Introducing the Corpus of the Canon of Western Literature : A corpus for culturomics and stylistics

Article *in* Language and Literature · December 2017

DOI: 10.1177/0963947017718996

Clarence Green
The University of Hong Kong
60 PUBLICATIONS **272** CITATIONS

GREEN, C. (2017). INTRODUCING THE CORPUS OF THE CANON OF WESTERN LITERATURE: A CORPUS FOR CULTUROMICS AND STYLISTICS. LANGUAGE AND LITERATURE, 26(4), 282-299.DOI:

https://doi.org/10.1177/0963947017718996

Abstract

This paper introduces the Corpus of the Canon of Western Literature (Version 1.0), accompanied by a demonstration of its potential uses. The Canon of Western Literature has been an important construct in the study of literature, long standing and long contested. It has been argued to represent many of the greatest works produced in the history of Western literature. This corpus operationalises the Western Canon based on Bloom (1994). The paper describes the development of the corpus, its organization and source material. Corpus procedures are applied to the corpus, such as word frequency analysis, lemmatization and keyness to demonstrate its potential uses in Culturomics and Corpus Stylistics, two interdisciplinary fields between the traditional and digital humanities, and the linguistic and literary approaches to literature. Culturomics is the study of culture and social psychology via the investigation of corpora of literature as cultural artifacts, while Corpus Stylistics is the application of corpus linguistics to traditional literary scholarship. The corpus introduced in this paper is open source and freely available.

Introducing the Corpus of the Canon of Western Literature: A corpus for Culturomics and Stylistics

1 Introduction

A relatively recent paper in *Science*, introducing the Google Books corpus with approximately 4% of books ever published, termed a new field of study: Culturomics (Michel et al., 2011). As originally framed, Culturomics was the use of the Google Books corpus to investigate the culture and social psychology of different times and places, with the corpus considered as a collection of cultural artifacts. While Culturomics is a new term, widely cited (Acerbi et al., 2013; Greenfield, 2013; Pechenick et al., 2015), using corpora for cultural studies is something corpus linguists have been doing for some time (e.g. Baker, 2003). Parallel to the rise of Culturomics has been the related field of Corpus Stylistics. Corpus Stylistics is the study of literary style via computational tools applied to machine readable literary works. It combines the science of linguistics with literary studies and like Culturomics is one of the growing interdisciplinary fields between the traditional and digital humanities.

This paper introduces the *Corpus of the Canon of Western Literature (Version 1.0),* with a demonstration of its potential in Culturomics and Corpus Stylistics. The Canon of Western Literature has been an important construct in the study of literature, long standing and long contested (Beach et al., 2016; Guillory, 2013). Speaking broadly, traditional-minded literature scholars have held the works of the Canon to be the greatest literature in the history of the West (Adler & Weismann, 2000; Bloom, 1994). By ‘greatest’, they tend to mean that such literature exhibits qualities such as aesthetic beauty, profound ideas, themes, notable characters and language, and impressive artistic skill. Canonical works are also those that have influenced other literature, e.g. exhibit intertextuality, and impacted culture, e.g. Aristotle’s *Politics* and Christendom. The *Corpus of the Canon of Western Literature* (henceforth CCWL) is an attempt to operationalise the construct of the Western Canon as defined by Bloom (1994). The paper first describes the development and organization of the CCWL. Next, to demonstrate its applications to Culturomics and Stylistics, some standard corpus procedures are reported, such as lemmatization, keyness, standardised type-token ratios (a measure of vocabulary range), as well as word and sentence length estimates across genres, authors and texts.

2 Corpus linguistics, Culturomics and Stylistics

Culturomics, as introduced by Michel et al. (2011), argued that the 5,195,769 texts in the Google Books corpus opened a new field of study in the digital humanities: the tracking of cultural trends and social psychology through linguistic artifacts in big data. In their introductory paper, they demonstrate how the relative frequencies of n-grams (words and phrases) map onto cultural phenomena. For example, the names for inventions in their corpus show that from the 1800’s onward the cultural adoption of technology has become more rapid. The frequency of reference to an invention first mentioned in the early 1800’s peaked around 66 years later in the corpus, yet by the 1900’s peak frequency occurred within 27. Other demonstrations of Culturomics in their paper include the tracking of censorship, evidenced by declining mentions of Jewish artists during Nazi Germany, the spread of scientific concepts throughout modernity such as *evolution*, and political concepts such as *feminism*, which has been taken up more rapidly in English books than French. Given the limitations of the Google Books corpus, e.g. prolific but unread authors affect frequency but not culture (Pechenick et al., 2015), Culturomics has expanded to other corpora. Samothrakis and Fasli (2015), for example, built a corpus from the digital repository Project Gutenberg consisting of 3403 public domain literary texts. They found that the frequency and dispersion of words associated with lexical domains such as *anger*, *fear*, *joy*, *surprise*, help predict publication periods of texts as these words tap into the changing cultural milieus of different historical periods (see also Hughes et al., 2012).

Corpus Stylistics is concerned with how the literary style of an author, text or genre is reflected in language, yet like Culturomics it is also interested in broader issues of how literature reflects culture, how ideas and themes pattern in texts, and how literature creates psychological effects in readers and characters (McIntyre, 2015). Even though corpus linguistics is advancing toward ever increasing complex quantitative research designs, the basic tool-kit of the field has provided much insight into literature. Stubbs (2005: 14), for example, shows that the application of what he calls “very simple frequency stuff” such as word lists and collocations capture important themes and style markers of Conrad’s *Heart of Darkness*. Amongst the most frequent words are *seem*, *like* and *looked*, as well as *something*, *somebody*, *sometimes*, *somewhere*, *somehow*, which Stubbs (2005) argues reflect the vagueness and sense of the inscrutable that has long been noted as a stylistic marker of Conrad’s novella (Leavis, 2011 [1948]).

Mahlberg and McIntyre (2011: 216) view Corpus Stylistics as “an approach that can link in with the concerns in literary stylistics and criticism”, rather than as field of study that competes with traditional literary studies (see also McIntyre, 2015). They demonstrate this in a corpus stylistic study of Fleming’s *Casino Royale* where, similar to Stubbs (2005), frequency information functions as evidence for arguments about theme, style and characterization. Beside raw frequency, they employ corpus linguistics procedures such as lemmatization and keyword analysis, which identify lexis associated with core themes (e.g. cards, casinos, spies), characters (Bond, Le Chiffre, Vesper) and the male viewpoint (e.g. the subjective pronoun he). Mahlberg and McIntyre (2011: 221) report that a key semantic domain in Fleming’s work is physicality, since there is high frequency of lemmas associated with the body. Further, the representation of the body is constructed differently according to gender. A collocational analysis of the n-gram *his body* (i.e. Bond's) compared to the central female character Vesper, reveals Bond's collocates emphasize his ability to separate his physical self from his mental and emotional self, while Vesper's body is presented either sexually, collocating with words such as *morals, bed, sheet, sensual, conquest,* or from Bond's point of view as *unemotional, cold, arrogant, remote*.

Not only do the above studies indicate the wide range of research applications for literary corpora once they are built, but also how the basic toolkit of corpus linguistics can produce insights into literature, culture and social psychology (Greenfield, 2013). The following sections describe a newly built literary corpus and, by way of introduction, apply some of the above procedures in the context of Culturomics and Corpus Stylistics.

3 The Canon of Western Literature

Unlike the corpora in the previous section, the corpus introduced in this paper represents a specific literary and cultural construct, i.e. it is a specialised corpus, and this construct is the Canon of Western Literature (Bloom, 1994). The Canon of Western Literature has been an influential idea in literary studies. It has been argued to consist of the core literary tradition of the west. Canonical literature has been defined as texts with great aesthetic beauty and important influence in shaping other literature, as well as western thought and culture in general. Leavis (2011 [1948]) argued it represents a 'Great Tradition', in which previous great works shape the style and form of the literature that follows. Adler and Weismann (2000) use a similar phrase: the 'Great Conversation'. They conceive of the Canon as an intertextual conversation between authors across centuries, where ideas, styles, characters, philosophies, and science are discussed, refined, rejected, and renewed. The Canon has an overall coherence, they believe, as literature that does not participate in this 'Great Conversation', either explicitly or implicitly via literary criticism, falls outside canonical literature. Bloom (1994), author of the influential *The Western Canon:*

The Books and School of the Ages, presents a similar definition, though he largely excludes scientific treatises as he argues that aesthetic beauty is a key inclusion criterion. Bloom (1994) is one of the staunchest current defenders of the Western Canon, and also offers one of the most cited taxonomies of canonical authors and texts.

The challenges and critiques of the Canon are well-known, part of the general culture wars of recent academia (Gorak, 2013), and include that the Canon overwhelmingly represents white male authors, characters and viewpoints, suppresses the voices of women, the cultures of minorities, the spiritual beliefs of those not consistent with an era's reigning (and often brutally enforced) theology etc. The canonicity of any text is debatable, and overrepresented is literature related to the Greco-Roman tradition, which partly reflects 19th century models of Liberal Arts education (Towheed & Owens, 2011). Further, there is a debate over who gets to choose the works in the Canon, as scholars who have proposed lists of canonical literature tend to be much like the authors they include, i.e. white, male, English speakers of European heritage. The current paper's introduction of a corpus of the Canon of Western Literature is not meant as a defense of the construct itself. Rather, the corpus is presented as an object of study for the empirical investigation of what has been held up to be literature of great importance to western culture (cf. Google Books).

4 The development and structure of the Corpus of the Canon of Western Literature

The corpus introduced here operationalises the construct of the Canon of Western Literature based on Bloom's (1994) description of the canon, chosen because he is a major contemporary literary scholar who specialises in canonical literature, because his work is highly cited and influential, and because his list can be operationalised since he offers an explicit taxonomy of thousands of texts and authors in Appendix A of his book. The structure of Bloom's (1994) canon has guided the structure of the CCWL. He organizes canonical literature into four chronological ages: 1. The Theocratic Age (2000 BCE to 1321 CE), 2. The

Aristocratic Age (1321 CE to 1832 CE), 3. The Democratic Age (1832 CE to 1900 CE), 4. The Chaotic Age (20th Century). The names of the ages, Bloom (1994) suggests, reflect important cultural or stylistic underpinnings of the literature in each era such as a heightened religiosity (the first literary age) or a lack of cultural coherence (the final age). He subdivides the four ages into different cultures/societies. For example, nested within the Theocratic Age are the Ancient Greeks and the Romans, while nested in the Democratic Age are works from Great Britain and the United States.

The majority of texts in the Canon are from the British Isles or the United States and originally written in English. Indeed, one might suggest that Bloom's (1994) Western Canon is more specifically a Western Canon of the English speaking peoples. Hundreds of literary works not originally in English from Homer to Proust are listed by Bloom (1994), and these have been included in the CCWL in translation. While Bloom (1994) might hold that the works should be read in the original languages (though this is not clear), others such as Adler and Weisman (2000) argue that translations still represent the 'Great Conversation', and so it was decided they have a place in the corpus. Of course, the style of the translator and era of translation influence these texts, but the CCWL has been designed for researchers to ignore translated texts if desired.

The development of the CCWL proceeded as follows. Every text listed in Bloom's (1994) Appendix A was searched for in Project Gutenberg (https://www.gutenberg.org/), a digital repository of public domain literature. Project Gutenberg texts are not copyrighted and available freely for research. Each text contains a license statement, and scholars who use this corpus should read the license, as countries vary on copyright. The CCWL is freely available under the standard licensing of Project Gutenberg upon request from the authors[1]. The corpus was tagged and cleaned to minimize non-target text. License statements were put behind the XML tags <License>; footnotes, endnotes, indexes, introductions, appendices and contents pages were tagged <notes>. Texts were also tagged for the genres: <fiction>, <non-fiction>, <play>, <poetry>, <prose>, <scripture>, <mixed genres>. When possible, regex scripts were written to remove noise such as line break characters, page numbers etc. Plays presented a particular challenge as Gutenberg editions standardly have a period immediately after a line initial speaking character's name. This skews estimates of mean sentence length, and such repetition affects type-token ratios. To minimize this, all plays (and works such as Plato's *Dialogues*), had the speaker's names put behind <character> tags. All files were Utf-8 encoded, which provides a standard and compact formatting for all characters in text files.

Text files were kept in-tact as much as possible; that is, sometimes a single volume in Project Gutenberg contained multiple target texts from an author listed in Bloom (1994). However, when a target text was only available in a collected volume, non-target texts within that file were removed. Files in the corpus were named according to Bloom's Appendix (i.e. author/title), rather than given codes. This was done in an interdisciplinary sprit, in the hopes that intuitive file names may make the corpus more accessible to non-corpus linguists such as literary scholars. When there were multiple versions of the same text available, it was decided to use the edition that had been most downloaded from Project Gutenberg. This is arbitrary, but it is possible the most downloaded version is more central to the Canon than less read editions. Bloom (1994) operates similarly, including only the King James version of the Bible. A supplementary part-of-speech tagged version of the corpus was also developed, with tagging by TagAnt (Anthony, 2015). Checks of random samples suggested that tag accuracy varies, with performance best on prose written after 1800. For example, within Chaucer's *Canterbury Tales* the tagger handled some archaic style with 100% accuracy, e.g. *Thus _RB can _MD Fortune _NP her _PP wheel _NN govern _VV*, while it was inaccurate with other sequences, e.g. *He _PP which _WDT that _DT misconceiveth _NN oft _RB misdeemeth _VVZ.* An examination the 100 most frequent NP tags in Greco-Roman sub-corpus (approximately 1.3 million words) indicated and error rate of around 6%. Given time and resource constraints in this phase of the project, machine tagging has not been checked by hand by independent raters and errors corrected.

The final corpus contains 805 individual files (many containing multiple works) in a flat structure and, excluding non-target text, approximately 73 million words, which compares favorably to large corpora such as the BNC at 100 million. Table 1 shows the organization of the corpus and the sample sizes for each literary age, society and culture listed in Bloom (1994).

Table 1. The Corpus of the Canon of Western Literature

A. The Theocratic Age	**Word Count**	**B. The Aristocratic Age**	**Word Count**
(2000 BCE to 1321 CE)		**(1321 to 1832)**	
A1. Ancient Near East	1 183 650	B1. Italy	2 062 754
A2. Ancient India	620 728	B2. Portugal	74 835
A3. Ancient Greeks	1 627 097	B3. Spain	720 886
A4. Hellenistic Greeks	951 025	B4. England and Scotland	14 512 256
A5. The Romans	808 185	B5. France	2 347 696
A6. The Middle Ages	1 307 171	B6. Germany	628 670
Total:	**6 702 973**	**Total:**	**20 347 097**
C. The Democratic Age	**Word Count**	**D. The Chaotic Age**	**Word Count**
(1832 to 1900)		**(20th Century)**	
C1. Italy	279 505	D1. Italy	64 119
C3. France	3 054 359	D4. Portugal	6 953
C4. Scandinavia	191 032	D5. France	331 477
C5. Great Britain	19 321 021	D6. Great Britain and Ireland	6 983 223
C6. Germany	1 139 020	D7. Germany	479 747
C7. Russia	3 976 265	D8. Russia	346 211
C8. United States	7 734 357	D9. Scandinavia	534 970
Total:	**35 695 559**	D15. Yiddish	96 361
		D23. Australia and New Zealand	212 723
		D24. The United States	1 889 639
		Total:	**9 945 423**
CCWL Word Count:	**72 691 052**		

Table 1 indicates significant word count differences exist in the representation of times and places, but this reflects the canon as described by Bloom (1994). Approximately 25% of the corpus is British literature from the Democratic Age (1832-1900 CE). The sample sizes for other periods and cultures/societies are quite good, nonetheless, with around half of the nested subcorpora around or greater than one million words. Corpora of a million words have been effectively used since the 1960's (e.g. Brown) until the current era (e.g. ICE). It is worth noting that Bloom (1994) is not strictly chronological in categorization, but considers also literary movement. For example, the romantic poets are nested in the

Democratic Age, as they were a reaction to neoclassicism and a style he considers of the Aristocratic Age. Not every text listed in Bloom (1994) was obtainable in Project Gutenberg. Literature from the Chaotic Age has the least coverage as many of the texts are still under copyright; yet, as Table 1 shows, the age nevertheless has sizable representation. Gaps in consecutive numbering (e.g. D2-3) indicate no available texts. The exact coverage of the Western Canon as described by Bloom (1994) can only be approximated. This is for two reasons. One is that Bloom is at times vague about the texts that are canonical; for example, while the specific titles of Charles Dickens are listed, for other authors he simply notes Selected Poems or Short Novels. The second issue relating to coverage is that where Bloom specifies the complete works of an author as canonical, Project Gutenberg did not always have all their work. If we estimate representation by authors, from the Theocratic Age, the CCWL represents 48 of 63 (76%) canonical authors mentioned by Bloom (1994); from the Aristocratic Age, 88 of 139 (63%); from the Democratic Age, 125 of 159 (79%); and finally from the Chaotic Age, where Bloom (1994) lists a total of 506 authors, only 58 (12%) are represented. Representation bias is thus toward literature before 1900. Bloom (1994: 548) leaves open whether Chaotic Age texts are technically canon, as he suggests they must also withstand the test of time: "I am not as confident about this list… Not all of the works here can prove to be canonical”.

5 Applications to Culturomics

This section applies a few standard corpus procedures to the CCWL, and illustrates how the corpus can be used for Culturomics. Simple frequency has its interest, but to hone in on the lexis of literature lemmatization and keyness procedures often provide more insights (McIntyre, 2015; Stubbs, 2005). Keyness highlights lexis in a corpus that stand out statistically in terms of relative frequency and dispersion compared to a larger reference corpus. Reported in Table 2 are the 20 highest ranked keywords in the CCWL, computed against the BNC. The BNC is a far from perfect reference corpus (indeed no currently available corpus would be) as it is a contemporary, mixed-genre corpus of speech and writing. Nevertheless, it is a well-known British corpus of a size larger than the CCWL, and the comparison for the generation of keywords, while problematic, is not meaningless. Consider that when a school student encounters Shakespeare, the lexis that stands out is that which is distinct from their everyday experience of English: e.g. *Shall* I compare *thee* to a summer's day?

Table 2. Highest ranked keywords in the CWCL

N	Keyword	Freq.	N	Keyword	Freq.
1	My	404811	11	Shall	104098
2	His	728170	12	And	2497534
3	I	1047223	13	Thee	49531
4	He	878865	14	Man	140533
5	Him	364066	15	Not	513961
6	Me	320533	16	Am	80451
7	Thou	77637	17	Ye	33394
8	Her	474415	18	Himself	81608
9	Thy	64120	19	All	347217
10	Upon	114796	20	Nor	53878

Table 2 shows that pronouns stand out as keywords in the CCWL. This likely reflects a property of literature that Stockwell and Mahlberg (2015) call the textual trace of characterization, i.e. characters display pronominal chains reflecting their participation in a narrative. Note that masculine pronouns are more key than female ones. In the top twenty keywords, five male referents occur, four being pronominal, and one superordinate *man*. There is only one female referent, the pronoun *her*, which is not subjective case; indeed, nominative *she* is only the 29th keyword of the CCWL compared to *he* ranked 4th. The subject of a clause is typically the agent, one who does, acts, perceives, thinks or senses (Givon, 1993), while the predicate is the part of the clause where propositions prototypically package those who are recipients, instruments, acted upon, or thought about (Halliday, 2003). Thus, Table 2 suggests that gender representation in canonical literature is qualitatively and quantitatively distinct. Of course, this observation is not necessarily true *only* of canonical literature, but it demonstrates nonetheless how the

CCWL can be used to bolster with supporting empirical evidence long-standing criticisms of the canon, such as that it is dominated by male characters, experience and viewpoints.

As discussed, Mahlberg and McIntyre (2011) effectively used lemmatization to highlight lexis associated with key themes, characters and semantic domains in their study of *Casino Royale*. A function word stoplist and the Someya (1998) list of 4,762 lemmas were therefore applied to the CCWL using Wordsmith v.7 (Scott, 2016). The Someya (1998) list, derived from modern corpora, lacks coverage of archaisms like in the Chaucer example above, but this seems a relatively minor limitation. Table 3 ranks the 25 most frequent lemmas in the CCWL.

Table 3. Most frequent lemmas in the CCWL

N	Lemma	Freq.	N	Lemma	Freq.
1	Man	215873	14	Hear	69375
2	Time	128363	15	Place	65498
3	Great	110725	16	Sir	64706
4	Day	105239	17	Speak	64465
5	Good	103873	18	God	68140
6	Hand	93038	19	Word	64163
7	Thing	92791	20	Feel	62563
8	Love	87536	21	House	60320
9	Life	85193	22	Call	58739
10	Find	84508	23	Lie	58308
11	Long	83144	24	Work	57283
12	Eye	73671	25	Heart	55041
13	Leave	71508			

A few interesting observations can be drawn from Table 3. The first is that canonical literature exhibits the Pollyanna Effect (Ingram et al., 2016). The Pollyanna Effect proposes that although human languages tend to have a wider range of words for negative experience, those for positive experience are much more frequent. In the CCWL, the most frequent lemmas reflect recurrent themes of *love* and *life*, things that are *great* and *good*, and discussions of the *heart* and *God*. This positivity bias is more marked than in a general corpus (Leech et al., 2001). For example, *good* occurs 1276 times per million words in the BNC, compared to 1523 p/m in the CCWL; *great* occurs 635 p/m words in the BNC, and 1523 p/m in the CCWL; *heart* 152 p/m in the BNC, and 757 p/m in the CCWL; and finally *love* occurs 150 times p/m in the BNC but 1204 times p/m words in canonical literature. This suggests that even though canonical literature from Homer to Hemmingway addresses death, war, heartache and tragedy, the overall cultural preoccupations of the western canon over history have been largely positive.

The list also shows many lemmas for body parts. Some of these lemmas are physical such as *hand*, *heart, eye*, and others are for bodily sensory experience such as *hear*, *speak*, *feel*. The reason why body part language plays such an important role is perhaps the cognitive poetic one noted by Stockwell and Mahlberg (2015: 132); namely, that effective characterization for mind-modelling requires more description of the body than non-literary language since the author needs to communicate what characters look like, how they move, what they are doing, in order to help readers create a cognitive representation. Table 3 reflects the (not surprising) fact that human experience is a major focus of canonical literature, and that this experience is embodied.

5.1 The decline in influence of the Greco-Romans and the Theocratic Age

Michel et al. (2011) argue that Culturomics can track the rise and fall of the cultural preoccupations of those who produced the texts in a corpus. This section explores two cultural preoccupations of canonical literature, namely religion and the Greco-Romans. Firstly, let us consider religion as a literary theme over time. As was reported in Table 3, *God* is the 18th most frequent lemma in the CCWL, indicating that religion is a canonical theme. Yet, the focus on religion wanes over time. Lemma lists computed for each age indicate that in the Theocratic Age, religion is a dominant topic, with *God* as the 2nd most frequent lemma, *lord* 3rd, and *soul* 35th. The top four keywords, computed against the rest of the corpus, are *God*, *son*, *lord* and *king* respectively. Bloom's (1994) intuitive naming of a Theocratic Age of canonical literature seems apt. However, in the Aristocratic Age, *God* is only the 19th most frequent lemma, *lord* 16th and *soul* 82nd. By the Democratic Age, *God* has slipped to 50th, *lord* 72nd, *soul* 107th; and by the Chaotic Age, *God* is 65th, *lord* 350th and *soul* 107th. While the influence and themes of the Theocratic Age decline, the rise of humanism appears to take its place. For example, even though *man* is the most frequent lemma in the Theocratic Age and all others, it is ranked 7 places (i.e. 8th) below *God* as a keyword for the era; however, by the Democratic Age, *God* is no longer within even the top 500 keywords. Further, in the Democratic and Chaotic Ages, the top 20 keywords and lemmas contain the following words which Theocratic Age literature does not: *eye*, *face*, *stand*, *sit*, *cry*, *feel*, *walk*, *laugh*- all related to human (bodily) experience. The data suggest a shift of focus in canonical literature across time from the spiritual to the representation of human experience. Arguably, the decline in religion evidenced in canonical literature is a reflection of the decline in its historical centrality to western culture (i.e. a Culturomic trend).

Let us consider the intertextual question of the influence of classical literature on the Western Canon. A long standing claim has been that the influence of the Greco-Romans has been unparalleled in terms of style, themes, philosophy, characters etc. (Highet, 2015 [1953]: 19). To compute literary connections to the classics, the Greco-Roman subcorpora of the CCWL were queried, approximately 3386307 words of texts nested within *A3: The Ancient Greeks*, *A4: The Hellenistic Greeks*, and *A5: The Romans*. To create a metric for tracking classical reference in subsequent literary eras, the 50 highest ranked keywords (computed against remaining eras) and the 100 most frequent proper nouns were extracted (from the POS tagged version, with tag accuracy checked by hand) and used as batch searches in Wordsmith 7 (Scott, 2016). The cutoff ranks are arbitrary (Mahlberg & McIntyre 2011), but the procedure produced a list of characters, places and historical figures central to Greco-Roman literature, as reflected in the sample in Table 4.

Table 4. Highest ranked keywords and proper nouns in Greco-Roman literature

N	Keyword	N	Keyword	Rank	Proper Noun	Rank	Proper Noun
1	Athenians	9	Ulysses	1	Athenians	9	Lacedaemonians
2	Socrates	10	Persians	2	Socrates	10	Troy
3	Hellenes	11	Army	3	Plato	11	Cato
4	War	12	Zeus	4	Athens	12	Greece
5	Lacedaemonians	13	Pompey	5	Caesar	13	Achilles

6	City	14	Caesar	6	Rome	14	Hector
7	Ships	15	Citizens	7	Pompey	15	Ulysses
8	Athens			8	Jove		

* Function words, *God*, *King*, and character names in plays were excluded

The keywords and proper nouns in Table 4 capture important classical characters (Achilles), places (Rome), gods (Zues), people (Socrates), as well as characteristics of the Greco-Romans such as the emphasis on the *city*, *ships*, *citizens*, and the valour of the *army* and *war*. Reported in Table 5, normalised per million words, are the keywords and proper nouns from Greco-Roman literature tracked across the literary ages.

Table 5. Frequencies (p/m) of Greco-Roman lexis across time in canonical literature

Literary Era	**Proper Nouns**	**Keywords**
Greco-Roman Age	10796	15598
Middle Ages	1464	6985
Aristocratic Age	1807	3397
Democratic age	605	2401
Chaotic age	349	2176

Table 5 suggests a general decline of the literary influence of the classics, or at least, with their literary preoccupations. Greco-Roman keywords steadily decline till the modern era, as do literary references to Greco-Roman characters, people and places. However, note how references to proper nouns from the classical period spike in the literature of the Aristocratic Age. This age, which in Bloom's (1994) estimation spans 1321 to 1832 A.D., represents the Late Middle Ages, Renaissance and the reestablishment of democracy. One of the defining characters of this period of western history was looking back to the classical world (Pitts & Versluys, 2014).

6 Stylistics with the CCWL

The above has largely used the CCWL for Culturomics, so let us conclude this paper with some uses of the corpus for stylistics. This section reports on: 1. The authors/texts in canonical literature with the most sesquipedalian style, which is a long word that refers to the overuse of long words; 2. Those with a preference for longer/shorter sentences; 3. Those with larger/smaller vocabulary ranges (measured by standardized type-token ratios). These may not be profound questions, but they are reflections of style that the CCWL can help us put on record. Since genre affects style, e.g. Tolstoy's sentence length in *Anna Karenina* is likely not comparable to his plays, the following reports genre estimates separately for texts tagged <poetry>, <prose>, <play>. Estimates have been computed by Wordsmith (Scott, 2016), excluding notes, license statements, and character names in plays. Table 6 reports the longest and shortest mean word lengths by author/text across genres.

Table 6. Mean word lengths in the CCWL

	Prose		**Play**		**Poetry**	
1	Edward Gibbon		Tommaso Campanella		Luis de Camoëns	
	Fall of the Roman Empire[2]	4.84	*The City of the Sun*	4.45	*The Lusiads*	4.63
2	Friedrich Nietzsche		Robert Garnier		Oliver Goldsmith	
	The Birth of Tragedy	4.79	*Mark Antony*	4.42	*The Deserted Village*	4.60
3	George Bernard Shaw		Christopher Marlowe		Aleksandr Pushkin	
	Essays (Vol 4)	4.78	*Tamburlaine the Great (1)*	4.40	*Eugene Onegin*	4.57
4	Percy Bysshe Shelley		Christopher Marlowe		Robert Graves	
	A Defence of Poetry	4.77	*Tamburlaine the Great (2)*	4.36	*Collected Poems (2)*	4.55
5	Samuel Taylor Coleridge		Jean Racine		Homer	
	Prose (Vol 5)	4.75	*Athaliah*	4.35	*Iliad*	4.51
6	Edgar Allan Poe		Goethe		Richard Crashaw	
	Eureka	4.74	*Egmont*	4.32	*Poems (Vol 1)*	4.51
7	Friedrich Nietzsche		Thomas Kyd		John Milton	
	Beyond Good and Evil	4.73	*The Spanish Tragedy*	4.30	*Paradise Lost*	4.50
8	George Bernard Shaw		Christopher Marlowe		Virgil	
	Essays (Vol 1)	4.78	*History of Dr Faustus*	4.29	*Georgics*	4.48
9	John Stuart Mill		Richard Wagner		Unkown	
	On Liberty	4.72	*Ring of the Nibelung (2)*	4.28	*Beowulf*	4.48
10	Thomas Carlyle		Schiller		William Cowper	
	Sartor Resartus	4.69	*Mary Stuart*	4.28	*Poetical Works*	4.46
	CCWL Mean	**4.33**		**4.10**		**4.27**
1	Mark Twain		John Millington Synge		Alfred Tennyson	
	Huckleberry Finn	3.80	*Collected Plays (3)*	3.83	*Lady Claire*	3.77
2	Charles Chesnutt		Leo Tolstoy		Geoffrey Chaucer	
	The Short Fiction (1)	3.80	*The Power of Darkness*	3.86	*Troilus and Criseyde*	3.81
3	Unknown		Oscar Wilde		Edwin A. Robinson	
	The Apocrypha (1)	3.85	*Plays (1)*	3.87	*Selected Poems (2)*	3.90

4	George MacDonald *Back of the North Wind*	3.87	John Millington Synge *Collected Plays (2)*	3.88	Robert Frost *The Poetry (3)*	3.90
5	Knut Hamsun *Pan*	3.87	Aleksandr Ostrovsky *The Storm*	3.90	Robert Frost *The Poetry (2)*	3.92
6	Robert Louis Stevenson *Kidnapped*	3.88	Beaumont and Fletcher *Plays (2)*	3.92	Edwin A. Robinson *Selected Poems (1)*	3.97
7	Samuel Richardson *Pamela*	3.89	Henrik Ibsen *The Master Builder*	3.93	A.E. Houseman *Collected Poems (2)*	4.00
8	Grimm Brothers *Fairy Tales*	3.90	John Millington Synge *Collected Plays (6)*	3.94	Edwin A. Robinson *Selected Poems (3)*	4.00
9	William Morris *Well at the World's End*	3.91	John Millington Synge *Collected Plays (5)*	3.94	Robert Frost *The Poetry (1)*	4.01
10	Daniel Defoe *Moll Flanders*	3.92	August Strindberg *Miss Julie; The Father*	3.96	Wolfram Eschenbach *Parzival*	4.02

Table 6 indicates that mean word length varies across genres. Plays use shorter words on average compared to poetry or prose, likely a stylistic marker of direct speech which correlates with high frequency, shorter words (Greenbaum & Nelson, 1995). Gibbon’s *Decline and Fall of the Roman Empire* uses the longest words, on average, of any author, which perhaps reflects a conscious (or unconscious) Latinate prose style related to his subject matter. Nietzsche also favors long words, which may partly be the influence of translation from German (see also Goethe’s and Wagner’s plays), a language with less analytic word building processes than English (Wierzbicka, 1997). However, it also seems to be a style associated with philosophy since J.S. Mill and Carlyle also have some of the longest average word lengths in the canon. Table 6 reflects authorial style more specifically; for example, the different plays of Synge are recurrent in the list of shortest mean words lengths, as are volumes of poems by Frost and Robinson. *The Adventures of Huckleberry Finn* uses the shortest words in prose, a style likely reflecting a conscious attempt by Twain at authenticity in the representation of the thoughts/conversations of central characters who would have used shorter, high frequency words: i.e. Huck Finn is child, Jim is a slave deprived of education (Wood, 2012). The style associates with children’s literature more generally, as Stevenson, Morris and Grimm’s *Fairy Tales* also make the list of shortest mean word lengths. Further, there is also perhaps reflection of the preferred styles of different literary ages, as the majority of authors with a preference for short words across genres, are generally more modern rather than (neo) classical.

In corpus-stylistics, sentence length has been correlated with the style of a range of authors from the short declarative sentences of Hemmingway (Toolan, 2009) to the long verbose sentences of Joyce (O'Halloran, 2007). Table 7 reports the authors/texts in the CCWL with the longest and shortest average sentence lengths.

Table 7. Mean sentence lengths in the CCWL

	Prose		**Play**		**Poetry**	
1	Thomas More		Christopher Marlowe		Torquato Tasso	
	Utopia	59.84	*Tamburlaine the Great (2)*	26.27	*Jerusalem Delivered*	54.91
2	Madame de La Fayette		Tommaso Campanella		John Milton	
	The Princess of Cleves	58.27	*The City of the Sun*	25.95	*Paradise Lost*	48.49
3	Herodotus		Christopher Marlowe		Oliver Goldsmith	
	The Histories	55.72	*Tragedy of Dido*	25.35	*The Deserted Village*	47.23
4	Giorgio Vasari		Christopher Marlowe		William Morris	
	Lives of the Painters	53.84	*Tamburlaine the Great (1)*	23.74	*Poems*	45.82
5	Lucian		William Shakespeare		Edmund Spenser	
	Satires	53.72	*Plays and Poems (3)*	23.30	*The Faerie Queene*	45.23
6	Daniel Defoe		Pedro de la Barca		Geoffrey Chaucer	
	Robinson Crusoe	53.55	*Life is a Dream*	22.82	*The Canterbury Tales*	45.04
7	Apuleius		Robert Garnier		George Byron	
	The Golden Ass	53.38	*Mark Antony*	21.76	*Poems (2)*	43.84
8	Erasmus		John Millington Synge		Samuel Butler	
	In Praise of Folly	51.61	*Collected Plays (6)*	20.40	*Hudibras*	41.77
9	Miguel de Cervantes		Pierre Corneille		Lucretius	
	Don Quixote	50.90	*The Cid*	18.65	*The Way Things Are*	41.15
10	Aristotle		Richard Sheridan		Michael Drayton	
	Ethics	49.97	*School for Scoundrels*	18.09	*Poems*	40.58
	CCWL Mean	**21.44**		**13.59**		**25.01**
1	Gertrude Stein					
	The Geographical		Frank Wedekind		Aleksandr Pushkin	
	History of America	9.22	*Lulu Plays (1)*	6.93	*Boris Godunov*	11.19
2	Lawrence, D. H		Henrik Ibsen		Robert Frost	
	Sons and Lovers	10.37	*The Lady from the Sea*	7.35	*The Poetry (3)*	11.41
3	Anton Chekhov		Oscar Wilde		S.T. Coleridge	
	The Tales (10)	10.90	*Plays (5)*	7.59	*Poems (2)*	12.18
4	Katherine Mansfield		Henrik Ibsen		Unknown	
	The Short Stories (1)	10.91	*The Master Builder*	7.60	*The Epic of Gilgamesh*	12.66
5	Arthur Schnitzler		Henrik Ibsen		S.T. Coleridge	
	Stories (4)	11.11	*Hedda Gabler*	7.66	*Poems (3)*	12.77
6	James Joyce		Leo Tolstoy		Robert Frost	
	Ulysses	11.25	*The Power of Darkness*	8.11	*The Poetry (2)*	15.17
7	E.M. Forster		Nikolai Gogol	8.16	W. Carlos Williams	

	Howard's End	11.65	*The Inspector-General*		*Collected Poems*	16.23
8	David Lindsay		Henrik Ibsen		Unknown	
	A Voyage to Arcturus	11.95	*When We Dead Awaken*	8.34	*The Poem of the Cid*	16.25
9	Henry James		Oscar Wilde		Vachel Lindsay	
	The Awkward Age	11.98	*Plays (2)*	8.37	*Collected Poems (3)*	16.36
10	Katherine Mansfield		August Strindberg		Vachel Lindsay	
	The Short Stories (2)	12.06	*To Damascus*	8.62	*Collected Poems (2)*	16.39

In Table 7, again one can see both styles of authors and genres reflected in sentence length. Plays have a much shorter mean sentence length than prose, though not it seems in the era of Shakespeare and Marlowe where the style was not intended to represent actual speech. This is unlike modern playwrights who use the shortest sentences, an imitation of spoken utterances which tend to be shorter and lack the syntactic complexity of writing (Greenbaum & Nelson, 1995). Ibsen's style of realism, with its truncated utterances to produce melancholic effects, is reflected in the fact that he has multiple plays within the ten texts with the shortest mean sentence length in the corpus. Poetry has generally longer sentences than prose, which one suspects reflects that a unit of scansion is more often offset from other text lines by a comma, or (semi)colon as in Milton (Fish, 2001), rather than sentence punctuation. Table 7 also suggests that long sentences pattern with the Greco-Roman or Aristocratic Ages. As the previous section indicated, the two periods appear to be intertextually and culturally related. Note that *Ulysses* had one of the shortest sentence lengths in the CCWL, despite the having one of the longest sentences in the history of literature. The estimate here, however, accords with previous reported estimates (Borja, 2014), and the novel did have the second highest standard deviation in the corpus.

Scholars have often used the literary output of authors to estimate their vocabulary size, Shakespeare being one frequently studied case (Craig, 2011). A common procedure for the estimate is the type-token ratio, which calculates how many different types of words there are in a text (i.e. lemmas) relative to how many actual words there are in the text (i.e. tokens). If an author's work has higher number of types to the overall number of tokens, this indicates it contains a wider vocabulary range (Holmes, 1994). Since text length affects the type-token ratio (Baker, 2004), i.e. texts with more words will have more words that occur only once, Table 8 reports a standardised TTR based on averages per 1000 words for the authors/texts in the CCWL.

Table 8. Vocabulary range in the CCWL

	Prose		**Play**		**Poetry**	
1	James Joyce, *Ulysses*	50.91	Robert Garnier, *Mark Antony*	50.89	Richard Crashaw, *Poems (Vo1 1)*	57.76
2	Thomas Carlyle, *Sartor Resartus*	50.72	Goethe, *Faust*	50.48	Virgil, *Georgics*	56.28
3	Juvenal, *Satires*	50.52	Seneca, *Tragedies*	49.32	Aleksandr Pushkin, *Eugene Onegin*	56.03
4	Robert Burton, *Anatomy of Melancholy*	49.54	Jean Racine, *Phaedra*	48.12	John Milton, *Minor Poems*	55.48
5	Gérard de Nerval, *Sylvie*	49.26	John Marston, *The Malcontent*	47.22	Catullus, *Attis and Other Poems*	54.73
6	Aleksandr Pushkin, *Prose Tales*	49.20	Jean Racine, *Athaliah*	47.12	John Keats, *Poems (3)*	54.64
7	Norman Douglas, *South Wind*	48.93	John Webster, *The White Devil*	46.92	George Byron, *Poems (2)*	54.58
8	Thomas Nashe, *Unfortunate Traveller*	48.71	Richard Wagner, *Ring of the Nibelung (2)*	46.74	Victor Hugo, *Selected Poems*	54.34
9	Jean de La Fontaine, *Fables*	48.63	Marlowe, *Tamburlaine the Great (1)*	46.67	Emerson, *Poems*	54.03
10	Gustave Flaubert, *Sentimental Education*	48.38	Thomas Otway, *Venice Preserv'd*	46.44	Wilfred Owen, *Collected Poems*	53.80
	CCWL Mean	**42.47**		**42.50**		**48.29**
1	Unknown, *The Apocrypha (1)*	29.65	Oscar Wilde, *Plays (1)*	32.14	Unknown, *The Epic of Gilgamesh*	32.90
2	Thomas Malory, *Le Morte D'Arthur*	31.43	Synge, *Collected Plays (3)*	33.36	E.A. Robinson, *Selected Poems (2)*	37.24
3	Gertrude Stein, *Geog. History*	31.56	Ibsen, *The Lady from the Sea*	34.86	Dante, *The New Life*	38.12
4	*Egyptian Book of the Dead*	32.00	Ibsen, *The Master Builder*	35.10	E.A. Robinson, *Selected Poems (1)*	38.72
5	Unknown, *The Apocrapha (2)*	32.21	Synge, *Collected Plays (2)*	35.43	Tennyson, *Poems (1)*	38.80
6	Plato,		Synge,	35.88	Edward Lear,	39.40

	Dialogues (1)	32.89	*Collected Plays(6)*		*Complete Nonsense*	
7	Aristotle, *Ethics*	33.09	Tolstoy, *The Power of Darkness*	36.00	Chaucer, *Troilus and Criseyde*	39.58
8	Rudyard Kipling, *Stories (2)*	33.57	Synge, *Collected Plays (5)*	36.05	Homer, *Odyssey*	40.03
9	Grimm Brothers, *Fairy Tales*	34.14	Oscar Wilde, *Plays (6)*	36.08	Wolfram Eschenbach, *Parzival*	39.20
10	Hobbes, *Leviathan*	34.57	Oscar Wilde, *Plays (6)*	36.46	Unknown, *The Poem of the Cid*	40.60

While *Ulysses* has one of the shorter average sentence lengths in canonical literature, Table 8 indicates the novel has the highest STTR of any prose work in the corpus. The finding is consistent with previous stylistic work that has emphasized Joyce's lexical complexity (O'Halloran, 2007). Generally, poets seem to have the widest vocabulary range in the Canon. There are few reasons for this. One is that poetry relies more heavily than other literature on the artistic choices made in relation to vocabulary, so rather than frequent words that come to mind easily, poets select words that are less common. Further, a poem is usually short, and the demands of the form sacrifice function words. A collection of poems also might not deal with same characters, places and things, thus decreasing STTR. Lexical range appears to be an element of the style of Ibsen, Synge and Oscar Wilde, at least in his plays, while authors such as Pushkin have a high STTR regardless of the form they are working in. Children's literature and religious prose, which had shorter words and sentences, tends to have a higher rate of lexical repetition.

The previous data have indicated that there is variation style according to genre and author across the three metrics of word length, sentence length and vocabulary range. However, some authors, e.g. Defoe, Joyce, Coleridge, appear multiple times across the measures, suggesting there may be relationship across these elements of style. A Pearson's product moment was therefore computed for all texts in the CCWL, finding the following general correlations: word and sentence length (r=.36, p< .01), word length and STTR (r=.49, p<.01), STTR and sentence length (r= .09, p < .01). In other words, canonical literature with longer sentences has a moderate tendency to also have longer sentences, higher vocabulary ranges tend to pattern with an increased use of longer words, and there is a weak but significant relationship between larger lexical ranges and longer sentences.

7 Conclusion

This paper has introduced the *Corpus of the Canon of Western Literature (Version 1)*, a corpus of approximately 73 million words that represents the construct of the Western Canon according to Bloom (1994). Future releases of the CCWL aim to add more markup to the files, such as date of publication, more genre categories, and when required the translators and original languages. Further markup will help researchers disambiguate how such variables affect canonical literature. A few limitations of the corpus and its analysis presented above are worth closing with. One general limitation on the corpus is the issue of translation for non-English texts. In translation, there is often a blend of the language and style of an era with that of the source material, the King James Bible being a good example. Also, as noted, the CCWL does not have complete representation of the Western Canon described in Bloom (1994). The open source nature of this corpus, however, allows for the CCWL to be updated (by anyone) with other editions, perhaps beyond Project Gutenberg, to improve coverage and quality. While much time and effort has tried to reduce noise and thus provide other researchers with accurate numbers and a useful corpus, noise still remains. It also should be noted that different corpus tools can produce variable estimates of word count, sentence length etc. Future releases will further reduce transcription errors, unwanted characters and any other non-target text that may still remain. While the Culturomic and stylistic analysis above has been introductory, future research can use this corpus for much more complex quantification of style and culture, e.g. which authors in the canon cluster together according to intertextuality or other style metrics? Are there differences in country of origin in literary preoccupations? Do male and female canonical authors (of which there are only approximately 7% for the latter) differ in their construction of themes, characters and narrative ideas? How have what Adler and Weismann termed the ‘great ideas’ contained in the Western Canon spread throughout literature across time and place? The Canon of Western Literature has been an important and contested idea in literary studies, and the corpus introduced in this paper is hoped to be of use to scholars interested in Culturomics and Stylistics.

Notes

1. Future releases and a permanent online repository to be announced via Corpora-list: mailman.uib.no/listinfo/corpora.

2. Gibbon's *Decline and Fall* is a single work across multiple volumes in the corpus. The reported mean is for the single work as a whole. This was also done for *Parizval*, *Lives of the Artists* and *Don Quixote*. It was not done for different works across multiple volumes by the same author.

References

Acerbi A, Lampos V, Garnett P and Alexander R (2013) The expression of emotions in 20th century books. *PloS One* 8(3): doi.org/10.1371/journal.pone.0059030.

Adler M J, and Weismann M (2000) *How to think about the great ideas: From the great books of Western civilization.* Chicago: Open Court Publishing.

Anthony L. (2015). *TagAnt (Version 1.2.0)* [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Baker P (2003) No effeminates please: a corpus-based analysis of masculinity via personal adverts in Gay News/Times 1973–2000. *The Sociological Review* 51(1): 243–260.

Baker P (2004) Querying keywords questions of difference, frequency, and sense in keywords analysis. *Journal of English Linguistics* 32(4): 346–359.

Beach R, Appleman D, Fecho B, and Simon R (2016) *Teaching Literature to Adolescents*. London: Routledge.

Bloom H (1994) *The Western Canon: The Books and School of the Ages*. New York: Harcourt.

Borja M (2014) How unreadable are James Joyce's novels? *Significance 11*(3). Retrieved from https://www.statslife.org.uk/culture/1572

Craig H (2011) Shakespeare's vocabulary: Myth and reality. *Shakespeare Quarterly* 62(1): 53–74.

Wood D (2012) Character synthesis in the adventures of Huckleberry Finn. *The Explicator* 70(2): 83–86.

Fish SE (2001) *How Milton Works*. Harvard: Harvard University Press.

Givón T (1993) *English Grammar: A Function–based Introduction*. Amstredam: Benjamins.

Gorak J (2013) *The Making of the Modern Canon: Genesis and Crisis of a Literary Idea*. London: Bloomsbury.

Greenbaum S and Nelson G (1995) Clause relationships in spoken and written English. *Functions of Language* 2(1): 1–21.

Greenfield PM (2013) The changing psychology of culture from 1800 through 2000. *Psychological Science* 24(9): 1722–1731.

Guillory J (2013) *Cultural Capital: The Problem of Literary Canon Formation*. Chicago: University of Chicago Press.

Halliday MAK (2003) *On Language and Linguistics*. London: A&C Black.

Highet G (2015) *The Classical Tradition*. New York: Oxford University Press.

Holmes DI (1994) Authorship attribution. *Computers and the Humanities* 28(2): 87–106.

Hughes JM, Foti N, Krakauer D and Rockmore D (2012). Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the National Academy of Sciences* 109(20): 7682–7686.

Ingram J, Hand C and Maciejewski G (2016) Exploring the measurement of markedness and Its relationship with other linguistic variables. *PloS One* 11(6): doi.org/10.1371/journal.pone.0157141.

Leavis FR (2011) *The Great Tradition: George Eliot, Henry James, Joseph Conrad.* London: Faber & Faber.

Mahlberg M and McIntyre D (2011) A case for corpus stylistics: Ian Fleming's Casino Royale. *English Text Construction* 4(2): 204–227.

McIntyre D (2015) Towards an integrated corpus stylistics. *Topics in Linguistics* 16(1): 59–68.

Michel JB, Shen YK, Aiden AP, Veres A, Gray M K, Pickett JP and Pinker S (2011) Quantitative analysis of culture using millions of digitized books. *Science* 331(6014): 176–182

O'Halloran K (2007) The subconscious in James Joyce's Eveline: a corpus stylistic analysis that chews on the Fish hook. *Language and Literature* 16(3): 227–244.

Pechenick E, Danforth C and Dodds P (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS One 10*(10): e0137041.

Pitts M and Versluys MJ (2014) *Globalisation and the Roman world: World history, Connectivity and Material Culture*. Cambridge: Cambridge University Press.

Rayson P, Wilson A and Leech G (2001) Grammatical word class variation within the British National Corpus sampler. *Language and Computers* 36(1): 295–306.

Samothrakis S and Fasli M (2015) Emotional sentence annotation. *Plos One* 10(11): doi.org/10.1371/journal.pone.0141922

Scott M (2016). *Wordsmith (Version 7)* [Computer Software]. Liverpool: OUP.

Someya Y(1998) *E-Lemma* [Data file].Retrieved from http://www.lexically.net/downloads/e_lemma.zip

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

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

Toolan M (2009) *Narrative Progression in the Short Story: A Corpus Stylistic Approach*. Amsterdam: John Benjamins.

Towheed S and Owens WR (2011) *The History of Reading: International Perspectives, c. 1550–1945*. London: Palgrave Macmillan.

Wierzbicka A (1997) *Understanding Cultures through their Key Words: English, Russian, Polish, German, and Japanese*. Oxford: Oxford University Press.