word, so an example such as 'the girl climbed the tree' comprises five tokens and contains four types: *the girl climbed tree* with 'the' occurring twice.

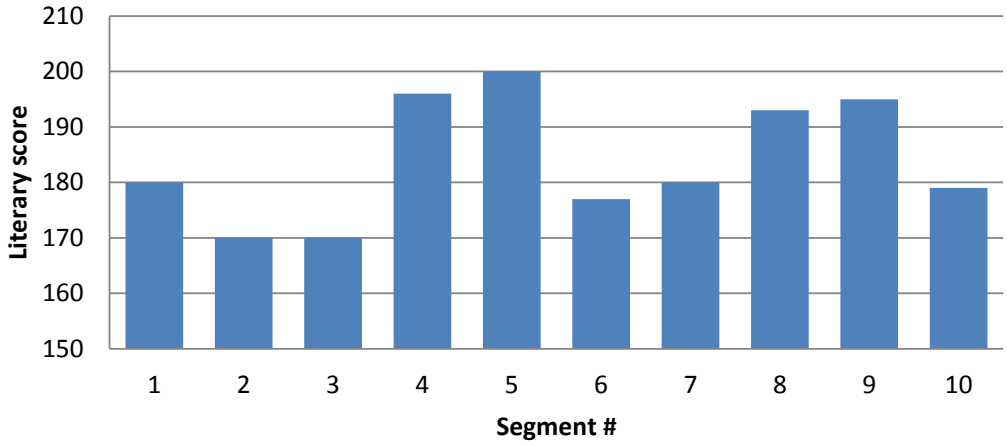


Figure 6.1: Literary score of each segment

Each segment was subjected to POS analysis using the NLTK and the POS showing the greatest correlation are shown in Table 6.2.

Experiments were carried out to determine the most efficient combination of POS features. Combining the qualifying features produced an eventual model that closely reflected the human survey results. Figure 6.2 shows the results of combining function words. A noticeable exception to the pattern was segment 9 which spiked higher than expected throughout many of the experiments, due in part to the high number of function words. It is of interest to note that this segment was considerably shorter than the others (214 words against an average word count of 683), suggesting that the percentage of function words is necessarily higher in a shorter segment.

Comparing translated results

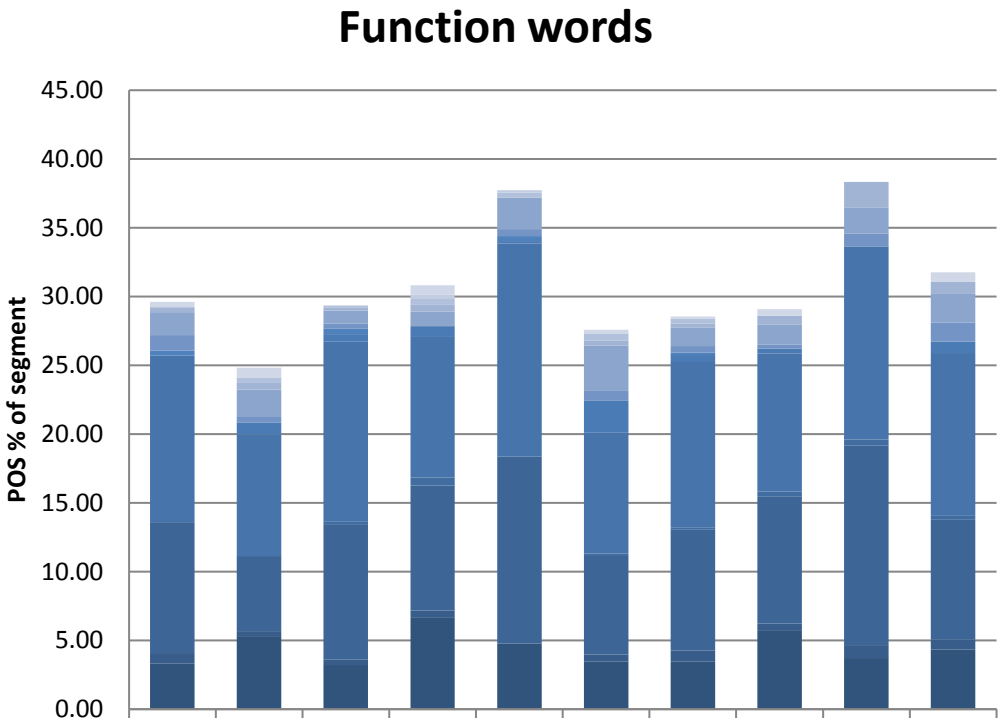
To examine this phenomenon more closely, the same texts were used but were subjected to the translation/re-translation process from the earlier study. As this was not the main focus of the study, for speed and simplicity only Norwegian and Catalan, the highest and lowest similarity languages for prose, respectively, were tried.

The function word spike was repeated in both languages, confirming the suspicion that a high ratio of function words is a facet of a shorter text. As in the previous study, the results included some native words that were not

Table 6.2: POS found to correlate to the human response to the text segments

Feature	Description
average sentence length (AvSentLen)	This has an impact on the rhythm of the text. Factual information is usually provided in short sentences, while news articles and advertising are sometimes delivered in a virtual staccato. Literature allows and encourages a lengthier sentence structure, so this was expected to be a strong indicator of literary quality.
lexical diversity (LexDiv)	Using the formula proscribed by Gonçalves and Gonçalves (2006), the ratio between word occurrence (hapax legomenon) and the total word count was calculated and applied to each segment.
CC (coordinating conjunctions)	These are words that combine two clauses. Examples are 'and', 'but', 'nor' and 'so'. As already observed, literary texts tend to be longer than non-literary ones, so the inclusion of conjunctions that create compound sentences is not surprising.
CD (cardinal number)	This was not an expected POS, but including it improved the overall accuracy.
DT (determiner)	Determiners reference the noun in a phrase and examples are 'the', 'a', 'my', 'some' and 'that'. A higher proportion indicates the existence of complex (multiple clause) sentences.
EX (existential 'there')	An instance of the word 'there' without a locative context. In an expression such as 'There is a place over there', the first 'there' is an EX. It will frequently occur in a descriptive context, and hence was an anticipated POS.
IN (preposition or subordinating conjunction)	An expression that introduces or a phrase, or a conjunction that introduces a dependent clause. Examples are 'if', 'because' and 'while'. This POS is indicative of a complex sentence.
JJ (adjectives)	As literature tends to be descriptive, this POS was fully anticipated.
NN (nouns)	Nouns were not anticipated in the framework. A news article or non-fiction text would contain a high proportion of nouns due to the factual nature; a literary text often contains more conceptual themes.
PRP$ (possessive pronoun)	This was also an unexpected POS, but inclusion improved the accuracy of the framework.
RB (adverb)	As with adjectives, and for the same reasons, this POS was expected.
VBN (verb, past participle)	It was anticipated that verbs would form part of the framework. However, including all variety of verbs proved unsuccessful. Because most literature is written from the point of view of things that happened (real or imaginary) in the past, this accounts for the appearance of this POS.

translated back into English and there were some nonsensical words that were an expected result of the process. However, the NLTK tagger did not pick these up as foreign words (FW in POS terms) as expected, tagging them instead as nouns (NN). This does not account for a spike in adjectives in segment 3 in the Catalan text which showed an unexpected jump of 3 per cent. Adjective results, however, remained unaffected.



	1	2	3	4	5	6	7	8	9	10
WRB	0.37	0.73	0.00	0.72	0.00	0.26	0.17	0.48	0.00	0.71
WP$	0.00	0.00	0.00	0.24	0.18	0.00	0.00	0.00	0.00	0.00
WP	0.00	0.36	0.19	0.48	0.35	0.52	0.33	0.00	0.00	0.00
WDT	0.37	0.48	0.19	0.48	0.00	0.39	0.33	0.64	1.87	0.82
TO	1.68	1.94	0.95	1.08	2.29	3.22	1.32	1.44	1.87	2.12
RP	1.12	0.48	0.38	0.00	0.53	0.77	0.50	0.32	0.93	1.41
POS	0.37	0.00	0.38	0.00	0.35	0.00	0.17	0.00	0.00	0.00
MD	0.00	0.85	0.57	0.72	0.18	2.32	0.50	0.32	0.00	0.82
IN	12.10	8.84	13.07	10.27	15.49	8.76	12.05	10.06	14.02	11.76
EX	0.00	0.00	0.19	0.60	0.00	0.13	0.17	0.32	0.47	0.35
DT	9.50	5.45	9.85	9.08	13.56	7.22	8.75	9.27	14.49	8.71
CD	0.74	0.36	0.38	0.48	0.00	0.52	0.83	0.48	0.93	0.71
CC	3.35	5.33	3.22	6.69	4.80	3.48	3.47	5.75	3.74	4.35

Figure 6.2: Percentage of function words

Apart from the adjective anomaly, the translated texts closely matched the untranslated versions, suggesting that the framework had potential as a tool. Although the small number of samples used meant there was a danger of

over-fitting the framework, the study demonstrated the feasibility of using this method to create a more complex framework to determine deeper stylistic indicators of literary quality.

6.2 Tools refinement

Readability scores were investigated as a way to qualify texts, an approach used by Ashok et al. (2013) that found that higher scores suggest a more literary work. There are three main tests: Gunning's FOG (Ashok et al., 2013; Afroz et al., 2012), the Flesch-Kinaid (Ashok et al., 2013; Afroz et al., 2012; Luyckx et al., 2006) and the SMOG[1] index (Aliu and Chung, 2010). These scores are frequently used to determine the reading level demanded of a reader by a text. The Flesch-Kincaid is widely used (it is the 'Readability Statistics' option used in Microsoft's Office products) using the following formula:

$$RE = 206.835 - (1.015 \times AVL) - (84.6 \times AVNS)$$

where RE is reading ease, AVL is the average sentence length and AVNS is the average number of syllables per word. The Flesch-Kinaid was tested against various texts, but it was found that although useful for determining whether a text is fiction or non-fiction (scores below 60 suggest non-fiction), there was little difference between the fiction texts. Similarly, Gunning's FOG and the SMOG indices showed large differences between fiction and non-fiction, with non-fiction texts scoring greater than 11 for the Gunning's FOG and greater than 9 for the SMOG, but there were inconsistent differences between fiction texts. Consequently, these tools were abandoned.

However, relative entropy (RelEnt) was included as a variable instead. This has been used effectively by Kan and Gero (2009) using a program written by Torres (2002) to determine the literary quality of songs and poems.

[1]Simple Measure of Gobbledygook

6.2.1 Factor analysis

To discover the factors that are most important in a book's popularity, the 100 most downloaded books were taken from The Gutenberg Project[2], an online resource of over 50,000 free ebooks that have been previously published by traditional means and are out of copyright. Using Gutenberg download counts as an indication of literary worth has been an effective measurement used in previous studies (Ashok et al., 2013). Plays, poetry and non-fiction (apart from one biography that has a story-like format) were discarded, leaving 75 books. A further 25 books were chosen that had multiple downloads (more than 200) but were not included in the top 100 to bring the total number of books to 100. Not only did this make a pleasing round number but it meant that a selection of texts were included that were not necessarily literary but had been deemed by a publisher to be of sufficient merit for investment.

Because of the large number of variables involved (those listed in Table 6.1 plus alliteration, average sentence length, lexical diversity, text entropy, relative entropy), principal component analysis was carried out to identify any correlation between them and reduce the number of observations. A scree plot can visually show how many factors are responsible for most of the variability by displaying the factors along the x-axis (in this case, 37 variables) and the calculated eigenvalues (the value of a vector whose direction remains the same even when a linear transformation is applied) along the y-axis. Those factors which form the "cliff face" i.e. that have a high eigenvalue are those variable combinations that are significant while factors that show a low eigenvalue are less important. The scree plot in Figure 6.3 shows the levelling off is at either six or nine principal components; however, the first six only account for 66 per cent of the variance and the nine for only 77 per cent (Table 6.3).

Ideally, three or four principal components would account for a much higher proportion of the variability, suggesting that there is not a great deal of opportunity to reduce the number of variables.

[2]http://www.gutenberg.org

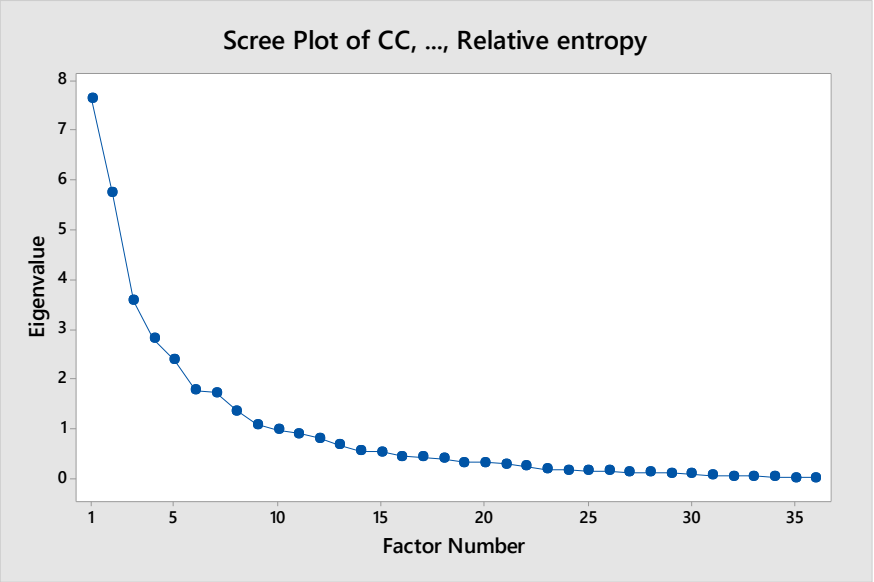


Figure 6.3: Scree plot indicating up to nine principal components

Table 6.3: Eigenanalysis of the correlation matrix with the cumulative variances at six and nine principal components in bold

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Eigenvalue	7.7312	5.7871	3.6815	2.9110	2.3941	1.9152	1.7259	1.3825	1.1139
Proportion	0.209	0.156	0.100	0.079	0.065	0.052	0.047	0.037	0.030
Cumulative	0.209	0.365	0.465	0.544	0.608	**0.66**	0.707	0.744	**0.774**

Loading plot

The loading plot shows how each variable influences each component so in Figure 6.4 NNS (circled in red) has a high negative eigenvalue in both components whereas PRP (circled in blue) has a high positive eigenvalue in the first component but a low negative in the second component. Lines that go in the same direction and are close to each other suggest that the factors are correlated. Although there are up to nine factors, only the first two (accounting for 36.5 per cent of the variation) are examined. Some of these groupings have obvious correlation: VB and TO, for example, form one group (circled in green) which is explained by use of the infinitive (e.g. 'to go', ' to be') and the group containing JJS, WDT, RBS, VBN and RBR (circled in orange) can be loosely described as 'descriptive' tools although JJ and JJR are less closely correlated to this group than expected. WRB and PDT (circled in yellow) are explaining and indicator words (e.g. 'how', 'however', 'whereby' and 'all', 'both', 'this'). Average sentence length and IN (circled in pink)

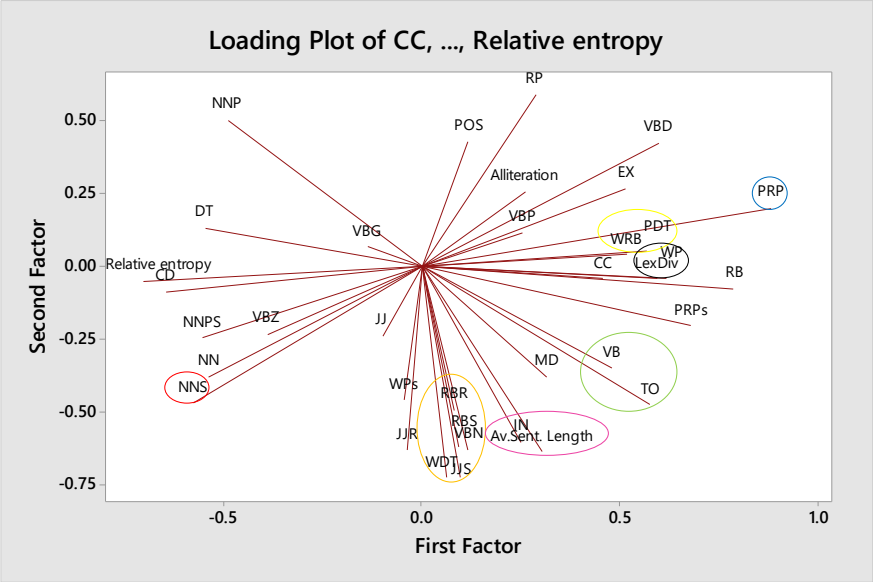


Figure 6.4: Loading plot with grouping

correlate because a subordinating preposition or conjunction (e.g. 'despite', 'like', 'until') can be used to join clauses, making one longer sentence where two shorter ones might be used. Other groupings such as LexDiv and WP (circled in black) are not intuitively clear.

Factor analysis allows examination of the data structure by showing correlations between variables. Some grouping was anticipated but not seen in the results, such as a clustering of verbs in their various forms. Instead, these are scattered across the range. It was thought that this may be due to the range of literary quality; a "bad" book may have too many verbs if the author is clumsily trying to create a sense of action, or too few if there is little plot movement. To find out if different factors affect books at either end of the quality range and cause there to be less grouping than anticipated, the data were divided into two parts: the top 50 and the bottom 50 ranked but although there was a little movement between groupings, the POS groups remained fairly consistent.

Score plot

A score plot visually projects the raw data onto the loading plot, giving a good indication of the degree to which a sample relates to the various components. The expectation in this case was that "good" books would be

clustered together around the 0:0 axes with the "bad" texts scattered further afield. Examination of the score plot indicates that stylistic tendencies are identifiable, as shown in Figure 6.5

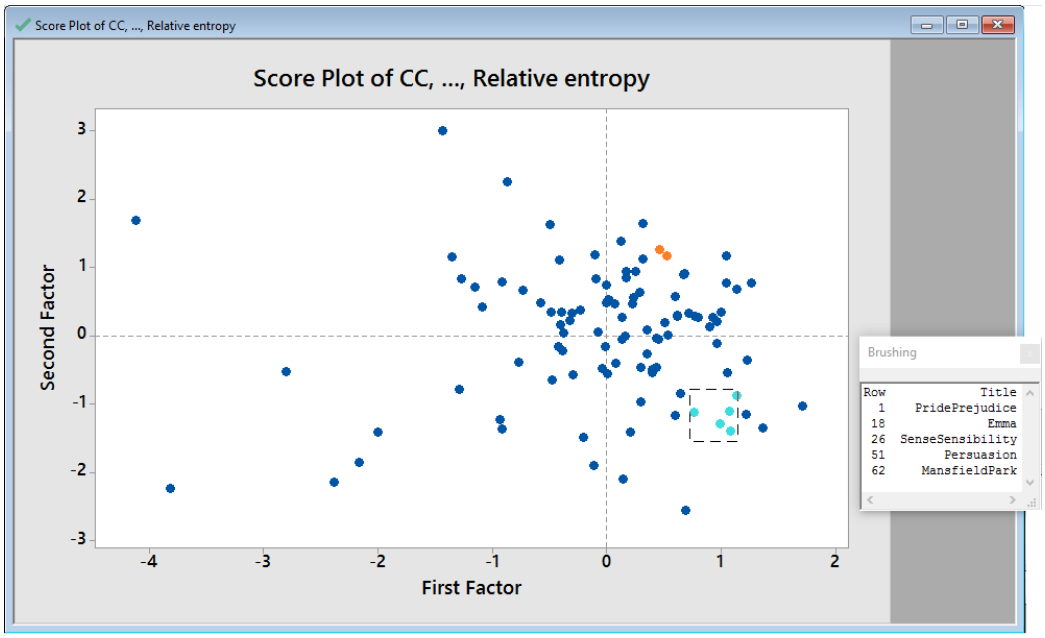


Figure 6.5: Score plot showing grouping of Austen novels (lighter blue dots) and Carroll novels (orange dots)

which indicates the Austen novels grouped together. The two Carroll texts are similarly clustered. However, the expected grouping of books according to their download ranking is not evidenced as clearly as expected.

This result was replicated when using the data split between the top and bottom ranked texts, with the Austen novels (three in the top 50 and two in the bottom 50) still bunched closely together, confirming that the process is valid. To ensure this, a number of non-fiction works were added to the collection; these ranged from news articles to instruction manuals. The score plot in Figure 6.6 shows the clear difference between the fiction and non-fiction texts. The Corsa text (a car manual, indicated by a pink dot) includes many tables and other numerical data that explain its isolation from the other non-fiction texts. As the clustering was clearly picking up stylistic traits, it was important to determine whether using the Gutenberg Project downloads as an indicator of literary quality is an inadequate measure.

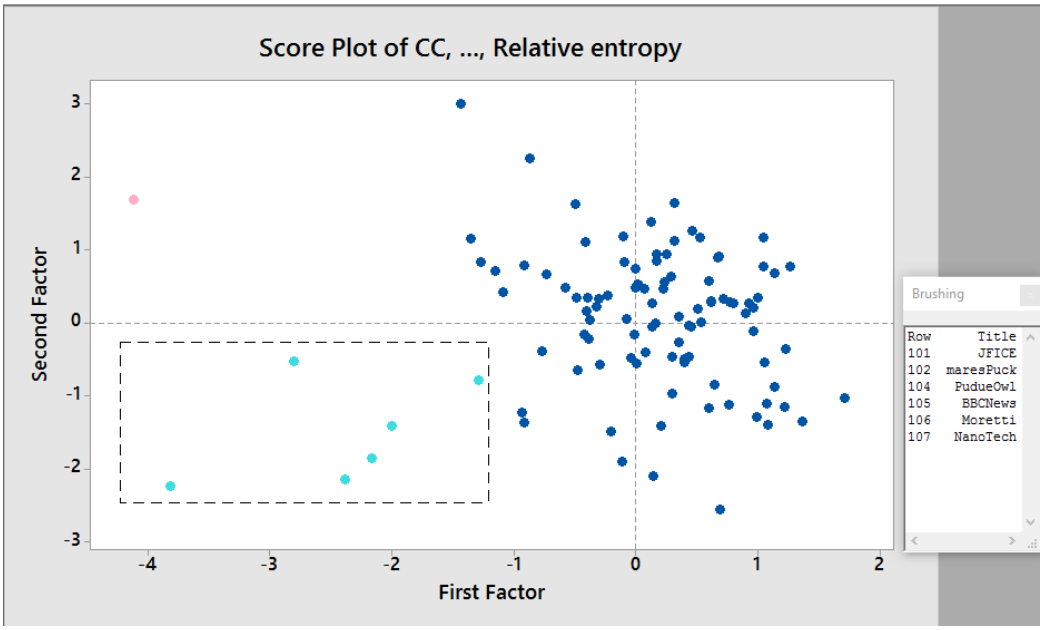


Figure 6.6: Score plot showing clear grouping of non-fiction works (lighter blue dots)

Human correlation

Consequently, a panel of seven literature graduate human experts were used to rank the fiction texts manually. Understandably, the panel members were not familiar with every text, particularly those novels with the fewest downloads which were by their very nature the least popular books. Nineteen of the books had not been read by any of the respondents; unfamiliar books were marked as N/R by participants so that books that had not been read by all of them were not penalised.

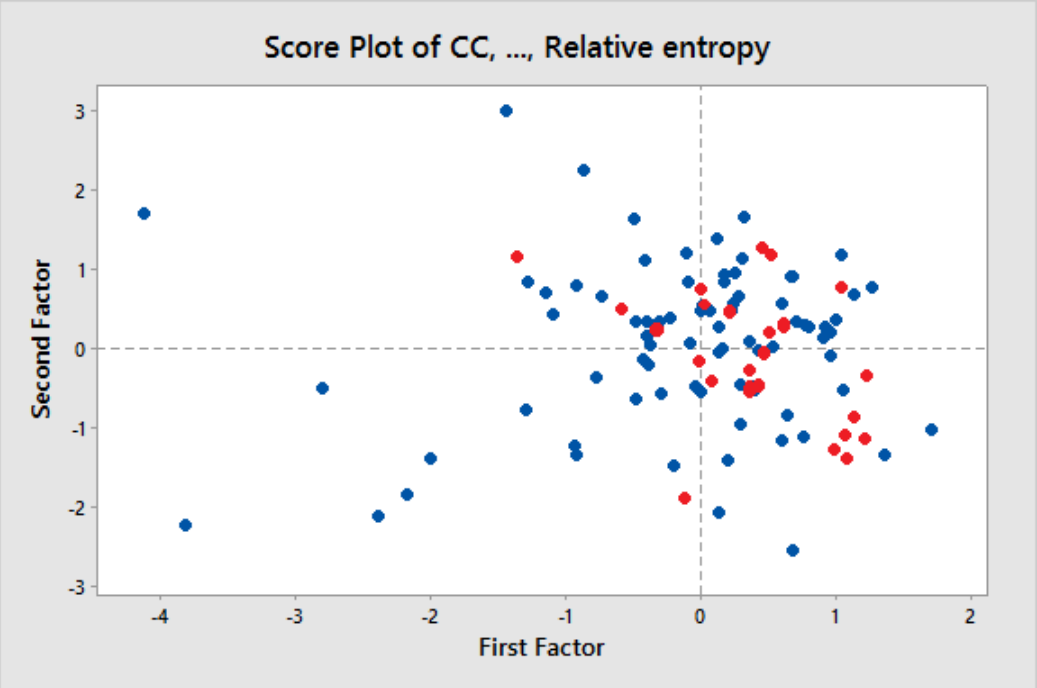


Figure 6.7: Score plot of first and second factors with the top 25 novels ranked by the human experts indicated by red dots

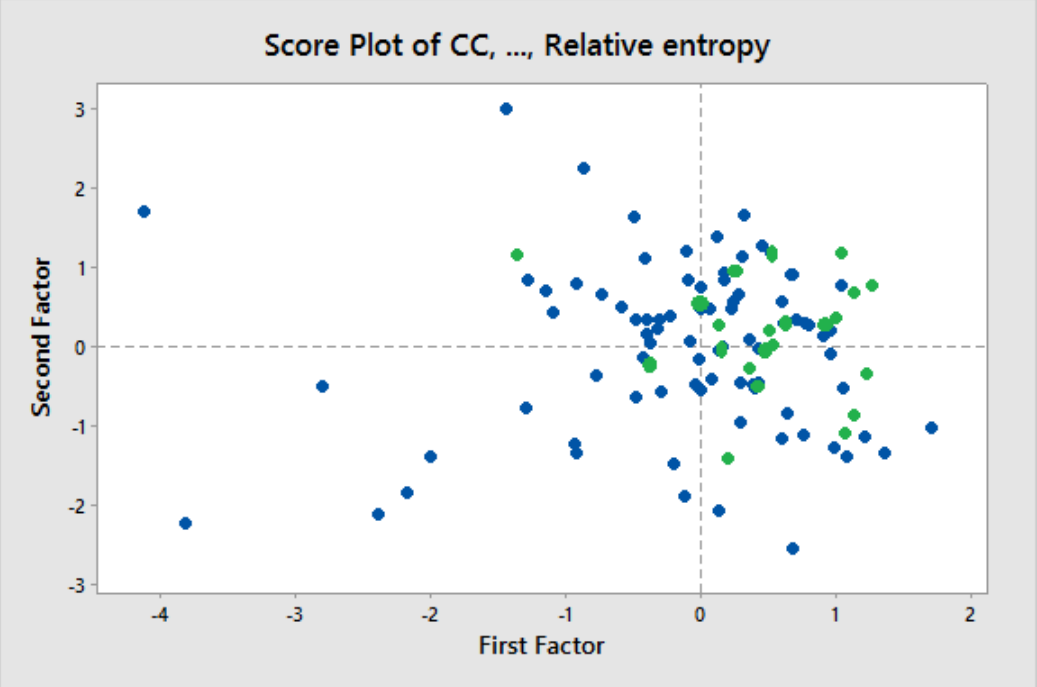


Figure 6.8: Score plot of first and second factors with the top 25 novels ranked by Gutenberg download indicated by green dots

Figures 6.7 and 6.8 show the placement of the top 25 novels according to the human panel and the Gutenberg downloads, respectively. These show that the choices made by the human panel and the ranking according to the Gutenberg Project's downloads are closely correlated, indicating that the number of downloads is a good indicator of literary quality and thereby confirming the findings of Ashok et al. (2013).

6.3 Feature selection

The factor analysis from Section 6.2.1 was used to confirm the most influential literary features. Tables 6.4 and 6.5 show the most significant loadings from Factors 1 and 2 (see Appendix F for full table of the first eight factors).

Table 6.4: Features with the greatest significance from the first factor

Variable	Factor 1	Variable	Factor 1
CD	-0.645	NN	-0.538
NNS	-0.575	NNPS	-0.552
PDT	0.566	PRP	0.881
PRP$	0.678	RB	0.787
TO	0.576	VBD	0.597
WP	0.616	WRB	0.516
LexDiv	0.593	RelEnt	-0.701

Table 6.5: Features with the greatest significance from the second factor

Variable	Factor 2	Variable	Factor 2
IN	-0.603	JJR	-0.633
JJS	-0.727	NNP	0.501
RBS	-0.621	RP	0.592
VBN	-0.630	WDT	-0.727
AvSentLen	-0.637		

6.3.1 Scoring the chosen variables

Although the significant variables were identified, some sort of scoring system was still required in order to grade the texts according to literary merit. To

facilitate this, the texts were sorted according to the number of Gutenberg Project downloads and graded into five categories: fiction texts were divided into four equal sections from Grade 1 to Grade 4 and non-fiction was added as an additional category. This division is not arbitrary; recall in Section 6.2.1 that 25 novels that were not part of the original top 100 were added to provide examples of lower quality texts so it is logical to divide the samples evenly. Dividing the sample into fewer groups would mean mixing "bad" texts with "good" ones. Experiments were carried out using finer subdivisions (i.e. more categories) but this did not improve the results and so this approach was discarded.

Each literary feature variable was averaged across the different grades (Table 6.6) and averages for each POS were calculated per grade and compared (Table 6.7).

Table 6.6: Average per grade of each literary feature. Figures are the percentage of text comprising alliteration, the calculated scores for LexDiv and RelEnt and the average sentence length for AvSentLen.

POS feature	Grade 1	Grade 2	Grade 3	Grade 4	Non-fiction
Alliteration	5.35	5.54	5.54	5.60	4.82
LexDiv	17.69	16.33	14.48	13.00	6.52
RelEnt	56.08	56.76	58.80	59.88	71.00
AvSentLen	23.12	25.03	26.45	20.31	24.26

In Table 6.7, the features that show a consistent difference between grades of fiction are shown in blue rows so, as an example, CC demonstrates a distinct trend with the percentage of CC decreasing as the texts become less literary. This then indicates whether a text containing a particular percentage of CC should be classified as literary or not for this specific POS.

Table 6.7: Average per grade of each literary feature. Figures are the percentage of the text each POS comprises

POS feature	Grade 1	Grade 2	Grade 3	Grade 4	Non-fiction
CC	4.48	4.41	4.36	3.95	3.03
CD	0.65	0.72	0.58	0.73	2.42
DT	9.19	9.59	9.72	9.37	15.24
EX	0.22	0.19	0.18	0.18	0.09
IN	11.52	12.04	12.31	11.39	10.59
JJ	5.06	5.39	5.76	5.54	5.59
JJR	0.24	0.24	0.27	0.23	0.34
JJS	0.19	0.20	0.20	0.18	0.21

POS feature	Grade 1	Grade 2	Grade 3	Grade 4	Non-fiction
MD	1.63	1.49	1.39	1.30	1.79
NN	12.65	12.56	13.56	12.56	14.78
NNP	11.71	11.71	11.15	14.90	15.81
NNS	2.95	3.22	3.17	3.03	5.53
NNPS	0.02	0.02	0.03	0.03	0.18
PDT	0.04	0.03	0.04	0.03	0.02
POS	0.49	0.39	0.63	0.65	0.23
PRP	8.05	7.45	6.79	7.19	2.48
PRP$	2.70	2.58	2.79	2.36	0.82
RB	5.60	5.37	5.23	4.88	3.06
RBR	0.13	0.12	0.14	0.11	0.18
RBS	0.04	0.04	0.04	0.03	0.04
RP	0.53	0.54	0.53	0.55	0.23
TO	2.67	2.64	2.71	2.57	2.43
VB	3.70	3.47	3.34	3.22	3.60
VBD	6.63	6.87	6.80	6.82	1.63
VBG	1.47	1.61	1.70	1.54	1.57
VBN	2.62	2.63	2.75	2.34	2.64
VBP	1.90	1.69	1.39	1.81	1.75
VBZ	1.34	1.16	0.91	1.06	2.46
WDT	0.50	0.54	0.56	0.46	0.58
WP	0.55	0.53	0.48	0.49	0.22
WP$	0.03	0.02	0.02	0.01	0.03
WRB	0.54	0.52	0.46	0.49	0.43

6.3.2 Observations on the chosen variables and their relationship to human preferences

Chapter 5 investigated the human reaction to literature and found that Plot, Theme and Description (and to a lesser extent, characterisation) are the main factors that mark out a novel as literary but the CoBAALT model is built through a less subjective approach by using factor analysis to determine the relevant variables. However, there are correlations between the human and computational choices, as discussed in the following sections and summarised in Table 6.8.

Table 6.8: Variables identified by factor analysis and their relation to human judgement

POS feature	Human choice
AvSentLen	Description
LexDiv	Description
RelEnt	Description
CD	Description
IN	Plot
JJR/JJS	Description
NN	Plot
NNP/NNS/NNPS	Theme
PDT	Description
PRP/PRP$	Theme
RB/RBS	Description
RP	Description
TO	Description
VBD/VBN	Plot and Theme
WDT/ WP/WRB	Theme and Description

Style features: Average sentence length, lexical diversity and relative entropy are not POS but specific style features. The average sentence length gives an indication of literary merit as shorter sentences suggest lower grade fiction or non-fiction; however, there is a danger of the narrative becoming lost if the sentences are too convoluted. Table 6.6 shows that Grade 2 and 3 books both have longer average sentence lengths than Grade 1, suggesting that "better" books are more tightly written and less likely to meander off in a purple haze of prose; this relates to the human choice of Description.

The LexDiv score indicates the richness of the vocabulary used by calculating how often words are used in the text; a higher score suggests a wider range of words used and hence a more literary work. RelEnt calculates the relative entropy of the text. A text that contains no repeated words would have 100 per cent entropy so it, too, is measuring the repeated use of words. Paradoxically, LexDiv and RelEnt scores trend in opposite directions: a high LexDiv indicates a "good" text yet a high RelEnt score suggests a "poor" text. This is because repetition is a highly effective literary device that incorporates anaphora, epistrophe and symploce (repetition of words at the beginning, end and both beginning and end of a clause, respectively) along with leitmotifs and repetition for emphasis, so some degree of repetition is highly desirable (Wales, 1990, pp. 402-403). Wales gives *Finnegans Wake* as

an example of an entropic novel, declaring it to be 'largely unread' as a result so it appears that a significant degree of word re-occurrence is desirable in a literary text.

CD: Cardinal numbers were not anticipated as an indicative POS nor suggested by the investigations in Chapter 5 but the result replicates the experience of the previous experiments in Chapter 6 (Table 6.2). A high preponderance in non-fiction is to be expected given that the example texts include instruction manuals and tables, but lower grade texts consistently have more CD than the Grade 1 works. It is suggested that this is the result of overzealous application of detail, of ignoring the writers' golden rule of 'show, don't tell' (Dynes, 2014, Chapter 19) and is a Description feature. A writer can provide a fuller description by adding a cardinal number (by writing 'five houses' rather than just 'some houses', for example).

IN: Prepositions and subordinating conjunctions serve to explain settings and move the story along as part of a narrative (Wales, 1990, p. 372) and this correlates to the human choice of Plot as an important feature.

JJR and JJS: These are comparative and superlative adjectives, respectively. Children are taught to add adjectives to their early creative writing attempts; unfortunately this is a lesson that is difficult for new writers to forget and writing courses must work hard to break a new writer's habit of throwing in adjective upon adjective (Dynes, 2014, Chapter 37). However, comparative and superlative adjectives can enhance descriptive text, as preferred by the focus groups and indicated as Description, and they appear to be less prone to excessive distribution.

NN: Nouns are associated with less literary texts, suggesting that a narrative is more concerned with verbs (action) than things which correlates to the human choice of Plot as an important feature although nouns also provide indicators of Theme.

NNP, NNS and NNPS: These are plural nouns and both singular and plural proper nouns, POS that point to Theme. These POS are found less frequently in the "good" books (although this is not consistent across all the grades) and relate mainly to character names. Too many characters or places can confuse a reader, a lesson never learnt by James Joyce. *Ulysses* contains 21 per cent NNP and *The Dubliners* almost 16 per cent. Of his novels included in the texts tested, *A Portrait of the Artist as a Young Man* has the lowest rate at just under 15 per cent, which may go towards explaining why

Joyce is a challenging read, '[pushing] language and linguistic experiment...to the extreme limits of communication' (Drabble, 1996, p. 528). Common plural nouns (NNS) are similarly found less frequently in the higher graded texts.

PDT: Pre-determiners are descriptive words that refine the noun reference in terms of quantity. Examples include 'all', 'half', 'quite'. As such they are used to elaborate a word picture as mentioned by the focus group participants, much in the way of adjectives (and therefore indicated as Description) but with less danger of over-use as they are function words, i.e. words with grammatical rather than lexical meaning (Wales, 1990, p. 199).

PRP and PRP$: Personal and possessive pronouns relate to people (characters) and so are anticipated POS. The research carried out in Chapter 5 suggested that characterisation is not as important to literary merit as other facets but it would have been surprising if there were no variables that relate to character in an investigation into fiction. As such, they are indicators of Theme.

RB and RBS: Adverbs (RB) are usually marked by the suffix *-ly* and are used to modify verbs. It is interesting that superlative adverbs are included but comparatives are not. These POS help to create the word pictures desired by the focus groups (Description).

RP: Particles are function words that have little lexical meaning on their own but add to the understanding of a noun phrase and as such are identified with Description.

TO functions both as a preposition and as part of the infinitive form of a verb. As such, it can be used to create adjectival and adverbial phrases by modifying the noun (e.g. 'it's good to talk') and verb ('I've had enough to eat'), respectively, thereby enhancing Description.

VBD and VBN: Verbs mean action and this in turn propels a story onwards in the form of Plot and Theme. Some verb forms were anticipated but not all variations are included as significant factors. VBD, for example, is the past tense form and was therefore anticipated as a POS to figure higher in literary text as most stories are told in the past tense, and this was a feature indicated by the factor analysis. In fact, the Grade 1 texts have the lowest of all the fiction grades. This is accounted for by the aforementioned 'show, don't tell' mantra (Dynes, 2014, Chapter 19) that is neglected by the less literary writer or by more simplistic stories written for younger readers. Gerunds (words

ending is 'ing' and indicated as VBG), for example, are found less in the literary texts than in non-fiction but Grade 2 texts actually contained the fewest instances of VBG.

WDT, WP and WRB: Pronouns and possessive pronouns relate to characters and are therefore found more in the higher grade texts while wh-adverbs help to build descriptions and provide explanations. This suggests that they satisfy the human demand for both Theme and Description.

Not all of the features indicated by the factor analysis show consistent trends across the different grades of text. In such cases, the trend is taken as the difference between Grade 1 and non-fiction.

One such inconsistent variable is VBD which is indicated as significant by the first factor analysis but this feature does not appear in blue in Table 6.7 where the average scores for graded texts are compared. The VBD (verb, past tense) anomaly is interesting. Here, the "good" novels show a lower percentage of this feature whereas the "less literary" texts have a higher percentage, yet non-fiction contains hardly any. It is logical that non-fiction contains less because stories are mainly told in the past whereas non-fiction (news, manuals, articles) are more likely to use present tense. Closer examination revealed that the fluctuation is due to specific novels containing a high percentage of VBD, which hiked the averages. Most of these were found to be stories for children. Wales observes that there are multiple shifts in temporal perspective within novels and cites *David Copperfield* as a specific example, 'Whether I shall turn out to be the hero of my own life, or whether that station will be held by anybody else, these pages must show', referring to a future outside the temporal reference of the novel, but she specifically indicates folk and fairy tales as being exceptions to this trend: precisely the types of story that caused the VBD anomaly (Wales, 1990, p. 458).

6.4 Summary

This chapter has outlined the steps taken to identify the features that can be combined to associate specific stylistic traits that are common in Classics. A human panel of experts identified ten passages from two separate books that they deemed to be particularly literary. These passages were then passed to an online survey for the general public to see which texts they thought were the most literary. The NLTK was used to break down the texts into

their component POS and to determine their lexical diversity; these stylistic entities were then tested against the survey's results to see where there was any correlation.

As the results of the above experiment showed that variables did correlate to the results of the online survey, a larger text sample was used to discover exactly which variables were influential in determining literary merit. To this end, 100 books were downloaded from the Gutenberg Project website. These comprised the top 75 downloaded works of fiction plus a further 25 books with more than 200 downloads each. Seven non-fiction texts were later added to see if there was a difference between fiction and non-fiction. Factor analysis was used to identify the variables with most influence on the literary qualities. Four grades of fiction and one of non-fiction were categorised and the averages of each grade for the variables selected by the factor analysis were calculated to see whether the presence of the variable had a positive or a negative effect on the literary merit.

With the relevant POS identified along with the other literary variables, progress can now be made on a conceptual model, given in the following chapter, that is able to determine the literary merit of a given text.

Chapter 7

CoBAALT: a Computer-Based Aesthetic Analysis of Literary Texts

This chapter gives the final selection of variables that allow the model, called CoBAALT, to judge a text for its literary merit. The model is tested in Sections 7.3.2 and 7.3.3 against two authors that are deemed by critics to be literary and a discussion of the findings follows.

CoBAALT is the result of the research carried out in the preceding chapters of this thesis. The literature review in Chapter 2 suggests that tools more frequently found in authorship attribution can be adapted to determine a stylistic map of literariness. The feasibility of this approach is tested in Chapter 4 by ensuring that texts can be parsed without 100 % accuracy and still retain their literary qualities to a measurable degree. The results suggested that a stylistic analysis was a feasible approach.

Chapter 5 serves to investigate the human perceptions of "good" literature that inform the understanding of *why* the selected variables are relevant to literary merit. Conducting focus groups found that description and the use of language are important considerations when deciding whether a book is literary and this fact was confirmed by using an online survey that was open to the general reading public.

Chapter 6 shows that the style of writing has a considerable impact on the reading experience and qualification of a book as literary. Some of the vari-

ables identified in this experiment were counter-intuitive to expectations so a decision was made to use computational analysis rather than to rely on the subjective choices of a human panel to identify relevant variables to use in the CoBAALT model. To this end, factor analysis was used to identify the most relevant literary features that constitute "good" literature and to create a grading system for these variables.

7.1 The CoBAALT model

From the experiments carried out in Chapter 6, the features from Tables 6.4 and 6.5 were identified by factor analysis as the variables that indicate literary quality. These were then categorised into four grades of fiction and one of non-fiction and each grade was averaged across all the fiction and non-fiction texts (Table 7.1). This grading indicates whether the presence of the variable has a positive or a negative effect on the literary merit of the text. Not all of the variables show a consistent trend; then, the trend is read as the difference between Grade 1 and non-fiction categories.

As an example, Table 7.1 shows that the baseline for the first variable, AvSentLen, is 23.12. This is the average percentage of this POS over the top 25 texts and the table shows how instances of this variable gradually increase across the first three grades of text. Grade 4 has a lower score and it then increases with non-fiction (23.12, 25.03, 26.45, 20.31 and 24.26, respectively) so here the tendency is taken between the Grade 1 and non-fiction texts, indicating that a lower AvSentLen is a more desirable feature for literary merit. Therefore, a text which contains 25.00 per cent of AvSentLen would score negatively (-1.88) because it is 1.88 from the baseline and is trending away from the "better" texts.

The CoBAALT model uses a toolbox of techniques and approaches examined in this thesis, according to the literary criteria given in Table 7.2. The variables are those identified by the factor analysis in the previous chapter (Tables 6.4 and 6.5). The baseline figures are the Grade 1 averages from Table 7.1.

Table 7.1: Average per grade of the variables selected by factor analysis. Grade 1 texts provide the baseline figure. The directional arrows indicate whether the trend is for a higher (↑) or a lower (↓) percentage to suggest literary quality.

Feature	Grade 1	Grade 2	Grade 3	Grade 4	Non-fiction	Literary (↑ *or* ↓)
AvSentLen	23.12	25.03	26.45	20.31	24.26	↓
LexDiv	17.69	16.33	14.48	13.00	6.52	↑
RelEnt	56.08	56.76	58.80	59.88	71.00	↓
CD	0.65	0.72	0.58	0.73	2.42	↓
IN	11.52	12.04	12.31	11.39	10.59	↑
JJR	0.24	0.24	0.27	0.23	0.34	↓
JJS	0.19	0.20	0.20	0.18	0.21	↓
NN	12.65	12.56	13.56	12.56	14.78	↓
NNP	11.71	11.71	11.15	14.90	15.81	↓
NNS	2.95	3.22	3.17	3.03	5.53	↓
NNPS	0.02	0.02	0.03	0.03	0.18	↑
PDT	0.04	0.03	0.04	0.03	0.02	↑
PRP	8.05	7.45	6.79	7.19	2.48	↑
PRP$	2.70	2.58	2.79	2.36	0.82	↑
RB	5.60	5.37	5.23	4.88	3.06	↑
RBS	0.04	0.04	0.04	0.03	0.04	↑
RP	0.53	0.54	0.53	0.55	0.23	↑
TO	2.67	2.64	2.71	2.57	2.43	↑
VBD	6.63	6.87	6.80	6.82	1.63	↑
VBN	2.62	2.63	2.75	2.34	2.64	↓
WDT	0.50	0.54	0.56	0.46	0.58	↓
WP	0.55	0.53	0.48	0.49	0.22	↑
WRB	0.54	0.52	0.46	0.49	0.43	↑

7.2 Implementation

The CoBAALT model is a collection of procedures that scores the output against the matrix of variables identified in Table 7.2. A video demonstration of CoBAALT is available at https://youtu.be/nInXEEx04hc.

Table 7.2: Features included in the literary criteria with their baseline figures. The directional arrows indicate whether a high proportion of this feature indicates literariness ↑ or whether a lower percentage is required ↓.

Feature	Baseline	Feature	Baseline	Feature	Baseline
AvSentLen	23.12 ↓	LexDiv	17.69 ↑	RelEnt	56.08 ↓
CD	0.65 ↓	IN	11.52 ↑	JJR	0.24 ↓
JJS	0.19 ↓	NN	12.65 ↓	NNP	11.71 ↓
NNS	2.95 ↓	NNPS	0.02 ↑	PDT	0.04 ↑
PRP	8.05 ↑	PRP$	2.70 ↑	RB	5.60 ↑
RBS	0.04 ↑	RP	0.53 ↑	TO	2.67 ↑
VBD	6.63 ↑	VBN	2.62 ↓	WDT	0.50 ↓
WP	0.55 ↑	WRB	0.54 ↑		

7.2.1 System architecture

This section outlines the system architecture that was used for the design and implementation of CoBAALT. The processes were carried out on an Acer Aspire One running Windows 10 Home edition.

Hardware

- Processor: Intel Pentium CPU 967 @ 1.30 GHz
- RAM: 4.0 GB
- System type: 64-bit operating system

Software

- Python 2.7.10 for win32
- NLTK 2.0 including optional NumPy and Matplotlib packages
- Code::Blocks version 16.01 rev 10702 SDK 1.29.0

7.2.2 Processes

Figure 7.1 shows a schematic of the CoBAALT process whereby the text is processed through a series of parsing processes to extract the following

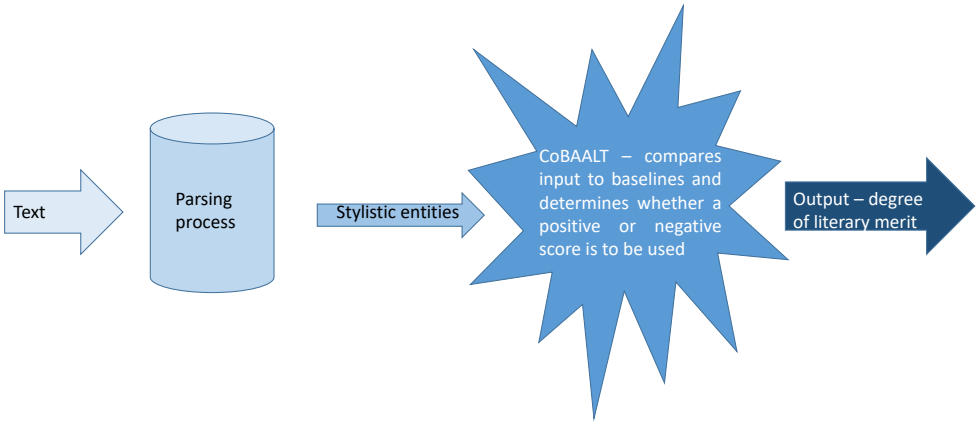


Figure 7.1: The CoBAALT process

stylistic entities:

- Using Code::Blocks, relative entropy is calculated using the formula given by Kan and Gero (2009) (the C code is given in Appendix D). This process determines the entropy of the text using the formula $H_{rel} = \frac{H_T}{H_{max}} \times 100$ where the relative entropy H_{rel} is the quotient between the text entropy H_T and the maximum entropy H_{max} multiplied by 100 to obtain a percentage. Maximum entropy would occur if all the words in the text were unique.

  Example output for this is shown in Figure 7.2.

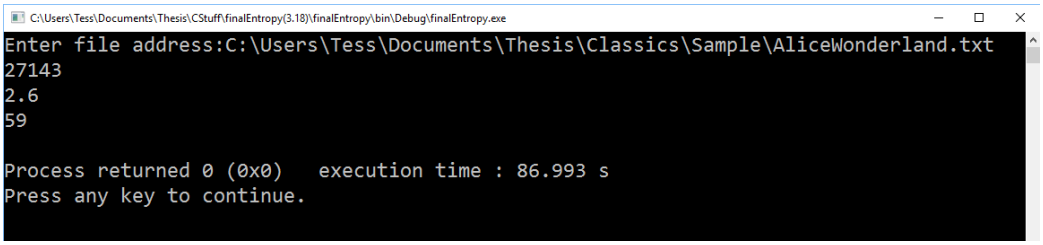


Figure 7.2: Relative entropy scores. The results show the total word count of the text, the entropy score and the relative entropy score which takes into account the length of the text.

- The average sentence length is calculated by NLTK as part of the lexical diversity output;

- Lexical diversity is calculated using the formula proposed by Gonçalves and Gonçalves (2006), $K = 100k(k = n/N)$, where lexical diversity K is the ratio between the number of types n and and number of tokens N.

  The NLTK process is shown in Figure 7.3. The results are shown in blue and indicate the average sentence length and the lexical diversity score, respectively.

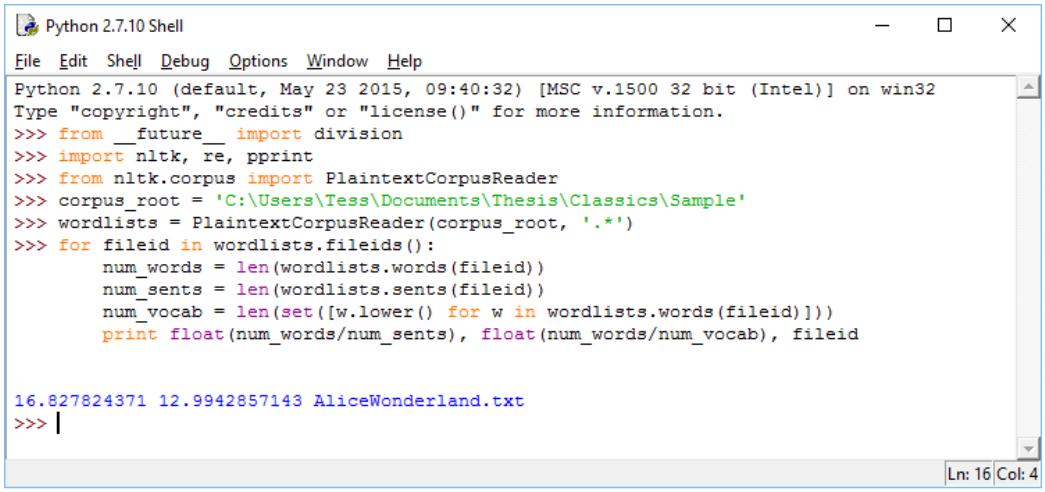


Figure 7.3: Python code for the average sentence length and the lexical diversity

- NLTK is used to extract the POS. An example of NLTK's POS output is given in Figure 7.4.

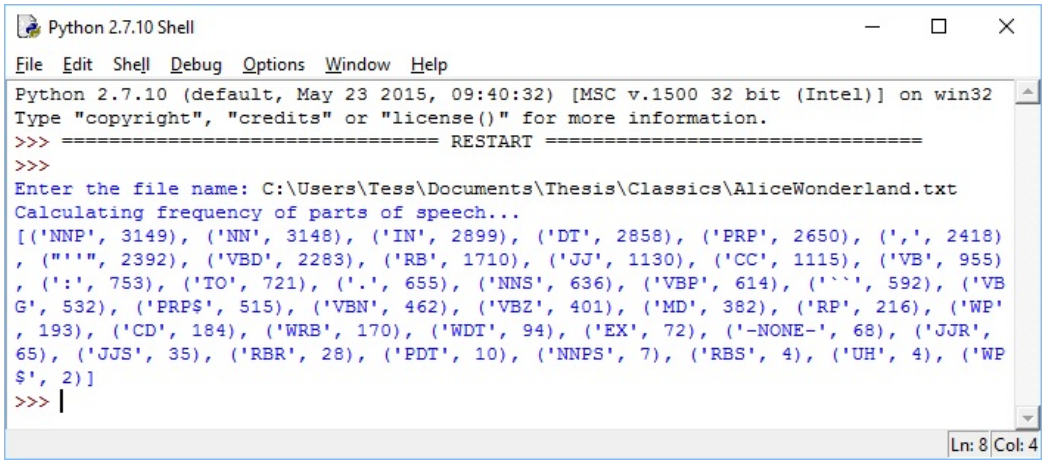


Figure 7.4: Sample output from *Alice in Wonderland*

Scoring

The results of the POS parsing are then transferred to an Excel spreadsheet and calculated as a percentage of the text's composition; the results are compared with the metrics determined as literary quality in Table 7.2 above. CoBAALT then determines whether the text meets the standards by comparing the stylistic entity scores and the POS percentages with the relevant baseline. Variables that are desirable in a literary text and are indicated with a (↑) in Table 7.2 score positively; those that are indicated with a (↓) score negatively. Figure 7.5 shows the Excel sheet that presents a total for *Alice in Wonderland* of -8.21, suggesting that the novel is not literary. As it was written as a children's story, this is not surprising.

POS	COUNT	PERCENT	BASELINE	SCORE
CC	1115	4.29		
CD	184	0.71	0.65	-0.06
DT	2858	10.99		
EX	72	0.28		
IN	2899	11.15	11.52	-0.37
JJ	1130	4.35		
JJR	65	0.25	0.24	-0.01
JJS	35	0.13	0.19	0.06
MD	382	1.47		
NN	3148	12.11	12.65	0.54
NNP	3149	12.11	11.71	-0.40
NNPS	7	0.03	0.02	0.01
NNS	636	2.45	2.95	0.50
NONE	68	0.26		
PDT	10	0.04	0.04	0.00
PRP	2650	10.19	8.05	2.14
PRP$	515	1.98	2.7	-0.72
RB	1710	6.58	5.6	0.98

POS	COUNT	PERCENT	BASELINE	SCORE
RBR	28	0.11		
RBS	4	0.02	0.04	-0.02
RP	216	0.83	0.53	0.30
TO	721	2.77	2.67	0.10
VB	955	3.67		
VBD	2283	8.78	6.63	2.15
VBG	532	2.05		
VBN	462	1.78	2.62	0.84
VBP	614	2.36		
VBZ	401	1.54		
WDT	94	0.36	0.5	0.14
WP	193	0.74	0.55	0.19
WP$	2	0.01		
WRB	170	0.65	0.54	0.11
Total	27308			
AvSenLen		19.29	23.12	3.83
LexDiv		8.09	17.69	-9.6
RelEnt		65	56.08	-8.92
				-8.20622

Figure 7.5: Excel spreadsheet showing the scoring from *Alice in Wonderland*. Those variables not used in the scoring are greyed out.

The CoBAALT workflow is shown in Figure 7.6.

1. Has the text has been processed to remove Project Gutenberg's introductory and legal information? If yes, proceed to #3; if no, continue.

2. Remove text that is not part of the literary work.

3. Calculate the RelEnt using the C program.

4. Parse the text using the NLTK.

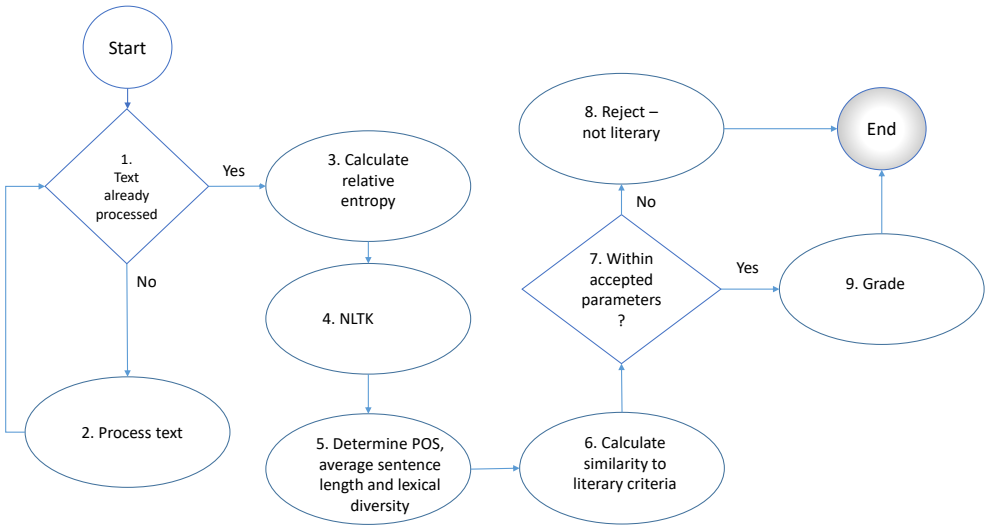


Figure 7.6: The CoBAALT flow process

5. Collect the Python-generated NLTK-calculated scores for POS as a percentage of the text, AvSentLen and LexDiv.

6. Compare to the literary criteria from Table 7.2 and score each variable accordingly. This is achieved as follows:

   - Each variable is compared to its baseline from Table 7.2 and the difference calculated.
   - If the variable score is higher than the baseline the difference is a positive if the Literary ($\uparrow$ *or* $\downarrow$) is $\uparrow$ and a negative if the Literary ($\uparrow$ *or* $\downarrow$) is a $\downarrow$ (Table 7.1).
   - After every variable has been scored, the results are totalled to give the literary score.

7. Is the score above 0.0? If yes, proceed to #9; if no, continue.

8. Reject as not literary.

9. Publish the score (grade).

7.3 Testing the model

7.3.1 Example of CoBAALT scoring

As an example, *Alice in Wonderland* is scored by CoBAALT. Figure 7.4 in the previous section shows the sample output for POS from the text. These are then expressed as a percentage of the total word count of the text. AvSentLen and LexDiv are calculated by the NLTK and RelEnt by the process in Appendix D. The results are seen in Figure 7.7 The score for this

Title	AvSentLen	LexDiv	RelEnt	CD	IN	JJR	JJS	NN	NNS	NNP	NNPS	PDT
Baseline	23.12	17.69	56.08	0.65	11.52	0.24	0.19	12.65	2.95	11.71	0.02	0.04
AliceWonderland	24.74	12.96	59.00	0.68	10.42	0.24	0.12	10.47	2.40	11.96	0.03	0.03
↕	↓	↑	↓	↓	↑	↓	↓	↓	↓	↓	↑	↑
Score	-1.62	-4.73	-2.92	-0.03	-1.10	0.00	0.07	2.18	0.55	-0.25	0.01	-0.01

Title	PRP	PRPs	RB	RBS	RP	TO	VBD	VBN	WDT	WP	WRB	
Baseline	8.05	2.70	5.60	0.04	0.53	2.67	6.63	2.62	0.50	0.55	0.54	
AliceWonderland	8.60	1.87	6.87	0.01	0.81	2.63	8.14	1.74	0.29	0.55	0.54	
↕	↑	↑	↑	↑	↑	↑	↑	↓	↓	↑	↓	
Score	0.55	-0.83	1.27	-0.03	0.28	-0.04	1.51	0.21	0.21	0.00	0.00	**-4.72**

TOTAL of AliceWonderland score is -4.72. Not literary.

Figure 7.7: The CoBAALT scores for *Alice in Wonderland*. The ↕ indicates whether the variable is more literary if the text's number is higher than the baseline (↑) or lower than the baseline (↓).

text is -4.72 and is therefore not considered to be literary.

To test the model, two established literary authors with whom the human panel were very familiar were selected. The panel ordered the novels under examination in order of perceived literary merit. The CoBAALT process was run and the output scores compared with the human panel's judgement. Additionally, published literary criticism was compared to the output scores.

7.3.2 Results using Austen novels

To test the model, eight of Jane Austen's novels were subjected to CoBAALT's process; the results with the literary scores are shown in Table 7.3 along with the order of preference given by the human panel. Of the eight Austen novels, six were included in the experiments in Chapter 6. The remaining two, *Lady Susan* which was not submitted for publication until long after Austen's death and *Love and Freindship* [sic], written when she was still a teenager, score considerably lower than the established works, with scores of -8.07 and -19.00, respectively. *Northanger Abbey* is a satire on the Gothic genre

Table 7.3: Austen novels with their CoBAALT scores and the rank order of the human panel

Text title	Grade	Panel ranking
Emma	14.31	3
Pride and Prejudice	10.26	1
Mansfield Park	6.92	4
Sense and Sensibility	4.34	2
Northanger Abbey	-3.40	5
Persuasion	-3.53	6
Lady Susan	-8.07	7
Love and Freindship [sic]	-19.00	8

and so is written in a specific style that does not indicate literary quality which may account for its low score of -3.40. *Emma* has the highest score at 14.31. Although it was only ranked second by the Gutenberg download citeria, Drabble declares that *Emma* 'is generally considered Jane Austen's most accomplished work' (Drabble, 1996, p. 321). *Persuasion* scores only -3.53. This was her final novel, written while she was already ill and therefore lacking the polish and editing that her previous works enjoyed.

The human panel broadly concurred with the results although their top three were in the order of *Pride and Prejudice*, *Sense and Sensibility* and *Emma*. It is suspected that the higher ranking of *Pride and Prejudice* and *Sense and Sensibility* are due to the popularity of the film versions. Although there has been a recent film of *Emma*, it uses more American actors, a move unlikely to endear it to the British serious reading public. *Mansfield Park* has been made into a film that has not been particularly successful, so it is unlikely to have made much impact on the scoring.

7.3.3 Results using Lawrence novels

As most of the Austen novels had been included in determining the literary criteria, it was decided to choose an established author whose books were not included. Five of D. H. Lawrence's classics were subjected to the CoBAALT process and the results of the scoring are shown in Table 7.4.

Sons And Lovers is a novel described by Becket as containing 'extremely accomplished writing' that she considers to be superior to works such as *The Trespasser* (Becket, 2002, p. 43); it scores the highest of the five texts with 12.94 points while *Women In Love* comes a close second. This is in keeping

Table 7.4: Lawrence novels with their CoBAALT scores and the rank order by the human panel

Text title	Grade	Panel ranking
Sons and Lovers	12.94	=1
Women In Love	11.30	=1
The Rainbow	7.65	3
The Trespasser	-4.52	4
Lady Chatterley's Lover	-11.31	5

with Lawrence's own opinion of the novel as he declared it to be his best work (Drabble, 1996, p. 1091). *The Rainbow* is rated by F.R. Leavis as one of his best (cited in Becket p. 127). *The Trespasser* is 'by Lawrence's own admission a juvenile work, and he quickly tired of it' but Becket, while agreeing it to be a minor work, proclaims it contains some excellent writing (Becket, 2002, p. 41). While *Lady Chatterley's Lover* is arguably the best-known of Lawrence's novels, it is more for its notoriety than its literary merit. Lawrence experiments in this novel with form 'bravely (some might say disastrously)' (Becket, 2002, p. 78), changing narrative for epistolary mode at the end of the book and this is reflected in the very low score of -11.31.

The human panel concurred with the results although *Sons and Lovers* and *Women in Love* were a dead heat for first place.

7.4 Observations

The CoBAALT output scores were in line with the findings of the human panel and with established literary opinion. The omission of D. H. Lawrence from the top 100 novels downloaded from Project Gutenberg had been noticed at the time and was remarkable for such a well-established author. The CoBAALT scores for *Sons and Lovers* and *Women in Love* suggest that these two novels at least should have been on a literary par with Austen's *Emma* or *Pride and Prejudice*. More recent investigation on Gutenberg's website shows that *Women in Love* has been downloaded over 900 times which would have put it well within the top 100 books. It is assumed that Lawrence was a more recent addition to the Gutenberg collection when the top 100 was identified in 2014, hence the low original number of downloads.

7.4.1 Fiction versus non-fiction

Although non-fiction was not included in the testing, CoBAALT shows that differences between fiction and non-fiction are quite marked for some POS, as shown in Figure 7.8.

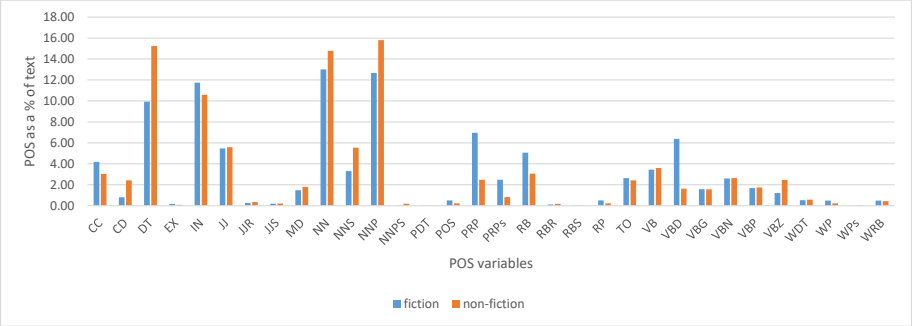


Figure 7.8: Fiction and non-fiction averages of POS

This shows, for example, that there are fewer CC in a non-fiction text but more CD and DT. This is logical; CC is a coordinating conjunction such as 'and', 'or', 'but'. Because literature aims to paint a picture in words for the reader, this is a feature used to link clauses and sometimes used as an adverbial constuct (Wales, 1990, p. 88) so it is normal that these would occur less frequently in non-fiction. CD and DT, however, as cardinal numbers and determiners, are used to specify nouns and are more likely to appear in non-fiction.

Features such as PRP and PRP$ demonstrate little variety between the fiction texts but in non-fiction they score very low. Again, this is a logical finding; personal and possessive pronouns relate to people (characters). Also indicated in the graph is the difference in occurrences of nouns and verbs. Non-fiction uses more nouns in all forms while fiction concentrates on verbs

that move the action along. The exception to this is seen for VBP and VBZ which indicate the present tense for non-third person and third person, respectively. As most fiction is still written in the past tense and non-fiction in the present, this is a perfectly understandable reversal of the trend.

Other differences include alliteration and average sentence length; these are lower for non-fiction. Alliteration was not one of the variables used to score fiction, however, it is an important variable to consider when examining non-fiction. With regard to average sentence length, one of the non-fiction texts, a BBC article, is a significant exception with an average of 34 words per sentence against a mean of 25.26 for non-fiction. Lexical diversity is very much lower for non-fiction with an average LexDiv score of 6.52 against a mean of 13.97 for all fiction scores. This test indicates the number of times words are used in the text so a high LexDiv score suggests a richer vocabulary.

There is a much higher score for relative entropy, averaging a score of 71.00 for non-fiction against an average of 59.02 for all fiction scores. Here, entropy is defined as the degree of predictability in a text. Literature is particularly entropic compared to everyday speech (Wales, 1990, p. 149) so this result confounds expectations. Maximum entropy would be achieved if all the words in the text were different. It appears that the length of the text has some impact on the RelEnt with regard to the higher scores for non-fiction. These texts were necessarily shorter, being mostly instruction manuals or news articles. The writers of these texts, therefore, are able to include more synonyms to vary their writing and express themselves without repeating words whereas this is more of a challenge in a 100,000 word novel, particularly if constantly referring to the same characters and places.

7.5 CoBAALT as a determiner of literary merit

The results of the experiments using novels by Austen and Lawrence are encouraging, corresponding with the opinions of the literary critics. Although there was some difference between the results and the opinions of the human panel and the Gutenberg downloads, these were minor. This suggests that the metrics chosen are exemplary tools to use when determining the literary 'worth' of a text. Whether or not CoBAALT can be used to predict the next *Harry Potter* or Man Booker winner remains to be seen but it certainly seems a plausible goal.

Chapter 8

Conclusion and Further Work

8.1 Summary of chapters

Chapter 1 provides an introduction to the thesis. The initial literature review in Chapter 2 to determine whether computational analysis of degrees of literariness was feasible comprehensively investigated authorship attribution and classification tasks, but there were some encouraging studies that were more concerned with the content or ranking of works of fiction, usually pertaining to a specific author (Li et al., 2004; Stubbs, 2005; Plaisant et al., 2006; Haiyan and Xiaohu, 2011; Hammond et al., 2013; Ashok et al., 2013). Chapter 3 outlines the methodology used in the thesis. As a pilot study, the work outlined in Chapter 4 was undertaken to investigate the durability of POS and literary features when the text is corrupted to varying degrees and to ensure that literary features would not be lost when computationally parsed without manual intervention. The facets recognisable as literary (Tables 4.1 and 4.2) turned out to be remarkably robust, suggesting that it would be possible to produce a quantitative procedure to measure literariness.

Chapter 5 gives an overview of schools of literary criticism and indicates which facets are computationally identifiable. In order to map the search for literary features, a series of focus groups, online surveys and interviews were carried out to determine the qualities that a human reader looks for in a literary work (Table 5.1). From these uncovered broad qualities, the features that are quantifiable were identified. Not every aspect was included; for example, "learning something new" figured in the survey but cannot be

quantified as this experience is different for each individual. As human readers rarely break down their reading experience into identifiable components, the results from the primary investigation were used to guide the next stage of the research and confirmed that a stylistic analysis was a valid approach for determining literary merit.

Once the broad literary characteristics were isolated, a basic framework was constructed to quantify each feature, outlined in Chapter 6. The initial results were piloted by an online survey, given in Section 6.1.1. Further refinement using factor analysis (Section 6.2.1) resulted in a model that can determine literary merit according to the metrics used. The variables identified are given in Section 6.3. These were built on examples of Classics and lesser-known texts that are available from the Gutenberg Project, as well as non-fiction texts. Classics were chosen because of the wealth of available expert literary criticism and the lesser-known books and non-fiction were used as examples of less literary texts.

Using the results found in the previous chapter, Chapter 7 describes the CoBAALT model, a toolbox of processes to identify and quantify specific literary features. The efficacy of CoBAALT is examined in three ways: by its ranking of texts and the extent to which this corresponds to Project Gutenberg downloads (a measurement used by Ashok et al. (2013)), by independent assessment by a human panel and by the extent to which the results correspond to existing literary criticism. CoBAALT was tested on novels by Jane Austen (Section 7.3.2) and D. H. Lawrence (Section 7.3.3). The results agreed to a greater extent with the Gutenberg Project and with the ranking done by the human literary panel and were in accord with published literary criticism (Drabble, 1996; Becket, 2002).

8.2 Contributions

The contribution of this study is an intellectual one, although there are practical examples of where the tool could be utilised, such as use as a teaching tool for creative writing or for refining translated texts.

- **Major - the development of a definitive model for application to a given text to qualify its degree of literary merit.**

State-of-the-art stylistic identification techniques are refined and a comprehensive examination of various tools is carried out (Chapter 6) to confirm a selection of computationally identified variables that are used to qualify the degree of literariness (Tables 6.4 and 6.5). The resulting model (Table 7.2) is tested on classic works of fiction written in or translated into English and its findings correspond to those of a human panel, established literary criticism and the number of downloads from the Gutenberg Project.

- **Minor - the integration of qualitative and quantitative text-analytical metrics are a contribution to knowledge and an enrichment of existing techniques in stylistic analysis.**

The features and POS identified by factor analysis and shown in Tables 6.4 and 6.5 define the characteristics of greatest impact on literary quality while Tables 6.6 and 6.7 indicate the desired proportion of each variable.

- **Minor - the literary devices that constitute "good" literature are identified and examined.**

An overview of literary criticism is given (Section 5.1) and fieldwork carried out to extract the features that humans relate to when reading (Section 5.2). These do not direct the choice of variables used in the CoBAALT model but serve to confirm those identified by the factor analysis.

- **Minor - use of the CoBAALT model provides a way to recognise non-fiction and fiction texts and categorise them accordingly.**

Although non-fiction works were included merely to provide examples of non-literary text, CoBAALT proved to be extremely adept at distinguishing between fiction and non-fiction (Section 7.4.1).

8.3 Conclusion

The research objectives of this thesis laid out in Section 1.1 were addressed as follows:

1. **To investigate the limitations of computers in interpreting text**. This was achieved through a comprehensive literature review in Chapter 2 that examined the tools available that could be used in stylistic analysis of text to determine aesthetic merit. Furthermore, an examination was carried out in Chapter 4 to test whether literary features are robust enough to withstand computational parsing without manual intervention. The success of this meant that parsing errors made by the computer would not detract from the overall value judgement of literary merit. Literary features were found to be remarkably robust, even when texts are altered by a translation and re-translation process. Extensive research suggested that authorship attribution tools could be adapted to create a model to identify stylistic features that indicate literary merit rather than a specific author.

2. **To develop a pragmatic definition of "good" literature and identify its features.** To confirm the hypothesis that a computer can determine literary merit, a definition of what constitutes "good" literature had to be established. This was done by an investigation into schools of literary criticism (Section 5.1) and into the human perspective of literary appreciation (Section 5.2). Fieldwork was carried out to discover the constituents of "good" literature as perceived by human readers. This was achieved by surveys, focus groups and interviews with both experts in literature (with at minimum a degree in Literature) and the general reading public. This found that a stylistic analysis is both possible and valid for proscribing a text as "good".

3. **To develop a framework of features that indicate "good" literature.** In Chapter 6, once key features had been identified, a framework was designed to identify and capture specific literary features. Factor analysis was used to identify the most significant variables that have the most influence on the degree of literariness of a text. These variables are substantiated by the data collected and identified in Section 5.2 as desirable features.

4. **To produce a model that determines aesthetic value according to the metrics proscribed.** The CoBAALT model scores each variable according to the metrics defined in Table 7.2. These scores may be positive for "good" literature and negative for less literary or non-fiction texts. The system was evaluated by testing on seen and unseen works by Jane Austen and D. H. Lawrence and the results confirmed the findings of a human panel, published literary criticism and download figures from the Gutenberg Project (Section 7.3).

8.4 Limitations and further work

The work in this thesis concentrated upon Classic works of fiction written in or translated into English that were out of copyright and therefore freely available. Investigation into more modern literature is desirable but impeded by the lack of machine-readable sources that are free of charge. This line of investigation would be useful in establishing the current state of publishing demands and determining the requirements for a successful modern literary submission.

However, there are plenty of texts available on the Gutenberg Project in different languages; these will have their own statistical fingerprint of literature that needs to be defined before proceeding with analysis but CoBAALT's basic framework should make a good foundation. Different languages were not tested but many of the rules of literature are universal, such as use of adjectives and adverbs to paint a richer word picture or varying the sentence length, so many of the features identified and used in CoBAALT could be used to investigate literature in other languages. Some POS, however, would not be transferable due to, for example, the lack of determiners in Slavic and articles in Chinese. A different human panel of experts in foreign literature is required before such an investigation can take place.

To summarise, the thesis is limited by the following factors:

- a lack of access to funds for modern day collections of literary texts. Writing styles and tastes change over time and this research has been conducted on works of fiction that are out of copyright (70 years after the death of the author under UK law) and hence freely available. This means that most of the literature investigated is at least 90 years old. Modern novels tend to be shorter and more concise, due in part to the use of technology for easy editing. Modern novels may require some tweaking of the variables to deliver an accurate literature score.

- a focus on texts written in or translated into English. The researcher lacks linguistic ability in other languages yet a sound understanding of English literature and grammar has been essential for this research. Other languages have their own specific grammar rules and these will necessarily have an impact on using CoBAALT for foreign literature. Some amendment of the variables is anticipated for texts written in languages other than English.

- recruitment of a human panel of literary experts that does not reflect cultural diversity or expertise in other languages. The researcher is aware that the focus groups and, to a lesser extent, the open surveys are biased towards a white, English-speaking, middle-class and middle-income demographic, these being the people who could be recruited for book surveys through personal contacts. The opinions of those reading in other languages would be a welcome addition to the research.

Further work may include:

- investigating more specific genres such as children's literature or Gothic horror. Similar styles mean that a more fine-grained approach could be investigated that may uncover further insights into the definition of "good" literature. CoBAALT has been developed using Classics and is tested on the same genre. Although it is expected that the same variables would be utilised for other genres, refinements in the metrics are anticipated.

- investigating current bookshelf literature including unpublished work. Literary tastes change over the years. A book written in the style of the nineteenth century would not be published as a current work due to the different readership; today's reading public demands a faster plot, less description and instant gratification. There is no longer patience for long introductions, literary segues that fail to move the story along or irrelevant back-story telling. The metrics used in Table 7.2 would need to be adjusted to encapsulate the tastes of a modern audience although it is anticipated that the variables used would remain essentially the same.

- using languages other than English. Although the development of the model utilised some texts in translation, all the texts were written in English. It is anticipated that similar languages (Germanic and Romance) would have very similar metrics to those used in the English CoBAALT; once dissimilar languages such as Chinese are incorporated it is expected that the metrics would change more radically due to the different emphasis on POS, e.g. no articles, and cultural expectations from literature. However, the basic concept of using stylistics to measure literary quality remains viable and factor analysis could again be used to identify the variables that constitute "good" literature in the appropriate language, with confirmation from a human panel or focus group.

8.5 Summary

This chapter concludes this research thesis. Investigation has been carried out into the robustness of literary features and the elements that constitute "good" literature. Experiments were carried out to determine such elements and a process developed to identify and qualify texts in degree of literariness. The hypothesis, that a computer can analyse literary text according to accepted criteria and make judgements of literary merit, has been shown to be valid. The work contained herein is limited to works in or translated into English that are out of copyright. Further investigation into different languages and cultural expectations of literature are future work but it is anticipated that the core CoBAALT model can provide a solid basis with minor adjustment and fine-tuning for any such future investigation and development into CoBAALTv.2.0.

Appendix A

Focus Groups

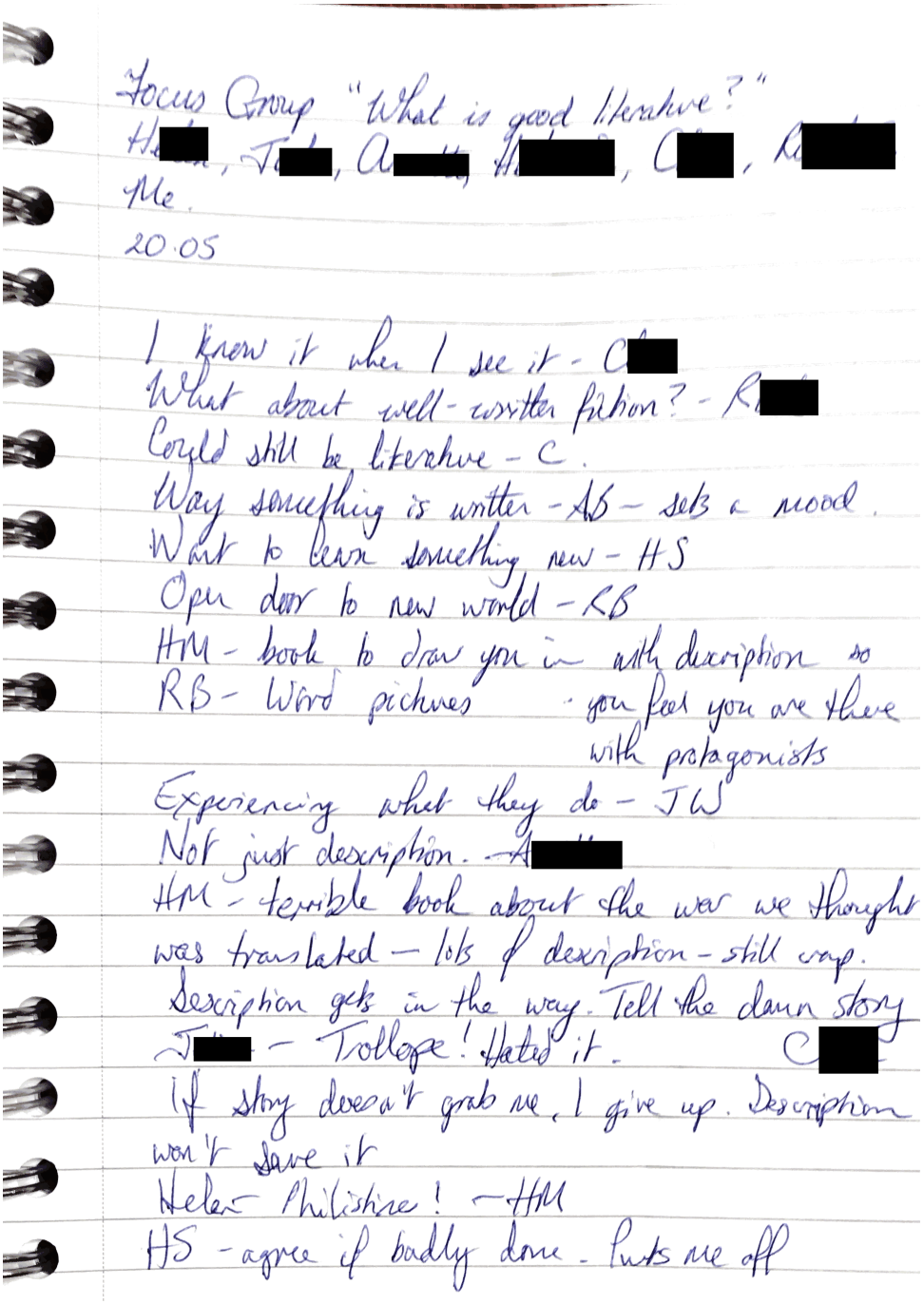
Two focus groups were held, the first in December 2013 and the second in June 2015. This appendix shows the information given to participants and an example of the moderator's notes. Names have been obscured to ensure anonymity.

Focus groups

Information for participants

Thank you for taking part in this focus group for my research. My name is Tess Crosbie and I am a PhD candidate at the University of Bedfordshire. The area I am investigating is to see whether a computer can recognise good literature from something that is less well-written and then apply some qualitative judgement to a text. Of course, the first question is “what is good literature?” and that is what this focus group is here to discuss. You all have responses to literature and I would like to understand how you decide whether a book is “good” or “bad” in the hope that a computer can be programmed to make the same value judgements.

Your contributions will be anonymised and not identified with you personally. You are under no obligation to answer questions you prefer not to and you may withdraw your participation at any time. The final thesis will be available after publication and I will send you a link at that time to read the paper if you would like to do so.


Focus Group "What is good literature?"
[redacted], [redacted], [redacted], [redacted], [redacted], [redacted]
Me.
20.05

I know it when I see it - [redacted]
What about well-written fiction? - [redacted]
Could still be literature - C.
Way something is written - AS - sets a mood.
Want to learn something new - HS
Open door to new world - RB
HM - book to draw you in with description so
RB - Word pictures - you feel you are there with protagonists
Experiencing what they do - JW
Not just description. - [redacted]
HM - terrible book about the war we thought was translated - lots of description - still crap.
Description gets in the way. Tell the damn story
[redacted] - Trollope! Hated it. [redacted]
If story doesn't grab me, I give up. Description won't save it
Helen Philistine! - HM
HS - agree if badly done. Puts me off

Figure A.1: Example of handwritten notes taken during the first focus group

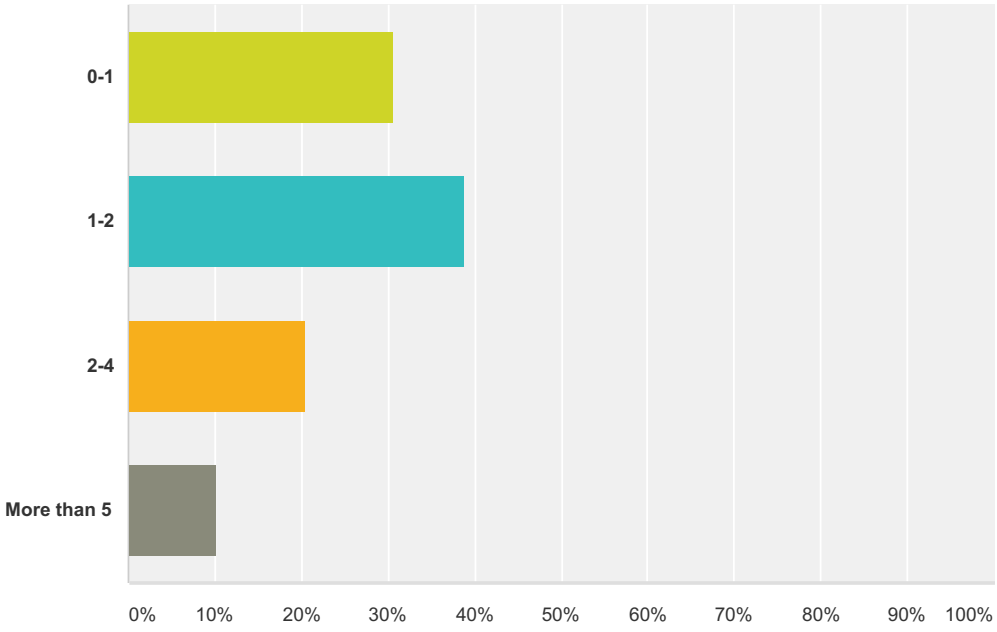
Appendix B

What Makes a Good Book?

This online survey was carried out in 2015 and this appendix includes the verbatim responses.

Q1 How many fiction books do you read in an average month?

Answered: 49 Skipped: 4

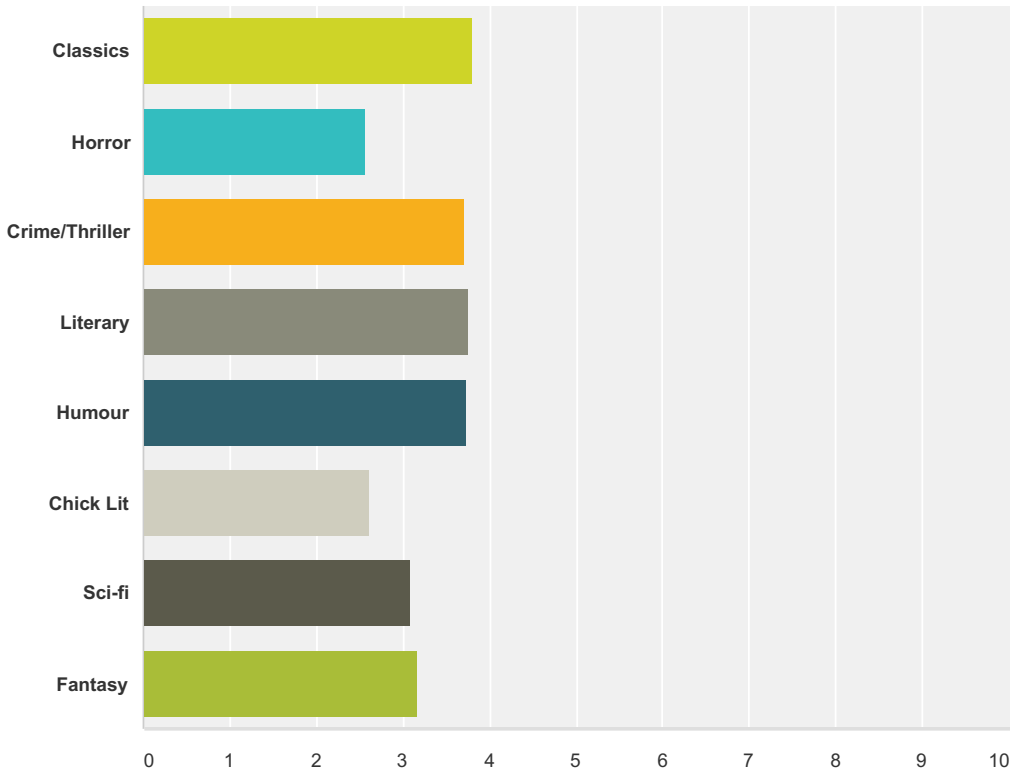


Answer Choices	Responses	
0-1	**30.61%**	15
1-2	**38.78%**	19
2-4	**20.41%**	10
More than 5	**10.20%**	5
Total		**49**

#	Optional comments	Date
1	Read 120+ last year, on track for about 70 this year	6/30/2015 10:36 AM
2	I listen to books more than read them theses days	6/26/2015 7:30 AM
3	Read 100+ books in 2014	6/25/2015 9:35 PM
4	Read all day for work; on holiday I read 2 books a week though.	3/3/2015 9:38 PM
5	I know I should read more.	3/3/2015 3:13 PM

Q2 What genres do you prefer?

Answered: 53 Skipped: 0



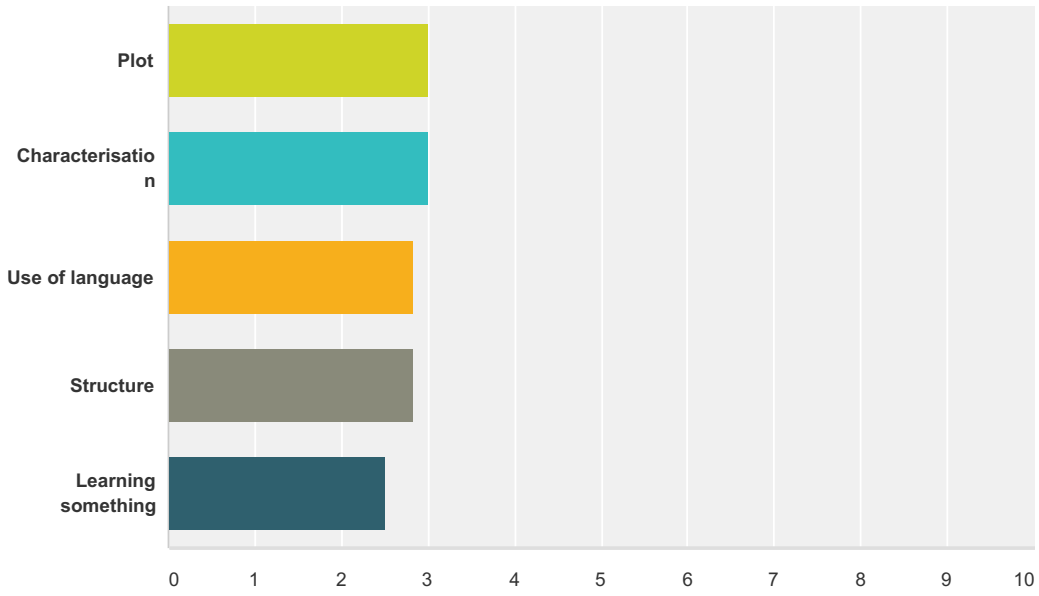
	Loathe it	Dislike it	Neutral	Like it	Love it	Total	Weighted Average
Classics	**3.92%** 2	**5.88%** 3	**25.49%** 13	**35.29%** 18	**29.41%** 15	51	3.80
Horror	**25.49%** 13	**29.41%** 15	**13.73%** 7	**25.49%** 13	**5.88%** 3	51	2.57
Crime/Thriller	**1.96%** 1	**13.73%** 7	**23.53%** 12	**33.33%** 17	**27.45%** 14	51	3.71
Literary	**0.00%** 0	**7.84%** 4	**31.37%** 16	**37.25%** 19	**23.53%** 12	51	3.76
Humour	**0.00%** 0	**7.69%** 4	**28.85%** 15	**46.15%** 24	**17.31%** 9	52	3.73
Chick Lit	**25.49%** 13	**15.69%** 8	**37.25%** 19	**15.69%** 8	**5.88%** 3	51	2.61
Sci-fi	**17.31%** 9	**23.08%** 12	**17.31%** 9	**17.31%** 9	**25.00%** 13	52	3.10
Fantasy	**11.54%** 6	**25.00%** 13	**19.23%** 10	**23.08%** 12	**21.15%** 11	52	3.17

#	Other genres (please specify)	Date
1	historical	7/1/2015 3:08 PM
2	YA love it	6/30/2015 10:36 AM
3	historical	6/26/2015 8:14 AM
4	YA love it, erotica neutral	6/25/2015 9:35 PM
5	historical fiction	6/25/2015 5:36 PM
6	Magical realism	6/25/2015 4:37 PM

7	Historical fiction	3/30/2015 4:20 PM
8	Black comedy	3/9/2015 3:01 PM
9	biographies, erotica, royal history	3/9/2015 1:16 PM

Q3 What do you look for in a good book?

Answered: 38 Skipped: 15



	Not important	Somewhat unimportant	Neutral	Somewhat important	Important	Total	Weighted Average
Plot	**0.00%** 0	**0.00%** 0	**0.00%** 0	**27.03%** 10	**72.97%** 27	37	3.00
Characterisation	**0.00%** 0	**0.00%** 0	**0.00%** 0	**31.43%** 11	**68.57%** 24	35	3.00
Use of language	**0.00%** 0	**0.00%** 0	**16.67%** 6	**25.00%** 9	**58.33%** 21	36	2.83
Structure	**0.00%** 0	**2.78%** 1	**13.89%** 5	**30.56%** 11	**52.78%** 19	36	2.83
Learning something	**8.11%** 3	**13.51%** 5	**18.92%** 7	**37.84%** 14	**21.62%** 8	37	2.51

#	Other (please specify)	Date
1	Pace	7/1/2015 3:15 PM
2	Optimism. For example, if the books scenario is an after the apocalypse, then now I won't touch it after reading too many such that start bad and go downhill from there.	6/25/2015 5:49 PM
3	Ideas	3/9/2015 8:17 PM
4	Believability	3/9/2015 1:37 PM
5	Witty or clever dialogue	3/3/2015 7:54 PM
6	It depends - sometimes I want it to conform satisfyingly to the genre. Sometimes I want it to surprise and challenge me.	3/3/2015 1:31 PM

Q4 For you, what makes a good book stand out from the rest of them?

Answered: 31 Skipped: 22

#	Responses	Date
1	A good book is one I don't want to put down.	7/11/2015 11:59 PM
2	Believable characters that act out a well structures plot and engage the reader to the extent that a book is memorable	7/8/2015 5:37 PM
3	Not being able to put it down!	7/1/2015 3:15 PM
4	New world, lack of predictability, rock-solid science base with interesting extrapolations	6/30/2015 10:03 PM
5	A cast of characters that you really connect to and great world-building	6/30/2015 10:47 AM
6	A good plot and beautiful writing	6/26/2015 5:21 PM
7	World building, natural dialogue	6/26/2015 12:54 PM
8	If its written around a subject that interests me	6/26/2015 7:10 AM
9	Something which gets my attention straight away	6/26/2015 12:56 AM
10	Has something that is distinctly 'different' from the norm be it tone, plot or other.	6/25/2015 8:20 PM
11	Something that grips me early on	6/25/2015 7:14 PM
12	a book where you can not predict the ending	6/25/2015 7:09 PM
13	Credible, interesting characters	6/25/2015 7:04 PM
14	A good plot with characterization such that I both feel I am in the story and am emotionally involved with the characters.	6/25/2015 5:49 PM
15	It stands the test of time. I love to re-read books again and again if I like them.	6/25/2015 4:27 PM
16	it's hard to distinguish particular features, i think most of the outstanding books for me come from my younger age when i was more impressionable.	6/25/2015 4:23 PM
17	The way language is used and expectations played with	6/25/2015 4:20 PM
18	Charm, memorable characters with morals who have interesting lives and events	5/8/2015 8:48 PM
19	gripping the imagination.	3/30/2015 4:27 PM
20	Characters that I care about / am interested in rather than whiney ones that I would happily kill myself.	3/9/2015 9:36 PM
21	When you don't want to put the book down	3/9/2015 9:07 PM
22	Ideas and a story	3/9/2015 8:17 PM
23	Good use of words and description imaginatively written	3/9/2015 2:16 PM
24	When it has a concept that's different from the usual.	3/9/2015 1:58 PM
25	Credibility of the storyline... I want to believe it can happen.	3/9/2015 1:57 PM
26	fast pace/no waffle	3/9/2015 1:37 PM
27	An unexpected turn of plot, and convincing characters	3/4/2015 3:32 PM
28	All the features above in a novel way.	3/3/2015 9:41 PM
29	Dithe language	3/3/2015 7:54 PM
30	A book where you really don't want to put it down. A book that catches you out.	3/3/2015 3:22 PM
31	Interesting and unexpected use of language/structure.	3/3/2015 1:31 PM

Q5 In no particular order, what are your three favourite fiction books?

Answered: 34 Skipped: 19

Answer Choices	Responses	
First book choice	**100.00%**	34
Why do you like this book?	**85.29%**	29
Second book choice	**97.06%**	33
Why do you like this book?	**79.41%**	27
Third book choice	**97.06%**	33
Why do you like this book?	**79.41%**	27

#	First book choice	Date
1	doomsday book by connie willis	7/11/2015 11:59 PM
2	Return of the Native by Thomas Hardy	7/8/2015 5:37 PM
3	Birdsong by Sebastian Faulks	7/1/2015 3:15 PM
4	Hitchhikers guide to the Galaxy	6/30/2015 10:03 PM
5	Harry Potter & the Half-Blood Prince	6/30/2015 10:47 AM
6	Tinker Taylor Soldier Spy	6/26/2015 5:21 PM
7	American Gods	6/26/2015 12:54 PM
8	The Hunger Games (The Hunger Games #1) by Suzanne Collins	6/26/2015 11:24 AM
9	The long winter Laura ingalls wider	6/26/2015 7:39 AM
10	fifty shades trilogy	6/26/2015 7:10 AM
11	Lord of the Rings	6/26/2015 12:56 AM
12	The Martian (Subject to change, all the time)	6/25/2015 8:20 PM
13	The English Patient	6/25/2015 7:14 PM
14	the Island by Victoria Hislop	6/25/2015 7:09 PM
15	Star of the Sea	6/25/2015 7:04 PM
16	The Lord of the Rings	6/25/2015 5:49 PM
17	The Plague, Albert Camus	6/25/2015 4:45 PM
18	Lord of the rings trilogy	6/25/2015 4:27 PM
19	The Days of Solomon Gursky	6/25/2015 4:23 PM
20	The NIght Circus	6/25/2015 4:20 PM
21	The Odyssey	5/8/2015 8:48 PM
22	Birdsong by Sebastian Faulks	3/30/2015 4:27 PM
23	Tom Jones	3/9/2015 9:36 PM
24	The Perfume	3/9/2015 9:07 PM
25	Consider phlebas	3/9/2015 8:17 PM
26	The English Patient	3/9/2015 2:16 PM
27	Cloud Atlas	3/9/2015 1:58 PM
28	Gullivers Travels	3/9/2015 1:57 PM
29	Christopher Reeve Still Me	3/9/2015 1:37 PM
30	The Red Tent	3/4/2015 3:32 PM

31	The Red Tent	3/3/2015 9:41 PM
32	Birds without wings	3/3/2015 7:54 PM
33	Pride and Prejudice	3/3/2015 3:22 PM
34	Vanity Fair	3/3/2015 1:31 PM
#	**Why do you like this book?**	**Date**
1	it was the fist time travel book i read and got very engrossed with both the characters and plot.	7/11/2015 11:59 PM
2	For it's intriging character driven tragic nature.	7/8/2015 5:37 PM
3	Evocative, moving.	7/1/2015 3:15 PM
4	Fantastically clever allegory for bureaucracy with wordplay and humour	6/30/2015 10:03 PM
5	Great world-building, characters you love, gripping plot	6/30/2015 10:47 AM
6	I enjoyed the main character Smiley	6/26/2015 5:21 PM
7	Interesting story	6/26/2015 12:54 PM
8	The Most Dangerous Game" meets Survivor. I loved it: deft characterization wrapped tightly around this lean, brutal plot that absolutely grabbed me and hung on. This was one that I started reading in the morning and snarled at all comers until I had finished it in the evening.	6/26/2015 11:24 AM
9	i read this as a child and as an adult and loved it both times	6/26/2015 7:39 AM
10	plot	6/26/2015 7:10 AM
11	love the descriptive writing	6/26/2015 12:56 AM
12	Funny sci fi, which is quite rare	6/25/2015 8:20 PM
13	The Italian setting and the language and romance	6/25/2015 7:14 PM
14	great story line and learnt something new	6/25/2015 7:09 PM
15	character, quality of writing and strong sense of place	6/25/2015 7:04 PM
16	It encompasses all I've said. Brilliant plot, though slow starting, great characterization and so many sub-plots that reading it again is not boring	6/25/2015 5:49 PM
17	Astonishing profundity, plus humour & great story	6/25/2015 4:45 PM
18	Outstanding in every sense!	6/25/2015 4:27 PM
19	Because it's a beautiful story of one man, his path and what matters for him.	6/25/2015 4:23 PM
20	Structure/plot	6/25/2015 4:20 PM
21	Rosy-fingered down, wine-dark sea: charming language, excellent story	5/8/2015 8:48 PM
22	Compelling and moving.	3/30/2015 4:27 PM
23	it is outrageously funny	3/9/2015 9:36 PM
24	unusual story	3/9/2015 9:07 PM
25	Ideas	3/9/2015 8:17 PM
26	wonderful writing. Italian setting	3/9/2015 2:16 PM
27	Very unusual structure, and original ideas.	3/9/2015 1:58 PM
28	Sometimes I feel small, sometimes I feel tall.	3/9/2015 1:57 PM
29	Well written, heart touching, no padding	3/9/2015 1:37 PM
#	**Second book choice**	**Date**
1	pillars of the earth by ken follett	7/11/2015 11:59 PM
2	The Goldfinch by Donna Tart	7/8/2015 5:37 PM
3	Cloud Atlas by David Mitchell	7/1/2015 3:15 PM
4	Lord of the Flies	6/30/2015 10:03 PM
5	Red Rising	6/30/2015 10:47 AM
6	Any Human Heart	6/26/2015 5:21 PM

7	Lord of the Rings	6/26/2015 12:54 PM
8	the hobbit	6/26/2015 7:39 AM
9	Danielle Steele Malice	6/26/2015 7:10 AM
10	Watership Down	6/26/2015 12:56 AM
11	Game of thrones (Subject to change, all the time	6/25/2015 8:20 PM
12	The Story of San Michele	6/25/2015 7:14 PM
13	The long Winter by Laura Ingells wilder	6/25/2015 7:09 PM
14	half of a yellow sun	6/25/2015 7:04 PM
15	On Basilisk Station	6/25/2015 5:49 PM
16	Aunt Julia & The Scriptwriter, Mario Vargas Llosa	6/25/2015 4:45 PM
17	Pride and Prejudice	6/25/2015 4:27 PM
18	Ring around the sun	6/25/2015 4:23 PM
19	Possesion	6/25/2015 4:20 PM
20	Gone with the Wind	5/8/2015 8:48 PM
21	Cloud Atlas by David Mitchell	3/30/2015 4:27 PM
22	The Tao of Pooh	3/9/2015 9:36 PM
23	The Sparrow	3/9/2015 9:07 PM
24	Adolf Hitler my part in his downfall	3/9/2015 8:17 PM
25	Gone With the Wind	3/9/2015 2:16 PM
26	The Conqueror series (Genghis Kahn)	3/9/2015 1:58 PM
27	Robinson Crusoe	3/9/2015 1:57 PM
28	Tell me your dreams by Sidney Sheldon	3/9/2015 1:37 PM
29	Harry Potter	3/4/2015 3:32 PM
30	The Poisonwood Bible	3/3/2015 9:41 PM
31	Green Dolphin Stret	3/3/2015 7:54 PM
32	Lives and loves of a She Devil	3/3/2015 3:22 PM
33	The Handmaid's Tale	3/3/2015 1:31 PM

#	Why do you like this book?	Date
1	as i enjoy history this was an ideal way to read about the social and cultural life of thevperiod.	7/11/2015 11:59 PM
2	The pace and sense of impending doom that hangs over the main protagonist keeps the reader interested	7/8/2015 5:37 PM
3	Provokes thought.	7/1/2015 3:15 PM
4	Simple premise well explored, great characterisation	6/30/2015 10:03 PM
5	Intense drama, character politics, great world-building	6/30/2015 10:47 AM
6	Liked the narrative and character	6/26/2015 5:21 PM
7	Epic scale	6/26/2015 12:54 PM
8	had everything I love in a book adventure,suspense and great writing	6/26/2015 7:39 AM
9	Plot	6/26/2015 7:10 AM
10	beautifully written, different	6/26/2015 12:56 AM
11	Very in depth charicterisations, and very tense	6/25/2015 8:20 PM
12	A love from my extreme youth	6/25/2015 7:14 PM
13	the description and stories are brilliant	6/25/2015 7:09 PM
14	same reason as above	6/25/2015 7:04 PM
15	A great SF military story that involves a complex plot with characters who are complex themselves, showing strengths and weaknesses	6/25/2015 5:49 PM

16	Funny, brilliant, dazzling, hooks you from the start, stories within the story.	6/25/2015 4:45 PM
17	Wonderful prose and great escapism	6/25/2015 4:27 PM
18	Language/structure	6/25/2015 4:20 PM
19	Learn more each time I re-read it	5/8/2015 8:48 PM
20	a different way of looking at the world	3/30/2015 4:27 PM
21	spiritual guidance i can relate to	3/9/2015 9:36 PM
22	fascinating	3/9/2015 9:07 PM
23	Humour and reality	3/9/2015 8:17 PM
24	The huge story of it all with drama and romance.	3/9/2015 2:16 PM
25	Very detailed, but fictionalised books about an interesting historical figure.	3/9/2015 1:58 PM
26	I want to live on an island and only have to talk on Friday's.	3/9/2015 1:57 PM
27	based on medical fact, fast paced, not a see through plot	3/9/2015 1:37 PM
#	**Third book choice**	**Date**
1	diary of a nobody by George and weedon grossmith	7/11/2015 11:59 PM
2	Behind the scenes at the Museum by Kate Atkinson	7/8/2015 5:37 PM
3	The hundred year old man who climbed out of the window and disappeared by Jonas Jonasson	7/1/2015 3:15 PM
4	1984	6/30/2015 10:03 PM
5	Assassin's Apprentice	6/30/2015 10:47 AM
6	Madame Bovary	6/26/2015 5:21 PM
7	Name of the Wind	6/26/2015 12:54 PM
8	the persuader lee child	6/26/2015 7:39 AM
9	Film books	6/26/2015 7:10 AM
10	King of the Wind	6/26/2015 12:56 AM
11	Station Eleven (Subject to change, all the time)	6/25/2015 8:20 PM
12	Gone with the Wind	6/25/2015 7:14 PM
13	Diana by RF Delderfield	6/25/2015 7:09 PM
14	germinal	6/25/2015 7:04 PM
15	The Stories of Sherlock Holmes	6/25/2015 5:49 PM
16	Chronicle of a Death Foretold. Gabriel Garcia Marquez	6/25/2015 4:45 PM
17	Life of Pi	6/25/2015 4:27 PM
18	The Master and Margarita	6/25/2015 4:23 PM
19	Life After Life	6/25/2015 4:20 PM
20	Summer- Edith Wharton	5/8/2015 8:48 PM
21	The Help by Kathryn Stockett	3/30/2015 4:27 PM
22	The Count of Monte Cristo	3/9/2015 9:36 PM
23	The Magus	3/9/2015 9:07 PM
24	Jonathan Livingston seagull	3/9/2015 8:17 PM
25	The Light Between the Oceans	3/9/2015 2:16 PM
26	Lord of the Rings	3/9/2015 1:58 PM
27	Candide	3/9/2015 1:57 PM
28	Kiss cut by Karin Slaughter	3/9/2015 1:37 PM
29	Northanger Abbey	3/4/2015 3:32 PM
30	Cutting for Stone	3/3/2015 9:41 PM
31	One hundred Years of Solitude	3/3/2015 7:54 PM

32	The Stand in	3/3/2015 3:22 PM
33	Pride and Prejudice	3/3/2015 1:31 PM
#	**Why do you like this book?**	**Date**
1	i just find this book so funny, easy to read and follow.	7/11/2015 11:59 PM
2	It is an extremely look at the weird and wonderful ways of an ordinary family. The stories she tells about some many members of the clan are all a delight	7/8/2015 5:37 PM
3	Made me laugh	7/1/2015 3:15 PM
4	Great extrapolation of the situation at the time to a potential future	6/30/2015 10:03 PM
5	Fantastic world-building and characters you care about	6/30/2015 10:47 AM
6	Beautifully written	6/26/2015 5:21 PM
7	Well written dialogue and characters	6/26/2015 12:54 PM
8	the whole family like his books fast paced and exciting	6/26/2015 7:39 AM
9	read the book then see the film	6/26/2015 7:10 AM
10	Favourite from my childhood (about horses!)	6/26/2015 12:56 AM
11	Interesting dystopian that changes viewpoint and time a lot	6/25/2015 8:20 PM
12	The huge sweep historically and emotionally	6/25/2015 7:14 PM
13	very descriptive and a long saga	6/25/2015 7:09 PM
14	as above	6/25/2015 7:04 PM
15	This work provides an enjoyable trip into Victorian England whilst providing enjoyable mysteries that stand up to being re-read time and again.	6/25/2015 5:49 PM
16	Probably the finest sustained prose in the history of literature. Every sentence tells you something new yet leaves you wanting to know more.	6/25/2015 4:45 PM
17	What a concept!	6/25/2015 4:27 PM
18	language/structure	6/25/2015 4:20 PM
19	Deals charmingly with a moral problem	5/8/2015 8:48 PM
20	Wonderful characterisation	3/30/2015 4:27 PM
21	thought it was going to be pants but surprisingly complex and intriguing	3/9/2015 9:36 PM
22	must-read	3/9/2015 9:07 PM
23	Ideas and an easy read	3/9/2015 8:17 PM
24	i have recently read this and just loved it	3/9/2015 2:16 PM
25	Immerses you in a whole new world.	3/9/2015 1:58 PM
26	A wonderful combination of optimism & sarcasm... Is everything for the best?	3/9/2015 1:57 PM
27	Good clear character back stories, grabbed me in the first 3 pages	3/9/2015 1:37 PM

Q6 This study is to determine to what extent computers can tell the difference between a good book and a poor one. Do you think this is feasible? Please give a reason for your answer.

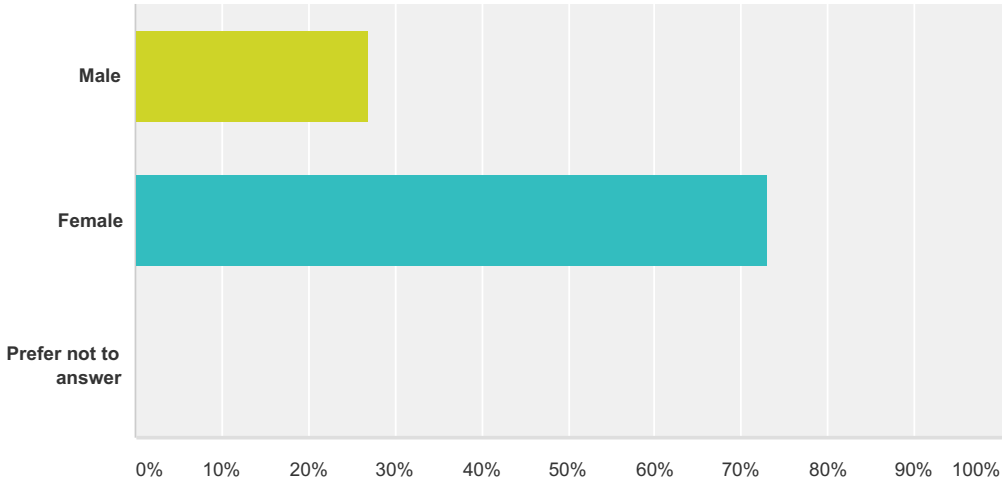
Answered: 33 Skipped: 20

#	Responses	Date
1	As a "good book" is dependent on personal taste, it wouldn't be feasible that a computer could predict what individual people would select as good books. After all, there are many books already that weren't expected to be best sellers, but proved to be popular with the public.	7/11/2015 11:59 PM
2	Appreciation of literature must surely be subjective. Until computers can experience emotion I would not think feasible that they can determine what makes a good book.	7/8/2015 5:37 PM
3	Only by analysing data provided by readers.	7/1/2015 3:15 PM
4	No, characterisation depends on the reader's understanding of people, then the reader's belief in the author's understanding of those same types of people	6/30/2015 10:03 PM
5	No. Firstly, this is based on the assumption that there is a clear divide between 'good books' and 'bad books' - in reality, people feel very differently about the same book. For example, Fifty Shades of Grey has been criticised for having been poorly written, full of unlikeable characters, glamourising abuse etc. It's also a huge publishing success, having sold more than 125 million copies. If you asked 100 people however, some would say it's a good book and others would say it's a bad book - it's not a black and white, binary concept, and a 'good book' is one that connects with you personally as a reader, not the one that is technically and grammatically correct, or the one that has the correct elements to make the formula of a 'good' book. Secondly, different people have different preferences. For example, characters I love and connect with are an absolute necessity to me, whereas a gripping plot would be necessary for my partner. Even among people who prioritise the same elements (e.g. characters), some people will connect with a cast of characters and others won't. Furthermore, most people have a book that they love, completely unexpectedly - someone who doesn't like sci-fi may be surprised to love The Hitchiker's Guide to the Galaxy, or a contemporary romance fan may love Harry Potter, even though they don't like fantasy and they prefer small casts of characters. The book doesn't tick the boxes they were necessarily looking for, and yet they love it.	6/30/2015 10:47 AM
6	Probably not as everyone has different views and I will probably choose 3 different favourite books next! People change their minds etc	6/26/2015 5:21 PM
7	I don't believe current algorithms are able to detect good dialogue flow. If the subject matter was more important then maybe.	6/26/2015 12:54 PM
8	hmmm my concern would be the algorithms used to programme the computers gives certain answers, how would programme for individual taste? You could use these results but how can you programme a readers mood, time to read to say what is a good or poor one	6/26/2015 7:39 AM
9	No, because no two people are the same and everyone has different tastes	6/26/2015 7:10 AM
10	no. I think it is very much the judgement of the reader (although a computer may well be able to judge the quality of writing, impact of the story is a personal and very human thing)	6/26/2015 12:56 AM
11	Until one passes the Turing test, I find it unlikely. Opinions for good and poor are far too subjective.	6/25/2015 8:20 PM
12	Not really.	6/25/2015 7:14 PM
13	No, because i think you books are very individual , My son can choose me books better than myself	6/25/2015 7:09 PM
14	No. I consider that whether we find a book good is driven by its quality but also the reader's emotional context and desires at the time which a computer cannot anticipate or emulate.	6/25/2015 7:04 PM
15	Possibly. In the past thirty years, computers have gone from just doing mathematical tasks for accountants to being present in nearly all areas of life. Their future potential only depends upon the skills of the programmers. As to whether they can tell the difference between a good and a poor book, the question has to be "Can humans tell the difference?" Because a computer can only be taught what a human knows how to do. If I teach a computer what they are, then I'll teach it what I know. But what suits me, doesn't suit everyone. Like with Weather prediction, book quality prediction will be general. You may get it better than pure chance, but you'll never get 100% accuracy.	6/25/2015 5:49 PM
16	No. This is a philistine idea. The soul exists. A great book taps into it, and a computer cannot. Better to ask a group of thoughtful readers. Claims for artificial intelligence are exaggerated because the comparison is always with a single human, not a group.	6/25/2015 4:45 PM

17	Tricky one! If a computer can figure my mood why not!	6/25/2015 4:27 PM
18	I think given enough money any problem is trivial :) But on a serious note: determining a good book is, to my opinion, a very subjective process so much as with wine and it will probably give similar results (people will disagree, basically).	6/25/2015 4:23 PM
19	Possible. Might be able to differentiate between literary/non literary language. Not sure if it can account for individual emotional reaction to a book.	6/25/2015 4:20 PM
20	I don't think they can - while they can recognise use of language, probably evaluate development of a plot, I rather doubt they can have that "Aaahh!" of charm and later remember it in a thinking way.	5/8/2015 8:48 PM
21	Only in as much as statistically they can tell how many people would recommend a particular book.	3/30/2015 4:27 PM
22	Probably but my head would explode if I thought about it for too long ... Computers are controlled by wee evil pixies?	3/9/2015 9:36 PM
23	No. Opinions are subjective.	3/9/2015 9:07 PM
24	A few years ago binatone tennis was a marvel. Now anything is possible.	3/9/2015 8:17 PM
25	Not really. Computers are not emotional	3/9/2015 2:16 PM
26	Not yet. Probably unable to understand the descriptions and tone of a book.	3/9/2015 1:58 PM
27	Yes providing that you can identify what is good or bad manually.	3/9/2015 1:57 PM
28	Books and the storyline are somewhat subjective to the reader. What 1 person may really pick up on, another may skim over but depending on the readers personal feelings depends on the spin the reader puts on the characters. The persons imagination fleshes out the images a book evokes. A computer can't do that.	3/9/2015 1:37 PM
29	They were easy to read, but still gripping.	3/4/2015 3:32 PM
30	They explored a world I had never encountered and they took me there.	3/3/2015 9:41 PM
31	The language, the magic, the emotion	3/3/2015 7:54 PM
32	Love story, love conquering all with a bit of fear thrown in. Observations on society and communities, and wry humour.	3/3/2015 3:22 PM
33	Social commentary and wonderful use of language.	3/3/2015 1:31 PM

Q7 I am

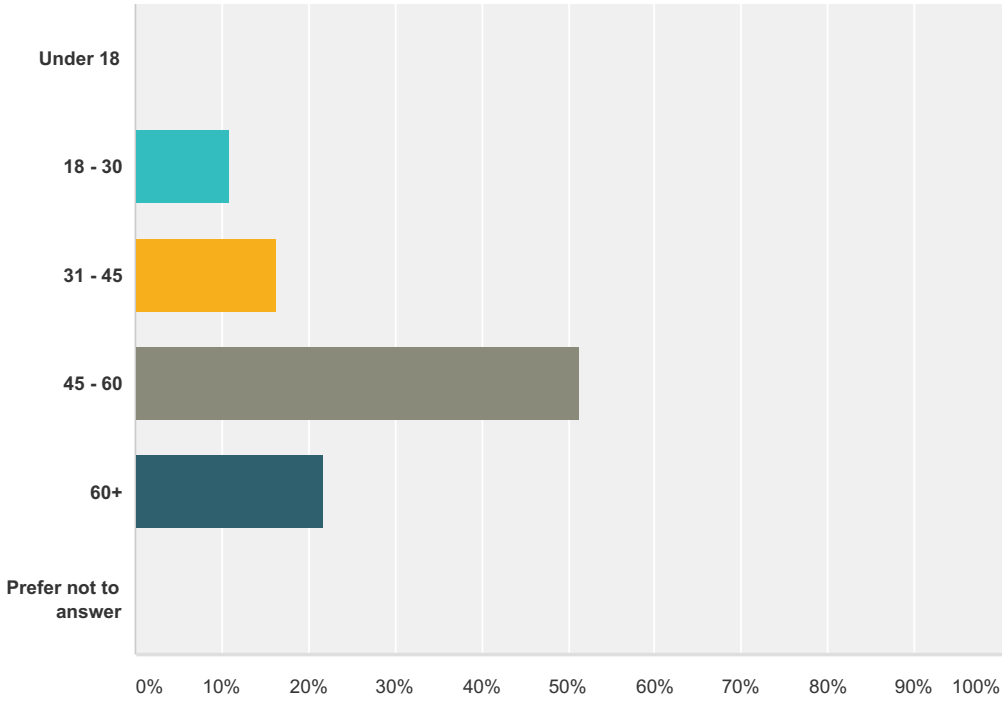
Answered: 37 Skipped: 16



Answer Choices	Responses	
Male	**27.03%**	10
Female	**72.97%**	27
Prefer not to answer	**0.00%**	0
Total		**37**

Q8 My age range is

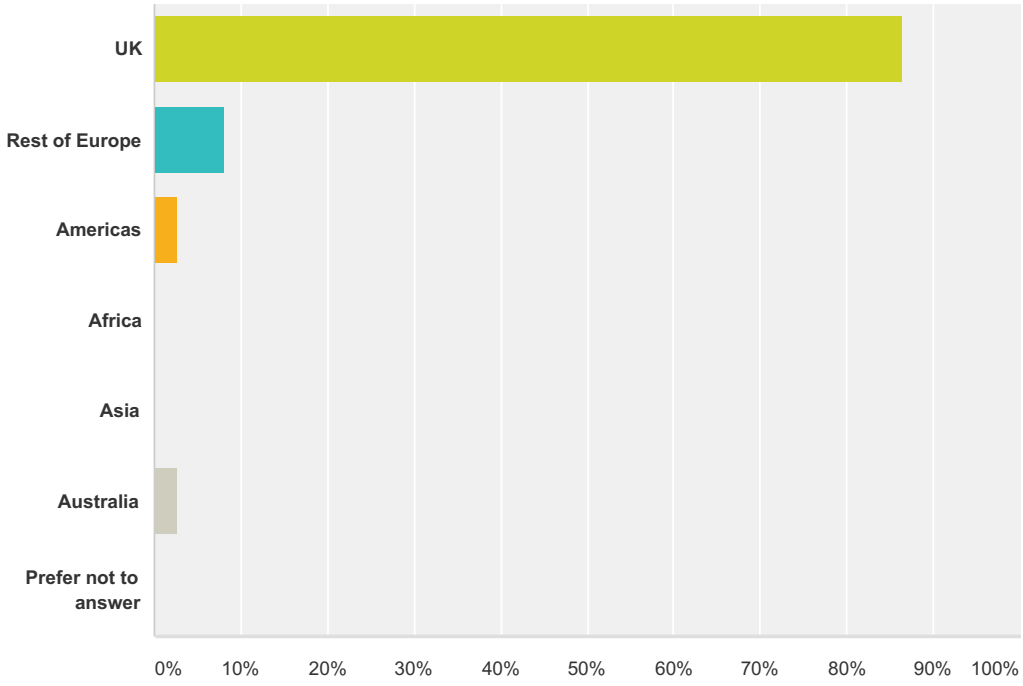
Answered: 37 Skipped: 16



Answer Choices	Responses	
Under 18	**0.00%**	0
18 - 30	**10.81%**	4
31 - 45	**16.22%**	6
45 - 60	**51.35%**	19
60+	**21.62%**	8
Prefer not to answer	**0.00%**	0
Total		**37**

Q9 I live in

Answered: 37 Skipped: 16



Answer Choices	Responses	
UK	**86.49%**	32
Rest of Europe	**8.11%**	3
Americas	**2.70%**	1
Africa	**0.00%**	0
Asia	**0.00%**	0
Australia	**2.70%**	1
Prefer not to answer	**0.00%**	0
Total		**37**

Appendix C

Interview Questions

Interview were carried out in 2013 and 2014. This appendix shows the information given to participants.

Semi-structured interviews

Information for participants

Thank you for your interest in being interviewed for my research. My name is Tess Crosbie and I am a PhD candidate at the University of Bedfordshire. The area I am investigating is to see whether a computer can recognise good literature from something that is less well-written and then apply some qualitative judgement to a text. Of course, the first question is “what is good literature?” and that is one of the aspects this interview will address, but my main interest here is to determine how English Literature is currently taught to children. If there are simple features we can teach children to look for in texts, the same processes may be useful in getting the computer to find them.

Your answers will be anonymised and not identified with you personally. You are under no obligation to answer questions you prefer not to and you may withdraw your participation at any time. The final thesis will be available after publication and I will send you a link at that time to read the paper if you would like to do so.

Questions to cover

What stylistic features do you get students to look for?

What techniques do you use to identify stylistic features?

How easy is it for students to find these features? Age ranges of abilities?

How much emphasis is put on:

- Plot
- Structure
- Language
- Theme

Appendix D

Entropy

The program to calculate entropy is the one developed by Kan and Gero (2009) from Torres (2002).

```
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <math.h>
#include <ctype.h>
/* word/frequency */
typedef struct pf {
	char * p;
	int f;
	struct pf * next;
} pafr;

void insere(char * p, pafr * * ps)
{
	int i = p[0] - 'a';
	pafr * pc = (i>=0 && i<=25) ? ps[i] : ps[26];
	pafr * * pa = (i>=0 && i<=25) ? ps+i : ps+26;
	while (pc) {
		if (!strcmp(pc->p,p)) {
			pc->f++;
			return;
		}
		pa = &(pc->next);
		pc = pc->next;
	}
	if (!(pc = (pafr *) malloc(sizeof(pafr))))
		fprintf(stderr,"Memory allocation error!\n"), exit(1);
	if (!(pc->p = (char *) malloc(strlen(p)+1)))
		fprintf(stderr,"Memory allocation error!\n"), exit(1);
	strcpy(pc->p,p), pc->f = 1, pc->next = 0;
	*pa = pc;
}
```

```
double calculaEntropia(pafr * * pal, double logntp)
{
	int i = 0;
	double et = 0;
	pafr * p = pal[0], * pa = p;
	for ( ; i < 27; p = pa = pal[++i]) {
		while (p) {
			et += (logntp-log10(p->f)) * p->f;
			free(p->p);
			p = p->next;
			free(pa), pa = p;
		}
	}
	return et;
}

int readInput(pafr * * pal)
{
	char pFilePath[255];
	FILE *fp;
	int ntp = 0;
	char palavra[21], c, *p;
	int i=0;
	int iEof = 0;

	printf("Enter file address:");
	scanf("%s",pFilePath);
	fp=fopen(pFilePath, "r");
	if (fp == NULL) {
		printf("Failed to open file");
		exit(0);
```

```
   }
        while (1) {
                i=0;
                do
                {
                        c = fgetc(fp);
                        iEof = feof(fp);
                        palavra[i] = c;
                        i++;
                }
                while ((!iEof)&&(!(c=='"'||c==' '||c=='\n'||c=='.'||c==','||c==':'||c==';'||c=='!'
||c=='?'||c=='('||c==')')));
      if (iEof && i==1) return ntp;
                else if(i==1) continue;
                palavra[i-1] = '\0';
                ntp++, p = palavra;
                while (*p != '\0')
      {
         *p= tolower(*p);
         p++;
      }
                insere(palavra,pal);
          }
}
int main()
{
        /* Array initialized with zeros because static */
        static pafr * palavras[27];
        int ntp = readInput(palavras);
        double logntp = log10(ntp), et = calculaEntropia(palavras,logntp)/ntp;
        printf("%d\n%.1f\n%.0f\n",ntp,et,(et/logntp)*100);
        return 0;
}
```

Appendix E

Literary Quality

This online survey was carried out during the latter half of 2012 and includes the instructions given and the passages used to assess literary merit.

Literary Quality

Welcome to my survey. This is part of a project investigating the ability of computers to recognise aesthetic qualities in literature. In order to build some rules into the computer model, we need to find out what humans think first. In the context of this survey, literariness is defined as something that is well written and that conveys the author's message to you. There are no right or wrong answers here - what you think is literary, someone else might disagree. That is not important. There are ten short passages.

The demographic questions allow us to ensure a spread of respondents. Your answers will remain anonymous. Results of the survey will be available at http://research.tesscrosbie.com/ by 31st December 2012.

Thank you.

What is your age band?

- 18-30
- 31-40
- 41-50
- 51-60
- 61-70
- 70+

Gender

- Male
- Female
- Prefer not to answer

Is English your native language?

- Yes
- No
- Prefer not to answer

To which level have you studied English Literature?

- Degree
- Postgraduate
- Neither of the above

Which area are you in?

- Europe
- Americas
- Asia
- Australia
- Africa
- Prefer not to answer

For the following passage, please rate it according to how literary you find it

- 1 (no literary quality)
- 2
- 3
- 4
- 5 (very literary)

[The following ten passages were used with the same question repeated for each]

1. "Black shapes crouched, lay, sat between the trees leaning against the trunks, clinging to the earth, half coming out, half effaced within the dim light, in all the attitudes of pain, abandonment, and despair. Another mine on the cliff went off, followed by a slight shudder of the soil under my feet. The work was going on. The work! And this was the place where some of the helpers had withdrawn to die. "They were dying slowly—it was very clear. They were not enemies, they were not criminals, they were nothing earthly now—nothing but black shadows of disease and starvation, lying confusedly in the greenish gloom. Brought from all the recesses of the coast in all the legality of time contracts, lost in uncongenial surroundings, fed on unfamiliar food, they sickened, became inefficient, and were then allowed to crawl away and rest. These moribund shapes were free as air—and nearly as thin. I began to distinguish the gleam of the eyes under the trees. Then, glancing down, I saw a face near my hand. The black bones reclined at full length with one shoulder against the tree, and slowly the eyelids rose and the sunken eyes looked up at me, enormous and vacant, a kind of blind, white flicker in the depths of the orbs, which died out slowly. The man seemed young—almost a boy—but you know with them it's hard to tell. I found nothing else to do but to offer him one of my good Swede's ship's biscuits I had in my pocket. The fingers closed slowly on it and held—there was no other movement and no other glance. He had tied a bit of white worsted round his neck—Why? Where did he get it? Was it a badge—an ornament—a charm—a propitiatory act? Was there any idea at all connected with it? It looked startling round his black neck, this bit of white thread from beyond the seas. "Near the same tree two more bundles of acute angles sat with their legs drawn up. One, with his chin propped on his knees, stared at nothing, in an intolerable and appalling manner: his brother phantom rested its forehead, as if overcome with a great weariness; and all about others were scattered in every pose of contorted collapse, as in some picture of a massacre or a pestilence. While I stood horror-struck, one of these creatures rose to his hands and knees, and went off on all-fours towards the river to drink. He lapped out of his hand, then sat up in the sunlight, crossing his shins in front of him, and after a time let his woolly head fall on his breastbone."
2. "I said I'd pack. I rather pride myself on my packing. Packing is one of those many things that I feel I know more about than any other person living. (It surprises me myself, sometimes, how many of these subjects there are.) I impressed the fact upon George and Harris, and told them that they had better leave the whole matter entirely to me. They fell into the suggestion with a readiness that had something uncanny about it. George put on a pipe and spread himself over the easy-chair, and Harris cocked his legs on the table and lit a cigar. This was hardly what I intended. What I had meant, of course, was, that I should boss the job, and that Harris and George should potter about under my directions, I pushing them aside every now and then with, "Oh, you—!" "Here, let me do it." "There you are, simple enough!"—really teaching them, as you might say. Their taking it in the way they did irritated me. There is nothing does irritate me more than seeing other people sitting about doing nothing when I'm working. I lived with a man once who used to make me mad that way. He would loll on the sofa and

watch me doing things by the hour together, following me round the room with his eyes, wherever I went. He said it did him real good to look on at me, messing about. He said it made him feel that life was not an idle dream to be gaped and yawned through, but a noble task, full of duty and stern work. He said he often wondered now how he could have gone on before he met me, never having anybody to look at while they worked. Now, I'm not like that. I can't sit still and see another man slaving and working. I want to get up and superintend, and walk round with my hands in my pockets, and tell him what to do. It is my energetic nature. I can't help it. However, I did not say anything, but started the packing. It seemed a longer job than I had thought it was going to be; but I got the bag finished at last, and I sat on it and strapped it. "Ain't you going to put the boots in?" said Harris. And I looked round, and found I had forgotten them. That's just like Harris. He couldn't have said a word until I'd got the bag shut and strapped, of course. And George laughed—one of those irritating, senseless, chuckle-headed, crack-jawed laughs of his. They do make me so wild. I opened the bag and packed the boots in; and then, just as I was going to close it, a horrible idea occurred to me. Had I packed my tooth-brush? I don't know how it is, but I never do know whether I've packed my tooth-brush. My tooth-brush is a thing that haunts me when I'm travelling, and makes my life a misery. I dream that I haven't packed it, and wake up in a cold perspiration, and get out of bed and hunt for it. And, in the morning, I pack it before I have used it, and have to unpack again to get it, and it is always the last thing I turn out of the bag; and then I repack and forget it, and have to rush upstairs for it at the last moment and carry it to the railway station, wrapped up in my pocket-handkerchief. ------Of course I had to turn every mortal thing out now, and, of course, I could not find it. I rummaged the things up into much the same state that they must have been before the world was created, and when chaos reigned. Of course, I found George's and Harris's eighteen times over, but I couldn't find my own. I put the things back one by one, and held everything up and shook it. Then I found it inside a boot. I repacked once more."

3. "Next day I left that station at last, with a caravan of sixty men, for a two-hundred-mile tramp. "No use telling you much about that. Paths, paths, everywhere; a stamped-in network of paths spreading over the empty land, through the long grass, through burnt grass, through thickets, down and up chilly ravines, up and down stony hills ablaze with heat; and a solitude, a solitude, nobody, not a hut. The population had cleared out a long time ago. Well, if a lot of mysterious niggers armed with all kinds of fearful weapons suddenly took to travelling on the road between Deal and Gravesend, catching the yokels right and left to carry heavy loads for them, I fancy every farm and cottage thereabouts would get empty very soon. Only here the dwellings were gone, too. Still I passed through several abandoned villages. There's something pathetically childish in the ruins of grass walls. Day after day, with the stamp and shuffle of sixty pair of bare feet behind me, each pair under a 60-lb. load. Camp, cook, sleep, strike camp, march. Now and then a carrier dead in harness, at rest in the long grass near the path, with an empty water-gourd and his long staff lying by his side. A great silence around and above. Perhaps on some quiet night the tremor of far-off drums, sinking, swelling, a tremor vast, faint; a sound weird, appealing, suggestive, and wild—and perhaps with as profound a meaning as the sound of bells in a Christian country. Once a white man in an unbuttoned uniform, camping on the path with an armed escort of lank Zanzibaris, very hospitable and festive—not to say drunk. Was looking after the upkeep of the road, he declared. Can't say I saw any road or any upkeep, unless the body of a middle-aged negro, with a bullet-hole in the forehead, upon which I absolutely stumbled three miles farther on, may be considered as a permanent improvement. I had a white companion, too, not a bad chap, but rather too fleshy and with the exasperating habit of fainting on the hot hillsides, miles away from the least bit

of shade and water. Annoying, you know, to hold your own coat like a parasol over a man's head while he is coming to. I couldn't help asking him once what he meant by coming there at all. 'To make money, of course. What do you think?' he said, scornfully. Then he got fever, and had to be carried in a hammock slung under a pole."

4. "They awe us, these strange stars, so cold, so clear. We are as children whose small feet have strayed into some dim-lit temple of the god they have been taught to worship but know not; and, standing where the echoing dome spans the long vista of the shadowy light, glance up, half hoping, half afraid to see some awful vision hovering there. And yet it seems so full of comfort and of strength, the night. In its great presence, our small sorrows creep away, ashamed. The day has been so full of fret and care, and our hearts have been so full of evil and of bitter thoughts, and the world has seemed so hard and wrong to us. Then Night, like some great loving mother, gently lays her hand upon our fevered head, and turns our little tear-stained faces up to hers, and smiles; and, though she does not speak, we know what she would say, and lay our hot flushed cheek against her bosom, and the pain is gone. Sometimes, our pain is very deep and real, and we stand before her very silent, because there is no language for our pain, only a moan. Night's heart is full of pity for us: she cannot ease our aching; she takes our hand in hers, and the little world grows very small and very far away beneath us, and, borne on her dark wings, we pass for a moment into a mightier Presence than her own, and in the wondrous light of that great Presence, all human life lies like a book before us, and we know that Pain and Sorrow are but the angels of God. Only those who have worn the crown of suffering can look upon that wondrous light; and they, when they return, may not speak of it, or tell the mystery they know. Once upon a time, through a strange country, there rode some goodly knights, and their path lay by a deep wood, where tangled briars grew very thick and strong, and tore the flesh of them that lost their way therein. And the leaves of the trees that grew in the wood were very dark and thick, so that no ray of light came through the branches to lighten the gloom and sadness. And, as they passed by that dark wood, one knight of those that rode, missing his comrades, wandered far away, and returned to them no more; and they, sorely grieving, rode on without him, mourning him as one dead. Now, when they reached the fair castle towards which they had been journeying, they stayed there many days, and made merry; and one night, as they sat in cheerful ease around the logs that burned in the great hall, and drank a loving measure, there came the comrade they had lost, and greeted them. His clothes were ragged, like a beggar's, and many sad wounds were on his sweet flesh, but upon his face there shone a great radiance of deep joy. And they questioned him, asking him what had befallen him: and he told them how in the dark wood he had lost his way, and had wandered many days and nights, till, torn and bleeding, he had lain him down to die. Then, when he was nigh unto death, lo! through the savage gloom there came to him a stately maiden, and took him by the hand and led him on through devious paths, unknown to any man, until upon the darkness of the wood there dawned a light such as the light of day was unto but as a little lamp unto the sun; and, in that wondrous light, our way-worn knight saw as in a dream a vision, and so glorious, so fair the vision seemed, that of his bleeding wounds he thought no more, but stood as one entranced, whose joy is deep as is the sea, whereof no man can tell the depth. And the vision faded, and the knight, kneeling upon the ground, thanked the good saint who into that sad wood had strayed his steps, so he had seen the vision that lay there hid. And the name of the dark forest was Sorrow; but of the vision that the good knight saw therein we may not speak nor tell."
5. "The day was ending in a serenity of still and exquisite brilliance. The water shone pacifically; the sky, without a speck, was a benign immensity of unstained light; the very mist on the Essex marsh was like a gauzy and radiant fabric, hung from the wooded rises inland, and draping the

low shores in diaphanous folds. Only the gloom to the west, brooding over the upper reaches, became more sombre every minute, as if angered by the approach of the sun. And at last, in its curved and imperceptible fall, the sun sank low, and from glowing white changed to a dull red without rays and without heat, as if about to go out suddenly, stricken to death by the touch of that gloom brooding over a crowd of men. Forthwith a change came over the waters, and the serenity became less brilliant but more profound. The old river in its broad reach rested unruffled at the decline of day, after ages of good service done to the race that peopled its banks, spread out in the tranquil dignity of a waterway leading to the uttermost ends of the earth. We looked at the venerable stream not in the vivid flush of a short day that comes and departs for ever, but in the august light of abiding memories. And indeed nothing is easier for a man who has, as the phrase goes, "followed the sea" with reverence and affection, that to evoke the great spirit of the past upon the lower reaches of the Thames. The tidal current runs to and fro in its unceasing service, crowded with memories of men and ships it had borne to the rest of home or to the battles of the sea. It had known and served all the men of whom the nation is proud, from Sir Francis Drake to Sir John Franklin, knights all, titled and untitled—the great knights-errant of the sea. It had borne all the ships whose names are like jewels flashing in the night of time, from the Golden Hind returning with her rotund flanks full of treasure, to be visited by the Queen's Highness and thus pass out of the gigantic tale, to the Erebus and Terror, bound on other conquests—and that never returned. It had known the ships and the men. They had sailed from Deptford, from Greenwich, from Erith—the adventurers and the settlers; kings' ships and the ships of men on 'Change; captains, admirals, the dark "interlopers" of the Eastern trade, and the commissioned "generals" of East India fleets. Hunters for gold or pursuers of fame, they all had gone out on that stream, bearing the sword, and often the torch, messengers of the might within the land, bearers of a spark from the sacred fire. What greatness had not floated on the ebb of that river into the mystery of an unknown earth!... The dreams of men, the seed of commonwealths, the germs of empires."

6. "I remember going to the British Museum one day to read up the treatment for some slight ailment of which I had a touch—hay fever, I fancy it was. I got down the book, and read all I came to read; and then, in an unthinking moment, I idly turned the leaves, and began to indolently study diseases, generally. I forget which was the first distemper I plunged into—some fearful, devastating scourge, I know—and, before I had glanced half down the list of “premonitory symptoms,” it was borne in upon me that I had fairly got it. I sat for awhile, frozen with horror; and then, in the listlessness of despair, I again turned over the pages. I came to typhoid fever—read the symptoms—discovered that I had typhoid fever, must have had it for months without knowing it—wondered what else I had got; turned up St. Vitus’s Dance—found, as I expected, that I had that too,—began to get interested in my case, and determined to sift it to the bottom, and so started alphabetically—read up ague, and learnt that I was sickening for it, and that the acute stage would commence in about another fortnight. Bright’s disease, I was relieved to find, I had only in a modified form, and, so far as that was concerned, I might live for years. Cholera I had, with severe complications; and diphtheria I seemed to have been born with. I plodded conscientiously through the twenty-six letters, and the only malady I could conclude I had not got was housemaid’s knee. I felt rather hurt about this at first; it seemed somehow to be a sort of slight. Why hadn’t I got housemaid’s knee? Why this invidious reservation? After a while, however, less grasping feelings prevailed. I reflected that I had every other known malady in the pharmacology, and I grew less selfish, and determined to do without housemaid’s knee. Gout, in its most malignant stage, it would appear, had seized me without my being aware of it; and zymosis I had evidently been suffering with from boyhood. There were no more diseases after zymosis,

so I concluded there was nothing else the matter with me. I sat and pondered. I thought what an interesting case I must be from a medical point of view, what an acquisition I should be to a class! Students would have no need to “walk the hospitals,” if they had me. I was a hospital in myself. All they need do would be to walk round me, and, after that, take their diploma. Then I wondered how long I had to live. I tried to examine myself. I felt my pulse. I could not at first feel any pulse at all. Then, all of a sudden, it seemed to start off. I pulled out my watch and timed it. I made it a hundred and forty-seven to the minute. I tried to feel my heart. I could not feel my heart. It had stopped beating. I have since been induced to come to the opinion that it must have been there all the time, and must have been beating, but I cannot account for it. I patted myself all over my front, from what I call my waist up to my head, and I went a bit round each side, and a little way up the back. But I could not feel or hear anything. I tried to look at my tongue. I stuck it out as far as ever it would go, and I shut one eye, and tried to examine it with the other. I could only see the tip, and the only thing that I could gain from that was to feel more certain than before that I had scarlet fever. I had walked into that reading-room a happy, healthy man. I crawled out a decrepit wreck."

7. "As we had plenty of wood, and caution was the word, I brought up in the middle of the stream. The reach was narrow, straight, with high sides like a railway cutting. The dusk came gliding into it long before the sun had set. The current ran smooth and swift, but a dumb immobility sat on the banks. The living trees, lashed together by the creepers and every living bush of the undergrowth, might have been changed into stone, even to the slenderest twig, to the lightest leaf. It was not sleep—it seemed unnatural, like a state of trance. Not the faintest sound of any kind could be heard. You looked on amazed, and began to suspect yourself of being deaf—then the night came suddenly, and struck you blind as well. About three in the morning some large fish leaped, and the loud splash made me jump as though a gun had been fired. When the sun rose there was a white fog, very warm and clammy, and more blinding than the night. It did not shift or drive; it was just there, standing all round you like something solid. At eight or nine, perhaps, it lifted as a shutter lifts. We had a glimpse of the towering multitude of trees, of the immense matted jungle, with the blazing little ball of the sun hanging over it—all perfectly still—and then the white shutter came down again, smoothly, as if sliding in greased grooves. I ordered the chain, which we had begun to heave in, to be paid out again. Before it stopped running with a muffled rattle, a cry, a very loud cry, as of infinite desolation, soared slowly in the opaque air. It ceased. A complaining clamour, modulated in savage discords, filled our ears. The sheer unexpectedness of it made my hair stir under my cap. I don't know how it struck the others: to me it seemed as though the mist itself had screamed, so suddenly, and apparently from all sides at once, did this tumultuous and mournful uproar arise. It culminated in a hurried outbreak of almost intolerably excessive shrieking, which stopped short, leaving us stiffened in a variety of silly attitudes, and obstinately listening to the nearly as appalling and excessive silence. 'Good God! What is the meaning—' stammered at my elbow one of the pilgrims—a little fat man, with sandy hair and red whiskers, who wore sidespring boots, and pink pyjamas tucked into his socks. Two others remained open-mouthed a while minute, then dashed into the little cabin, to rush out incontinently and stand darting scared glances, with Winchesters at 'ready' in their hands. What we could see was just the steamer we were on, her outlines blurred as though she had been on the point of dissolving, and a misty strip of water, perhaps two feet broad, around her—and that was all. The rest of the world was nowhere, as far as our eyes and ears were concerned. Just nowhere. Gone, disappeared; swept off without leaving a whisper or a shadow behind."

8. "It was the dead body of a woman. It lay very lightly on the water, and the face was sweet and calm. It was not a beautiful face; it was too prematurely aged-looking, too thin and drawn, to be that; but it was a gentle, lovable face, in spite of its stamp of pinch and poverty, and upon it was that look of restful peace that comes to the faces of the sick sometimes when at last the pain has left them. Fortunately for us—we having no desire to be kept hanging about coroners' courts—some men on the bank had seen the body too, and now took charge of it from us. We found out the woman's story afterwards. Of course it was the old, old vulgar tragedy. She had loved and been deceived—or had deceived herself. Anyhow, she had sinned—some of us do now and then—and her family and friends, naturally shocked and indignant, had closed their doors against her. Left to fight the world alone, with the millstone of her shame around her neck, she had sunk ever lower and lower. For a while she had kept both herself and the child on the twelve shillings a week that twelve hours' drudgery a day procured her, paying six shillings out of it for the child, and keeping her own body and soul together on the remainder. Six shillings a week does not keep body and soul together very unitedly. They want to get away from each other when there is only such a very slight bond as that between them; and one day, I suppose, the pain and the dull monotony of it all had stood before her eyes plainer than usual, and the mocking spectre had frightened her. She had made one last appeal to friends, but, against the chill wall of their respectability, the voice of the erring outcast fell unheeded; and then she had gone to see her child—had held it in her arms and kissed it, in a weary, dull sort of way, and without betraying any particular emotion of any kind, and had left it, after putting into its hand a penny box of chocolate she had bought it, and afterwards, with her last few shillings, had taken a ticket and come down to Goring. It seemed that the bitterest thoughts of her life must have centred about the wooded reaches and the bright green meadows around Goring; but women strangely hug the knife that stabs them, and, perhaps, amidst the gall, there may have mingled also sunny memories of sweetest hours, spent upon those shadowed deeps over which the great trees bend their branches down so low. She had wandered about the woods by the river's brink all day, and then, when evening fell and the grey twilight spread its dusky robe upon the waters, she stretched her arms out to the silent river that had known her sorrow and her joy. And the old river had taken her into its gentle arms, and had laid her weary head upon its bosom, and had hushed away the pain. Thus had she sinned in all things—sinned in living and in dying. God help her! and all other sinners, if any more there be. "
9. "One ship is very much like another, and the sea is always the same. In the immutability of their surroundings the foreign shores, the foreign faces, the changing immensity of life, glide past, veiled not by a sense of mystery but by a slightly disdainful ignorance; for there is nothing mysterious to a seaman unless it be the sea itself, which is the mistress of his existence and as inscrutable as Destiny. For the rest, after his hours of work, a casual stroll or a casual spree on shore suffices to unfold for him the secret of a whole continent, and generally he finds the secret not worth knowing. The yarns of seamen have a direct simplicity, the whole meaning of which lies within the shell of a cracked nut. But Marlow was not typical (if his propensity to spin yarns be excepted), and to him the meaning of an episode was not inside like a kernel but outside, enveloping the tale which brought it out only as a glow brings out a haze, in the likeness of one of these misty halos that sometimes are made visible by the spectral illumination of moonshine."
10. "I was sitting on the bank, conjuring up this scene to myself, when George remarked that when I was quite rested, perhaps I would not mind helping to wash up; and, thus recalled from the days of the glorious past to the prosaic present, with all its misery and sin, I slid down into the boat and cleaned out the frying-pan with a stick of wood and a tuft of grass, polishing it up

finally with George's wet shirt. We went over to Magna Charta Island, and had a look at the stone which stands in the cottage there and on which the great Charter is said to have been signed; though, as to whether it really was signed there, or, as some say, on the other bank at"Runningmede," I decline to commit myself. As far as my own personal opinion goes, however, I am inclined to give weight to the popular island theory. Certainly, had I been one of the Barons, at the time, I should have strongly urged upon my comrades the advisability of our getting such a slippery customer as King John on to the island, where there was less chance of surprises and tricks. There are the ruins of an old priory in the grounds of Ankerwyke House, which is close to Picnic Point, and it was round about the grounds of this old priory that Henry VIII. is said to have waited for and met Anne Boleyn. He also used to meet her at Hever Castle in Kent, and also somewhere near St. Albans. It must have been difficult for the people of England in those days to have found a spot where these thoughtless young folk were not spooning. Have you ever been in a house where there are a couple courting? It is most trying. You think you will go and sit in the drawing-room, and you march off there. As you open the door, you hear a noise as if somebody had suddenly recollected something, and, when you get in, Emily is over by the window, full of interest in the opposite side of the road, and your friend, John Edward, is at the other end of the room with his whole soul held in thrall by photographs of other people's relatives. "Oh!" you say, pausing at the door, "I didn't know anybody was here." "Oh! didn't you?" says Emily, coldly, in a tone which implies that she does not believe you. You hang about for a bit, then you say: "It's very dark. Why don't you light the gas?" John Edward says, "Oh!" he hadn't noticed it; and Emily says that papa does not like the gas lit in the afternoon. You tell them one or two items of news, and give them your views and opinions on the Irish question; but this does not appear to interest them. All they remark on any subject is,"Oh!" "Is it?" "Did he?" "Yes," and "You don't say so!" And, after ten minutes of such style of conversation, you edge up to the door, and slip out, and are surprised to find that the door immediately closes behind you, and shuts itself, without your having touched it. Half an hour later, you think you will try a pipe in the conservatory. The only chair in the place is occupied by Emily; and John Edward, if the language of clothes can be relied upon, has evidently been sitting on the floor. They do not speak, but they give you a look that says all that can be said in a civilised community; and you back out promptly and shut the door behind you. You are afraid to poke your nose into any room in the house now; so, after walking up and down the stairs for a while, you go and sit in your own bedroom. This becomes uninteresting, however, after a time, and so you put on your hat and stroll out into the garden. You walk down the path, and as you pass the summer-house you glance in, and there are those two young idiots, huddled up into one corner of it; and they see you, and are evidently under the idea that, for some wicked purpose of your own, you are following them about."

Confirmation page

Thank you for your participation. Results will be posted at http://research.tesscrosbie.com/ by 31st December 2012.

If you are interested, the texts were taken from "Three Men in a Boat" and "Heart of Darkness", available from Project Gutenberg at http://www.gutenberg.org/ebooks/308 and http://www.gutenberg.org/ebooks/219 respectively.

Appendix F

Factor Analysis

Factor analysis was used to determine the variables that combine to indicate literary merit.

Unrotated Factor Loadings and Communalities
103 cases used 4 cases contain missing values

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8
CC	0.455	-0.044	-0.233	-0.007	-0.144	-0.479	0.391	-0.135
CD	-0.645	-0.088	0.410	0.426	-0.044	-0.259	-0.194	0.049
DT	-0.547	0.131	0.026	0.583	0.123	-0.271	-0.032	-0.219
EX	0.514	0.268	0.123	-0.226	0.068	-0.193	0.082	0.232
IN	0.248	-0.603	-0.563	-0.202	-0.169	0.051	-0.009	0.130
JJ	-0.100	-0.238	-0.125	-0.505	-0.324	-0.045	-0.078	-0.228
JJR	-0.036	-0.633	0.062	-0.336	0.348	-0.189	0.058	0.242
JJS	0.098	-0.727	-0.023	-0.243	0.279	-0.117	0.367	-0.061
MD	0.314	-0.378	0.646	0.019	0.240	-0.052	-0.194	0.134
NN	-0.538	-0.379	-0.405	0.152	-0.025	-0.045	-0.148	0.018
NNS	-0.575	-0.471	-0.083	-0.245	-0.334	-0.181	-0.072	0.037
NNP	-0.490	0.501	0.264	-0.301	0.089	0.351	0.140	0.214
NNPS	-0.552	-0.246	0.144	-0.214	-0.288	0.018	0.395	0.021
PDT	0.566	0.050	0.058	-0.048	-0.062	-0.496	0.320	0.049
POS	0.117	0.427	-0.060	-0.332	0.446	0.207	-0.053	-0.068
PRP	0.881	0.199	0.114	-0.014	-0.011	0.036	0.091	-0.061
PRPs	0.678	-0.205	-0.391	0.157	-0.061	0.304	-0.046	0.041
RB	0.787	-0.077	0.092	-0.216	0.192	-0.064	0.083	-0.080
RBR	0.082	-0.493	0.119	-0.370	0.521	0.050	-0.239	-0.271
RBS	0.092	-0.621	0.138	0.002	0.428	0.104	0.431	-0.025
RP	0.286	0.592	-0.189	-0.330	-0.196	-0.346	-0.181	-0.172
TO	0.576	-0.476	0.212	-0.028	-0.043	0.095	-0.496	0.060
VB	0.480	-0.350	0.645	-0.032	0.073	-0.086	-0.324	0.091
VBD	0.597	0.422	-0.534	-0.028	0.145	0.062	0.034	0.085
VBG	-0.138	0.066	-0.376	-0.351	0.286	0.108	-0.323	-0.244
VBN	0.117	-0.630	-0.209	0.262	-0.145	0.220	-0.062	0.448
VBP	0.254	0.115	0.758	-0.234	-0.282	0.164	0.156	0.046
VBZ	-0.390	-0.232	0.565	-0.281	-0.364	0.092	0.171	-0.252
WDT	0.062	-0.727	-0.229	0.161	-0.306	0.123	0.039	0.196
WP	0.616	-0.040	0.049	0.077	-0.313	0.365	0.248	-0.084
WPs	-0.046	-0.459	-0.079	-0.173	-0.354	0.299	-0.084	-0.531
WRB	0.516	0.040	0.271	-0.028	-0.426	-0.265	-0.342	0.035
Alliteration	0.260	0.256	-0.194	-0.497	-0.377	-0.067	-0.161	0.285
LexDiv	0.593	-0.044	0.187	0.401	-0.076	0.195	0.050	-0.295
Av.Sent. Length	0.304	-0.637	-0.183	0.140	0.060	-0.379	-0.096	-0.217
Relative entropy	-0.701	-0.050	-0.022	-0.535	0.065	-0.117	-0.053	0.142
Variance	7.6409	5.7301	3.5697	2.7956	2.3700	1.7658	1.7242	1.3509
% Var	0.212	0.159	0.099	0.078	0.066	0.049	0.048	0.038

Bibliography

Afroz, S., Brennan, M. and Greenstadt, R. (2012), Detecting hoaxes, frauds, and deception in writing style online, *in* 'Proceedings of the IEEE Symposium on Security and Privacy', IEEE, pp. 461–475.

Aliu, O. and Chung, K. C. (2010), 'Readability of ASPS and ASAPS educational websites: An analysis of consumer impact', *Plastic and Reconstructive Surgery* **125**(4), 1271–1278.

Anyan, F. (2013), 'The influence of power shifts in data collection and analysis stages: A focus on qualitative research interview', *The Qualitative Report* **18**(Article 36), 1–9.

Ashok, V. G., Feng, S. and Choi, Y. (2013), Success with style: Using writing style to predict the success of novels, *in* 'Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, pp. 1753–1764.

Aydoğan, E. and Akcayol, M. A. (2016), A comprehensive survey for sentiment analysis tasks using machine learning techniques, *in* '2016 International Symposium on Innovations in Intellient Systems and Applications (INISTA)', IEEE.

Balahur, B., Mihalcea, R. and Montoyo, A. (2014), 'Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications', *Computer Speech and Language* **28**(1), 1–6.

Banea, C., Mihalcea, R., Wiebe, J. and Hassan, S. (2008), Multilingual subjectivity analysis using machine translation, *in* 'Proceedings of the Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, pp. 127–135.

Barrie, J. (2014), 'Computers are writing novels: Read a few samples here'. Online at http://uk.businessinsider.com/novels-written-by-computers-2014-11 (Accessed 25th October, 2016).

Barry, P. (2009), *Beginning Theory: an Introduction to Literary and Cultural Theory*, third edn, Manchester University Press.

Becket, F. (2002), *The Complete Critical Guide to D. H. Lawrence*, Routlege.

Bird, S., Loper, E. and Klein, E. (2009), *Natural Language Processing with Python*, O'Reilly Media Inc.

Boychuck, E., Paramonov, I., Kozhemyakin, N. and Kasatkina, N. (2014), Automated approach for rhythm analysis of French literary texts, *in* 'Proceeding of the 15th Conference of FRUCT Association', IEEE, pp. 15–23.

Burrows, J. (2002), 'Delta': a measure of stylistic difference and a guide to likely authorship', *Literary and Linguistic Computing* **17**(3), 267 – 287.

Cambria, E. (2016), 'Affective computing and sentiment analysis', *IEEE Intelligent Systems* **31**(2), 102–107.

Cambria, E., Schuller, B., Xia, Y. and Havasi, C. (2013), 'New avenues in opinion mining and sentiment analysis', *IEEE Intelligent Systems* **28**(2), 15–21.

Campbell, S. M., Braspenning, J., Hutchinson, A. and Marshall, M. (2002), 'Research methods used in developing and applying indicators in primary care', *Quality and Safety in Health Care* **11**, 358–364.

Conrad, J. (1899), *Heart of Darkness*. Online at http://www.gutenberg.org/ebooks/219 (Accessed 15th September, 2014).

Creswell, J. W. (2014), *Research Design: Qualitative, Quantative, and Mixed Method Approaches*, fourth edn, SAGE.

Crosbie, T., French, T. and Conrad, M. (2013*a*), 'Stylistic analysis using machine translation as a tool', *International Journal for Infonomics (IJI)* **Special issue 1**(1).

Crosbie, T., French, T. and Conrad, M. (2013*b*), Towards a model for replicating aesthetic literary appreciation, *in* 'Proceedings of the Fifth Workshop on Semantic Web Information Management', ACM, p. 8.

Crotty, M. (1998), *The Foundations of Social Research: Meaning and Perspective in the Research Process*, SAGE Publications.

Culler, J. (1992), *The Pursuit of Signs*, Routledge.

de Saussure, F. (1983), *Course in General Linguistics*, Duckworth.

Decrop, A., Pizam, A. and Mansfield, Y. (2000), *Consumer Behavior in Travel and Tourism*, Haworth Hospitality Press.

DiCicco-Bloom, B. and Crabtree, B. F. (2006), 'The qualitative research interview', *Medical Education* **40**, 314–321.

Drabble, M., ed. (1996), *The Oxford Companion to English Literature*, Oxford University Press.

Dynes, B. (2014), *Masterclasses in Creative Writing*, Constable and Robinson.

Eagleton, T. (2008), *Literary Theory: an Introduction*, Blackwell Publishing.

Eco, U. (2012), *On Literature*, Random House.

Feng, S., Banerjee, R. and Choi, Y. (2012*a*), Characterizing stylistic elements in syntatic structures, *in* 'Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning', Association for Computational Linguistics, pp. 1522–1533.

Feng, S., Banerjee, R. and Choi, Y. (2012*b*), Syntatic stylometry for deception detection, *in* 'Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers', Vol. 2, Association for Computational Linguistics, pp. 171–175.

Forsyth, R. and Holmes, D. (1996), 'Feature-finding for test classification', *Literary and Linguistic Computing* **11**(4), 163–174.

Gamon, M. (2004), 'Linguistic correlates of style: Authorship classification with deep linguistic analysis features', *Proceedings of the 20th International Conference on Computational Linguistics* **4**(1), 611.

Golban, P. and Ciobanu, E. A. (2008), *A Short History of Literary Criticism*, Üç Mart Press.

Gonçalves, L. and Gonçalves, L. (2006), 'Fractal power law in literary English', *Physica A: Statistical Mechanics and its Applications* **360**, 557–575.

Goss, J. D. and Leinbach, T. R. (1996), 'Focus groups as alternative research practice: Experience with transmigrants in Indonesia', *Area* **28**(2), 115.

Habib, M. A. R. (2005), *A History of Literary Criticism: from Plato to the Present*, Blackwell Publishing.

Haiyan, L. and Xiaohu, Y. (2011), Quantifying the vicissitude of Fitzgerald's creativity, *in* 'Information Technology and Artificial Intelligence Conference (ITAIC)', IEEE, pp. 123–127.

Hall, S. (1973), Encoding/decoding (1973), *in* S. Hall, D. Hobson, A. Love and P. Willis, eds, 'Culture, Media, Language: Working Papers in Cultural Studies (1980)', Hutchinson, pp. 128–138.

Hammond, A., Brooke, J. and Hirst, G. (2013), A tale of two cultures: Bringing literary analysis and computational linguistics together, *in* 'Proceedings of the Second Workshop on Computational Linguistics for Literature', Association for Computational Linguistics, pp. 1–8.

Hardy, T. (2003), *Far From the Madding Crowd*, Penguin Books Ltd.

Hawkes, T. (1977), *Structuralism and Semiotics*, Methuen and Co. Ltd.

Holmes, D. (1985), 'The analysis of literary style: A review', *Journal of the Royal Statistical Society* **Series A**(148), 328 – 341.

Hoppe, M. J., Wells, E. A., Morrison, D. M., Gillmore, M. R. and Wilsdon, A. (1995), 'Using focus groups to discuss sensitive topics with children', *Evaluation Review* **19**(1), 102–114.

Hudson, A. (2012), 'Man or machine - can robots really write novels?'. Online at http://news.bbc.co.uk/1/hi/programmes/click_online/9764416.stm (Accessed 25th October, 2016).

Hudson, L. A. and Ozanne, J. L. (1988), 'Alternative ways of seeking knowledge in consumer research', *The Journal of Consumer Research* **14**(4), 508–521.

Hurtado, J., Taweewitchakreeya, N. and Zhu, X. (2014), Who wrote this paper? Learning for authorship de-identification using stylometric features, *in* '15th International Conference on Information Reuse and Integration (IRI)', IEEE, pp. 859–862.

Jerome, J. K. (1889), *Three Men in a Boat.* Online at http://www.gutenberg.org/ebooks/308 (Accessed 15th September, 2014).

Jockers, M. L. and Mimno, D. (2013), 'Significant themes in 19th century literature', *Poetics* **41**, 750–769.

Johansson, V. (2008), *Lexical Diversity and Lexical Density in Speech and Writing: a Developmental Perspective*, Lund University, Department of Linguistics and Phonetics: Working Papers 53.

Johnson, R. B. and Onwuegbuzie, A. J. (2004), 'Mixed methods research: A research paradigm whose time has come', *Educational Researcher* **33**(7), 14–26.

Jungman, R. E. (2003), 'Trimming Shakespeare's Sonnet 18', *ANQ: A Quarterly Journal of Short Articles, Notes and Reviews* pp. 18–19.

Kan, J. and Gero, J. (2009), Using entropy to measure design creativity using a text based analysis tool on design protocols, *in* 'Digital Proceedings of the International Association of Societies on Design Research', Vol. 1.

Kaplan, D. and Blei, D. (2007), A computational approach to style in American poetry, *in* 'IEEE International Conference on Data Mining (ICDM)', pp. 553–558.

Keim, D. and Oelke, D. (2007), Literature fingerprinting: a new method for visual literary analysis, *in* 'VAST IEEE Symposium on Visual Analytics Science and Technology 2007, Proceedings', IEEE, pp. 115–122.

Kelley, K., Clark, B., Brown, V. and Sitza, J. (2003), 'Good practice in the conduct and reporting of survey research', *International Journal for Quality in Helath Care* **15**(3), 261–266.

Kitzinger, J. (1994), 'The methodology of focus groups: the importance of interaction between research participants', *Sociology of Health* **16**(1), 103–121.

Kitzinger, J. (1995), 'Introducing focus groups', *British Medical Journal* **311**, 299–302.

Krueger, R. A. (2002), 'Designing and conducting focus group interviews'. Online at http://www.eiu.edu/ihec/Krueger-FocusGroupInterviews.pdf Accessed 2nd June, 2015.

Krueger, R. A. and Casey, M. A. (2009), *Focus Groups: a Practical Guide for Applied Research*.

Kubát, M. and Milička, J. (2013), 'Vocabulary richness measure in genres', *Journal of Quantiative Linguistics* **20**(4), 339–349.

Li, J., Zheng, R. and Chen, H. (2006), 'From fingerprint to writeprint', *Communications of the ACM* **49**(4), 76–82.

Li, L., He, Z. and Yi, Y. (2004), 'Poetry stylistic analysis technique based on term connections', *Proceedings of the 2004 International Conference on Machine Learning and Cybernetics* **5**, 26–29.

Luyckx, K., Daelemans, W. and Vanhoutte, E. (2006), Stylogenetics: Clustering-based stylistic analysis of literary corpora, *in* 'Workshop: Towards Computation Models of Literary Analysis', pp. 30–35.

Mackenzie, N. and Knipe, S. (2006), 'Research dilemmas: Paradigms, methods and methodology', *Issues in Educational Research* **16**(2).

McSporran, C. (2005), *Daughters of Lilith: Witches and Wicked Women in The Chronicals of Narnia*, BenBella Books.

Morgan, D. L. (1988), *Focus Groups as Qualitative Research*, SAGE.

Mosteller, F. and Wallace, D. L. (1963), 'Inference in an authorship problem', *Journal of the American Statistical Association* **58**(302), 275–309.

Muralidharan, A. and Hearst, M. (2013), 'Supporting exploratory text analysis in literature study', **28**(2), 283–295.

NaNoGenMo (2016). Online at https://github.com/NaNoGenMo/2016 (Accessed 25th October, 2016).

Nelson, K. (n.d.), 'Timeline of literary theory'. Online at https://uk.pinterest.com/pin/553520610430070021/ (Accessed 19th April, 2016).

Neuman, W. L. (2013), *Social Research Methods: Qualitative and Quantitative Approaches*, Pearson Education.

Nield, D. (2016), 'A novel written by AI passes the first round in a Japanese literary competition'. Online at http://www.sciencealert.com/a-novel-written-by-ai-passes-the-first-round-in-a-japanese-literary-competition (Accessed 25th October, 2016).

Peng, R. and Hengartner, N. (2002), 'Quantitative analysis of literary styles', *The American Statistician* **56**(3), 175–185.

Plaisant, C., Rose, J., Yu, B., Auvil, L., Kirschenbaum, M. G., Smith, N., Clement, T. and Lord, G. (2006), Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces, *in* 'Proceedings of the Sixth ACM/IEEE-CS Joint Conference on Digital Libraries', ACM/IEEE-CS, pp. 141–150.

Powell, R. and Single, H. M. (1996), 'Focus groups', *International Journal of Quality in Health Care* **8**(5).

Ramezani, R., Sheydaei, N. and Kahani, M. (2013), Evaluating the effects of textual features on authorship attribution accuracy, *in* 'Third International Conference on Computer and Knowledge Engineering (ICCKE 2013)', IEEE, pp. 108–113.

Ransom, J. C. (1937), 'Criticism, inc.', *Virginia Quarterly Review* **13**.

Ray, R. H. (1994), 'Shakespeare's Sonnet 18', *The Explicator* p. 10.

Richards, L. (2009), *Handling Qualitative Data: A Practical Guide*, 2nd edn, Sage.

Roque, A. (2012), Towards a computational approach to literary text analysis, *in* 'Workshop on Computational Linguistics for Literature', Association for Computational Linguistics, pp. 97–104.

Ross, S. (2014), 'In praise of overstating the case: A review of Franco Moretti, *Distant Reading* (London: Verso, 2013)', *Digital Humanities Quarterly* **8**(1).

Rudman, J. (2012), 'The twelve disputed federalist papers: A case for collaboration'.

Sample, I. and Hern, A. (2014), 'Scientists dispute whether computer 'Eugene Goostman' passed Turing test', *The Guardian Online* . Online at http://www.theguardian.com/technology/2014/jun/09/scientists-disagree-over-whether-turing-test-has-been-passed (Accessed 29th August, 2014).

Sarndal, C.-E. (1967), 'On deciding cases of disputed authorship', *Applied Statistics* (16), 251–268.

Savoy, J. (2012), 'Authorship attribution based on specific vocabulary', **30**(2), 1–30.

Savoy, J. (2013), 'The federalist papers revisited: A collaborative attribution scheme', *Proceedings of the ASIST Annual Meeting* **50**(1), 1–8.

Sebastiani, F. (2002), 'Machine learning in automated text categorization', *ACM Computing Surveys* **34**, 1–47.

Shakespeare, W. (1968), Sonnet 18, *in* P. Alexander, ed., 'William Shakespeare: The Complete Works', Collins, p. 1311.

Simpson, P. (2004), *Stylistics: a Resource Book for Students*, Routledge.

Sofaer, S. (1999), 'Qualitative methods: What are they and why use them?', *Health Services Research* (34(5 Pt. 2)), 1101–1118.

Stamatatos, E. (2009), 'A survey of modern authorship attribution methods', **60**, 538–556.

Steadman, J. M. (1976), The idea of satan as the hero of "paradise lost", *in* 'Proceedings of the American Philosophical Society, Symposium on John Milton', Vol. 120, American Philosophical Society, pp. 253–294.

Steele, R. (1709), The Tatler, August 1709, *in* A. Partington, ed., 'The Oxford Dictionary of Quotations', revised 4th, 1998 edn, The Softback Preview, p. 662.

Stuart, L. M., Tazhibayeva, S., Wagoner, A. R. and Taylor, J. M. (2013*a*), On identifying authors with style, *in* 'International Conference on Systems, Man and Cybernetics', IEEE, pp. 3048–3053.

Stuart, L. M., Tazhibayeva, S., Wagoner, A. R. and Taylor, J. M. (2013*b*), Style features for authors in two languages, *in* 'International Conferences on Web Intellienge (WI) and Intelligent Agent Technology (IAT)', IEEE, pp. 459–464.

Stubbs, M. (2005), 'Conrad in the computer: Examples of quantitative stylistic methods', *Language and Literature* **14**(1), 5–24.

Thomasson, A. (2004), Fictional characters as abstract artefacts, *in* E. John and D. McIver Lopes, eds, 'Philosophy of Literature', Blackwell Publishing Ltd, p. 152.

Torres, D. F. M. (2002), Entropy text analyzer, *in* '2nd Portuguese National ACM Programming Contest'.

Tyson, L. (1999), *Critical Theory Today: a User-friendly Guide*, Garland Publishers.

U, A. and Thampi, S. M. (2015), Hallmarking author style from short texts by multi-classifier using enhanced feature set, *in* 'Proceedings of the Third International Symposium on Women in Computing and Informatics', pp. 284–289.

Wales, K. (1990), *A Dictionary of Stylistics*, Longman.

Zhang, J., Zeng, G. and Zhang, J. (2011), Quantitative evaluation of writing styles based on text analysis: Methods and case study, *in* '2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference', Vol. 1, IEEE, pp. 181–185.

Zhao, Y. and Zobel, J. (2007), Searching with style: Authorship attribution in classic literature, *in* 'Proceedings of the Thirtieth Australasian Computer Science Conference (ACSC2007)', Vol. 62, CRIPT, pp. 59–68.

DECLARATION

I, Tess Crosbie, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

A Computer Assisted Analysis of Literary Text: From Feature Analysis to Judgements of Literary Merit

I confirm that:
1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have cited the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Either none of this work has been published before submission, or parts of this work have been published as indicated on pages 28, 33 and 61.

Name of candidate: Tess Crosbie

Signature: Date: