Korean Journal of English Language and Linguistics, Vol 21, November 2021, pp. 1128-1144
DOI: 10.15738/kjell.21..202111.1128

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474
http://journal.kasell.or.kr

A Corpus Stylistic Analysis of Writing Style in English Literary Texts over Time

Jisu Ryu (Konkuk University) **Soonbae Kim** (Chungbuk National University)
Moongee Jeon (Konkuk University)

This is an open-access article distributed under the terms of the Creative Commons License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received: October 11, 2021
Revised: November 15, 2021
Accepted: November 27, 2021

Ryu, Jisu (first author)
Lecturer, Dept. of English Language & Literature, Konkuk University
jsryu0508@konkuk.ac.kr

Soonbae Kim
Professor, Dept. of English Language & Literature, Chungbuk Nat'l University
pearlpoet@chungbuk.ac.kr

Moongee Jeon (corresponding author)
Professor, Dept. of English Language & Literature, Konkuk University
mjeon1@konkuk.ac.kr

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020S1A5A2A01043615).

ABSTRACT

Ryu, Jisu, Soonbae Kim and Moongee Jeon. 2021. A corpus stylistic analysis of writing style in English literary texts over time. *Korean Journal of English Language and Linguistics* 21, 1128-1144.

This study aims to analyze the characteristics of the writing style of Jane Austen, a well-known novelist, based on the corpus linguistic method. Specifically, the current study compares Austen's *Pride and Prejudice* (1813) with *Persuasion* (1817) written with a four-year time gap using Coh-Metrix. This study analyzed whether Austen's writing style changes within a short period of four years despite the fact that the two literary works belong to the same genre. The Coh-Metrix measures selected for this study include basic counts, word features (word frequency, lexical diversity, imageability, concreteness, age of acquisition, familiarity, meaningfulness), personal pronouns, connectives, main parts of speech, readability indices, syntactic features, and reference cohesion measures. The findings of the study show that Austen's writing style changes at the vocabulary level (average word and sentence length, nouns, verbs, adjectives, pronouns, word frequency, lexical diversity, imageability, concreteness, age of acquisition, familiarity, meaningfulness, causal and additive connectives), but is maintained at the sentence level (readability indices, reference cohesion, syntactic complexity and similarity). In other words, Austen's writing style dramatically changes in terms of vocabulary use over a relatively short period of four years. The results of this study provide linguistic and pedagogical implications related to the analysis of literary texts.

KEYWORDS

corpus stylistics, corpus linguistics, stylistics, writing style analysis, writing style changes

© 2021 KASELL All rights reserved

1. Introduction

Corpus stylistics, which has recently attracted attention within linguistics, provides a new methodological paradigm that is applied to examine the writing style of writers by linking linguistics and literature (Hardy 2003, Hoover 1999, Louw 1993, Mahlberg 2007, 2013, Mastropierro 2017, McEnery, Xiao and Tono 2006, Semino and Short 2004, Stubbs 2005, Studer 2014, van Peer 1989, Wynne 2006). Corpus stylistics refers to a field of study that investigates electronic literary texts based on various types of statistical models (Wynne 2006). In other words, corpus stylistics is a discipline that analyzes the style of literary texts by means of qualitative and quantitative corpus linguistic methods (Lindquist 2009, McEnery and Wilson 2007, Meyer 2002).

Corpus linguistics (Lindquist 2009, McEnery and Wilson 2007, Meyer 2002) provides a new language analysis paradigm that is applied to statistically analyze multifarious types of text data based on methods which have been widely used in the field of computational linguistics (Jurafsky and Martin 2008). Since the corpus linguistic techniques are used to analyze multilevel linguistic features inherent in the text by computer tools and statistics, they contribute to increasing the reliability and validity of data analysis (Lindquist 2009, McEnery and Wilson 2007, Meyer 2002). Because the corpus stylistic approach is based on the corpus linguistic methods using computer tools and statistical analyses, it is effectively used to delve into writers' writing style. Traditionally, stylistic researchers (Enkvist 1964, Fowler 1971, Hough 1972, Ullmann 1964, Verdonk 2002) have studied the stylistic characteristics and changes of literary texts grounded on linguistic features that influence the writing style. Considering this, corpus stylistics is essentially a field in which corpus linguistic methods are applied effectively to examine the stylistic features of literary texts (Mahlberg 2007).

Corpus stylistic techniques are used in various stylistic studies (Louw 1993, Mahlberg 2007, McEnery et al. 2006, Semino and Short 2004, Stubbs 2005, Studer 2014, van Peer 1989, Wynne 2006). These studies include a study that analyzes the semantic prosodies inherent in literary texts (Louw 1993), a study that examines Leech and Short's (1981) *Speech and Thought Presentation model* using annotated corpus data (Semino and Short 2004), a study that compares British and American novel corpora with William Golding's book *The Inheritors* (Hoover 1999), a study that scrutinizes the linguistic patterns reflected in Flannery O'Connor's novels (Hardy 2003), a study that analyzes the influence of time and genre on writers' writing styles (Hartley and Jeon 2010), and a study that compares Alan Lightman's novel *Einstein's Dreams* with narrative and scientific text corpora (Graesser et al. 2008).

The studies presented above attempt to convert various types of printed literary texts into an electronic form that is processed with a computer platform, and then analyze these computerized corpus data using computer tools. The computer utilities currently useful for handling corpus data include WordSmith Tools (Scott 2004), TextSTAT (Bennett 2010), etc. These tools serve to increase the reliability of data analysis because they deal with massive corpus documents in an objective manner. The main functions of the WordSmith Tools and TextSTAT programs, however, are limited to concordance analyses. In other words, they are mainly used for searching keywords and applied for

© 2021 KASELL All rights reserved

frequency examinations of specific linguistic expressions in corpora. Therefore, they intrinsically reveal limitations in detecting various linguistic characteristics in discourse (Graesser, Jeon, Yan and Cai 2007).

A computer tool called Coh-Metrix (Graesser et al. 2007, Graesser, McNamara, Louwerse and Cai 2004, Jeon 2014) based on advanced technologies and multilevel linguistic indices presents a methodological alternative to the extant concordance-based text analysis programs. The Coh-Metrix system is a computer program developed by the Institute for Intelligent Systems (IIS) at the University of Memphis, and is used to illuminate the linguistic properties of corpus data by a wide range of linguistic measures. The Coh-Metrix system provides basic count measures, word frequency measures, vocabulary diversity measures (type-token ratios), vocabulary characteristic measures, pronouns, conjunctions (the total number of conjunctions, causal conjunctions, additive conjunctions, temporal conjunctions), content words, standard readability indicators (Flesch Kincaid Grade Level score, Flesch Reading Ease score), sentence syntax structure measures, and cohesion measures (referential and semantic cohesion). Given a myriad of linguistic measures the Coh-Metrix system affords, it can be efficaciously utilized to observe linguistic patterns on multiple levels.

In this study, two pieces of Jane Austen's literary texts, which occupy a key position in the 19th century English literature, were examined by the corpus stylistic methods. Specifically, Jane Austen's fiction *Pride and Prejudice* (1813) and *Persuasion* (1817) were analyzed on the primary measures provided by the Coh-Metrix system. Both literary works published solely four years apart belong to the same genre (i.e., romance novels), which makes them a particularly interesting comparison. The primary goal of the current study is to explore whether Jane Austen's writing style changes over a relatively short time based on specific linguistic features which are subject to change or not. This issue of possible shifts in writing style over time has been widely discussed in previous corpus stylistic studies (e.g., Altintas, Can and Patton 2007, Can and Patton 2004). The general conclusion is that writing styles tend to vary significantly over a long period of time. In this respect, previous studies fail to observe stylistic changes within a relatively short period of time (Can and Patton 2004, Lee, Park and Seo 2006, Pennebaker and Stone 2003). This study basically attempts to analyze the effect of time on Austen's writing style reflected in her two literary texts while controlling a potential compounding variable (i.e., genre type). In other words, this study aims to analyze whether a writer's style changes in a short time based on the corpus stylistic methodology. Unlike previous studies, this study tries to investigate the stylistic transformations of literary texts written in a short time period by the same author using many linguistic measures, and this is the justification for differentiating the current study from previous studies (e.g., Altintas, Can and Patton 2007, Can and Patton 2004), based mainly on lexical measures, which show that the style of literary texts changes over a long period of time.

© 2021 KASELL All rights reserved

2. Previous Studies

2.1 Corpus Stylistic Approach

In order to explore the stylistic characteristics of a writer, stylistic researchers pay attention to linguistic features included in his or her literary texts. According to stylistic researchers (Enkvist 1964, Fowler 1971, Gibbons and Whiteley 2018, Hough 1972, Verdonk 2002), the style of certain literary texts is clearly revealed through comparison with other literary texts. In other words, the style of a specific literary text is uncovered by comparing the linguistic features explicitly presented in it with those of another literary text to be compared. Stylistic researchers are inclined to use some proportional values to examine these linguistic features (Enkvist 1985, Gibbons and Whiteley 2018, Spencer and Gregory 1964). For example, if a specific linguistic expression of a literary text appears more frequently than in other texts, then this linguistic expression tends to be a significant element characterizing the style of the literary text (Enkvist 1985). This approach, which quantifies linguistic features to examine the style of literary texts, fits well with the approach of corpus linguistics (Meyer 2002). In essence, corpus stylistics can be defined as a field based on qualitative analyses depending on literary theories as well as quantitative analyses depending on corpus linguistic theories (Mahlberg 2007, Semino and Short 2004, Wynne 2006).

Since corpus stylistics is based on corpus linguistics, it reflects the advantages of the corpus linguistic approach. Corpus refers to an electronic text that is processed on a computer (Meyer 2002). Corpus linguistics is a field of study that analyzes linguistic features in the corpus data using computer tools (Meyer 2002). Thus, the corpus linguistic approach includes many advantages in conducting language studies (Lindquist 2009, McEnery and Wilson 2007, Meyer 2002). First, in the corpus linguistics research, a text is examined by computer tools in an objective way (WordSmith Tools, TextSTAT, Coh-Metrix, etc.), so the accuracy and speed of data analysis can be facilitated. Second, if a study uses the corpus linguistic method, it can be free from the bias of data analysis. Humans tend to have a cognitive bias when evaluating data (Goldstein 2015), thereby indicating that their experiences including background knowledge are reflected in data evaluation. If the corpus linguistic method, however, is used in data processing, the cognitive bias can be effectively removed. Third, since the corpus linguistic approach is based on a vast amount of language data and computer tools, it aids in revealing meaningful linguistic features that have not yet been discovered due to the limitations of human attention and perceptual ability (Goldstein 2015). Therefore, the corpus stylistic approach serves to improve the quality of literary text analysis by combining the advantages of corpus linguistics and literary theories (Mahlberg 2007).

Despite positive aspects of the corpus stylistic approach, there are also skeptical aspects. In the corpus stylistic approach, the linguistic features that characterize the style of literary texts are quantified and then analyzed with computer tools. Therefore, there is a possibility that the qualitative aspects existing in the literary texts may be overlooked. In other words, quantitative analyses are likely to offset qualitative aspects such as the context and meaning of texts that are critical for understanding literary texts (van Peer 1989, Wynne 2006). The corpus stylistic approach, however, essentially emphasizes the balance between a quantitative approach and a qualitative approach (McEnery 2006). In the corpus stylistic approach, only the positive aspects of the quantitative approach

© 2021 KASELL All rights reserved

are not emphasized. The corpus stylistic approach aims to improve the quality of literary text analysis based on the efficiency and advantages of the corpus linguistic approach. In other words, the corpus stylistic approach contributes to a comprehensive understanding of literary texts by combining a quantitative approach and a qualitative approach. This study attempted to maintain a balance between quantitative and qualitative approaches when analyzing *Pride* and *Prejudice and Persuasion* in terms of the corpus stylistic approach.

2.2 Studies in Corpus Stylistics

There are various studies based on the corpus stylistic approach (Graesser et al. 2008, Martin 2011, Mastropierro 2017, Semino and Short 2004, Stubbs 2005, Studer 2014, van Peer 1989, Wynne 2006). Specifically, there are empirical studies that analyze various linguistic expressions and characteristics that exist in the literary corpus by using concordancing programs such as WordSmith Tools and TextSTAT. For example, Louw (1993) suggested that the phenomenon of semantic prosody in text could be clearly revealed by a computational method, that is, a corpus stylistic method. Semantic prosody refers to a phenomenon in which the meaning of a vocabulary used with a specific vocabulary is recognized close to the meaning of that specific vocabulary (Partington 1998). Looking at the case of 'symptomatic of' among the examples (utterly, bent on, symptomatic of) presented by Louw, the meaning of 'symptomatic of' is itself neutral, but 'symptomatic of' tends to be usually accompanied by negative vocabularies in text. So, the meaning of 'symptomatic of' can be perceived negatively due to the influence of the negative words. Therefore, Louw mentioned that if positive words were used after 'symptomatic of', the meaning of 'symptomatic of' could be perceived ironically. Louw indicated that the corpus stylistic approach could be effectively utilized because these paradoxical semantic prosody phenomena may not be recognized otherwise.

Mahlberg (2007) used the corpus stylistic method to analyze linguistic expressions that perform textual functions in the novel, *Bleak House* by Charles Dickens. Mahlberg used the WordSmith Tools program, a concordancing program, to search for linguistic expressions included in the *Bleak House* corpus. Mahlberg was particularly interested in vocabulary groups consisting of five words among the linguistic expressions that existed in the *Bleak House* corpus. Mahlberg classified the vocabulary groups consisting of five words found in the *Bleak House* corpus into 5 categories (labels, speech, body parts, as if, time and place). Examples of labels consist of 'Mr. Pickwick and his friends', 'the Lady of the caravan', 'man with the wooden leg, etc. Examples of speech include 'do me the favor to, 'all I can say is', etc. Examples of body parts consist of 'with his hand to his', etc. Examples of as if embrace 'as if he would have', etc. Finally, examples of time and space include 'a quarter of an hour', 'up and down the room', etc. Mahlberg's study suggested that these vocabulary groups performed important textual functions in the *Bleak House* corpus.

Martin (2011) analyzed comparative structures used in the play texts (e.g., *The Admirable Bashville*, *Androcles and the Lion*, *Annajanska*, *The Apple Cart*, *Arms and the Man*, *Augustus Does His Bit*, *Back to Methuselah*, etc.) of the playwright Bernard Shaw based on the corpus stylistic method. Martin used the WordSmith Tools program to search for comparison phrases used in the play corpus. Specifically, Martin analyzed vocabularies related to similes (*as if*, *like the*, *like a/an*, *as the*, etc.) and vocabularies related to the

© 2021 KASELL All rights reserved

comparative degree (*than*, *most*, etc.) extracted from the play corpus. The results showed that these comparative phrases used in Bernard Shaw's play texts played an important role in characterizing his writing style.

In addition to these concordancing studies, there are studies using the Coh-Metrix system (Graesser et al. 2004) which was effectively used for various corpus analyses based on a wide range of linguistic measures. In particular, they investigate the effect of genre and time on writers' writing style by anatomizing corpora taken from various authors and genres (Hartley and Jeon 2010, Graesser et al. 2008, Hartley, Branthwaite, Ganier and Heurley 2007).

Graesser et al. (2008), for instance, analyzed the novel *Einstein's Dreams* written by Alan Lightman, a physicist at the Massachusetts Institute of Technology (MIT), using the Coh-Metrix system. They compared the novel with narrative and science texts. The main purpose of the study was to analyze the influence of genre on writing style. If Alan Lightman's *Einstein's Dreams* is similar to the narrative texts, the genre can influence Alan Lightman's writing style. On the other hand, if *Einstein's Dreams* is similar to science texts, Alan Lightman may write the novel as a scientific writing style because he is a physicist (i.e., scientist). Graesser et al. extracted the narrative and science corpora from the Touchstone Applied Science Associates (TASA) corpus for the purpose of comparing them with *Einstein's Dreams*. The Coh-Metrix measures analyzed for their study were basic counts (word count, sentence count, average word length, average sentence length), word frequency, conjunctions (all conjunctions, causal conjunctions, additive conjunctions, temporal cohesion), situation model (causal cohesion, intentional cohesion, temporal cohesion, spatial cohesion), standard readability indices (Flesch Reading Ease score, Flesch-Kincaid Grade Level score), syntactic complexity (noun phrase density, the number of words before the main verbs), lexical diversity (type-token ratios), pronoun ratios, etc. The results showed that *Einstein's Dreams* was similar to the narrative texts for most Coh-Metrix measures except for spatial cohesion, lexical diversity, additive conjunctions, and pronoun ratios. These findings indicated that the genre of fiction strongly influenced the writing style of Alan Lightman.

In addition to the studies of genre on writing style, researchers have been extensively concerned with the issue of time that induces shifts in writing style. For example, Can and Patton (2004) analyzed the works (old vs. new texts) of two Turkish authors. Specifically, they analyzed whether the writing style of the authors changes over time for the frequencies of word lengths. The results showed that longer vocabularies were used frequently in their new works than old works, indicating that their writing styles changed over time. Contrary to their findings, other researchers demonstrate that there are studies showing that author's writing style does not change over time (Hartley, Howe and McKeachie 2001, Hartley and Jeon 2010). Hartley and Jeon (2010), for instance, analyzed weekly newsletters written by Alistair Cooke, a well-known journalist, to study the changes of the author's writing style over time. In detail, they examined the newsletters written over 50 years based on sentence length, Flesch Reading Ease (FRE) score, Flesch-Kincaid Grade Level (FKGL), and passive sentence measures. The findings showed that there were no significant changes in these indices over time. These results suggest that the author's writing style has not changed significantly over time, thereby indicating that Alistair Cooke's writing style remained almost the same for a long time. In summary, author's writing style can alter (Can and Patton 2004, Lee, Park and Seo 2006, Pennebaker and Stone 2003) or remain unaffected by time (Hartley, Howe and McKeachie 2001, Hartley and Jeon 2010). So far, a number of studies based on the corpus stylistic theory were

© 2021 KASELL All rights reserved

presented. These studies suggested that corpus linguistic methods can be usefully applied in stylistic research. The current study attempts to analyze whether Jane Austen's writing style changes over time for literary texts of the same genre based on the multilevel linguistic features of Coh-Metrix.

3. Method

3.1 Corpora for this Study

The *Pride and Prejudice* and *Persuasion* corpus used for this study were obtained through the Gutenberg Project homepage (http://www.gutenberg.org/) where classical literary works whose copyrights had been expunged were available for free in an electronic form. *Pride and Prejudice* consists of a total of 61 chapters, and *Persuasion* contains a total of 24 chapters. Therefore, a total 26 electronic text files for *Pride and Prejudice* and a total of 24 electronic text files for *Persuasion* were created for this study. However, since the length of each file was different, each file was divided into approximately 3KB units to control the effect of the length. So, a total of 322 electronic text files for *Pride and Prejudice* and a total of 204 electronic text files for *Persuasion* were used for analyzing differences between two literary works over time.

3.2 Corpus Construction and Tool

In this study, the Coh-Metrix system (Graesser et al. 2004) was used to compare *Pride and Prejudice* and *Persuasion*. Since the Coh-Metrix system for desktop computers used in this study only can process files in Unicode format, all corpus files applied in this study were saved as text files in Unicode format (i.e., txt files). For this study, a total of 322 corpus files of *Pride and Prejudice* and a total of 204 corpus files of *Persuasion* were analyzed. The Coh-Metrix system used in this study is a program developed by the Institute of Intelligent Systems at the University of Memphis, and is used to analyze various types of corpus data. The Coh-Metrix system provides various types of linguistic and psycholinguistic measures (Graesser et al. 2004, Jeon 2014). Among the measures provided by the Coh-Metrix system, which have been extensively used for Coh-Metrix studies (Crossley and McNamara 2011, Graesser et al. 2004, Graesser et al. 2007, Graesser et al. 2008, Hartley and Jeon 2010) and are related to the analysis of the corpus data of this study, a total of 26 indices were selected for this study. Specifically, Coh-Metrix indices used for this study included basic counts (the number of words, the number of sentences, average sentence length, average word length), word frequency, lexical diversity (type-token ratio), word characteristics (imageability, concreteness, age of acquisition, familiarity, meaningfulness), content words (nouns, verbs, adjectives, adverbs), personal pronouns (the first-person, the second-person, the third-person, all pronouns), connectives (causal connectives, additive connectives, temporal connectives, all connectives), standard readability indices (Flesch Reading Ease score, Flesch-Kincaid Grade Level), sentence syntax analysis (syntactic complexity, syntactic similarity), and reference cohesion (argument overlap) measures. Detailed descriptions of each measure are presented in the result section for convenience.

© 2021 KASELL All rights reserved

4. Results and Discussions

In this study, one-way ANOVA was conducted to compare the corpus of *Pride and Prejudice* and *Persuasion*. Specifically, corpus type (*Pride and Prejudice* vs. *Persuasion*) was set as an independent variable, and each measure of the Coh-Metrix system was set as a dependent variable. In this study, SPSS 16 was used for statistical analyses which were performed at $p = .05$.

4.1 Basic Counts

The basic counts include the number of words, the number of sentences, the average word length, and the average sentence length. As presented in Table 1, there was a significant difference between *Pride and Prejudice* and *Persuasion* for the number of words. Specifically, *Persuasion* had more words than *Pride and Prejudice.* Interestingly, *Pride and Prejudice* contained longer words and shorter sentences than *Persuasion*. These results suggest that Jane Austen used shorter words and longer sentences over time when writing novels, thereby indicating that her writing style was changed within a short period of four years (Altintas, Can and Patton 2007, Can and Patton 2004, Lee, Park and Seo 2006, Pennebaker and Stone 2003).

Table 1. Results of Basic Counts

Indices	Pride and Prejudice (1813) (n = 322)	Persuasion (1817) (n = 204)	*F*	*p*	*Eta*	*power*
Number of words	349(51)	394(50.2)	98.003	.000*	.155	1.000
Number of sentences	17.3(6.63)	17.2(6.85)	.002	.967	.000	.050
Average word length	1.44(.07)	1.42(.07)	11.631	.001*	.021	.926
Average sentence length	23(8.78)	26.7(11.7)	17.615	.000*	.032	.987

Note. * $p < .05$

4.2 Readability Indices

The standard readability indices provided by the Coh-Metrix system are the Flesch-Kincaid Grade Level and the Flesch Reading Ease scores (Graesser et al. 2004). The Flesch-Kincaid Grade Level measure is expressed as a score between 1 and 12, indicating that the higher the number is, the higher the reading difficulty is. The Flesch-Kincaid Grade Level score represents the U.S. grade level (grades 1-12), respectively. On the other hand, the Flesch Reading Ease score is displayed as a number between 0-100, indicating that the higher the number, the lower the reading difficulty.

As shown in Table 2, there were no statistically significant differences between *Pride and Prejudice* and *Persuasion* for both the Flesch-Kincaid Grade Level and the Flesch Reading Ease scores. These indicators are related to reading difficulty, and the findings of this study suggest that Jane Austen's writing style did not change for these measures. In other words, her writing style was maintained at the sentence level (Hartley, Howe and McKeachie 2001, Hartley and Jeon 2010).

© 2021 KASELL All rights reserved

Table 2. Results of Readability Indices

Indices	Pride and Prejudice (1813) (n = 322)	Persuasion (1817) (n = 204)	*F*	*p*	*Eta*	*power*
FRE	61.7(12.6)	59.6(15.1)	2.969	.085	.006	.405
FKGL	9.49(2.39)	9.89(2.23)	3.741	.054	.007	.488

Note. * $p < .05$

4.3 Word Features

Regarding the word features, the Coh-Metrix system provides word frequency, lexical diversity, and word features. The results of the word frequency analysis for both content and all words were presented in Table 3. The word frequency score provided by Coh-Metrix is a logarithmic value of the raw word frequency score. The reason for using transformed frequency values is that the distribution of the transformed values becomes close to the normal distribution, thereby indicating the converted frequency values are more suitable for statistical analyses (Graesser et al. 2004). As shown in Table 3, statistically significant differences were found between the two works for the word frequency scores for the content and all words.

Table 3. Results of Word Features

Indices	Pride and Prejudice (1813) (n = 322)	Persuasion (1817) (n = 204)	*F*	*p*	*Eta*	*power*
Frequency						
Word frequency (content words)	2.44(.16)	2.47(.15)	5.499	.019*	.010	.648
Word frequency (all words)	3.20(.07)	3.22(.07)	8.080	.005*	.015	.810
Lexical diversity						
Type-token ratio (content words)	.81(.05)	.78(.06)	31.919	.000*	.056	1.000
Word characteristics						
Imageability	382(17.3)	385(15.6)	4.419	.036*	.008	.555
Concreteness	341(18.6)	346(17.2)	10.379	.001*	.019	.895
Age of acquisition	343(28.5)	337(26.7)	5.589	.018*	.010	.655
Familiarity	575(6.12)	575(5.37)	2.895	.089	.005	.397
Meaningfulness	410(12)	413(11.5)	7.978	.005*	.015	.805

Note. * $p < .05$

Specifically, more low-frequency words were used for *Pride and Prejudice*. This result is similar to the result of the word length analysis. As a result of the word analysis, more longer vocabularies were used for *Pride and Prejudice*. In general, longer vocabularies are more likely to be low frequency words. The lexical analysis results indicate that Jane Austen used easier language in *Persuasion*, thus suggesting that her writing style was changed for these measures.

As an index of lexical diversity for content words, Coh-Metrix uses a type-token ratio value. The type refers to an individual word contained in a text, and the token refers to the word when the individual word is reused in the text (Graesser et al. 2004). For example, if the vocabulary ‘time’ is repeatedly used 7 times in a text, the type

© 2021 KASELL All rights reserved

value of 'time' is 1 and the token value of 'time' is 7. The type-token ratio value is a value obtained by dividing the type value by the token value. Therefore, the higher the value of this ratio, the greater the likelihood that various words will be used in the text. As shown in Table 3, the type-token ratio value of *Pride and Prejudice* is higher than *Persuasion*. These results suggest that more various vocabularies were used in *Pride and Prejudice*, thereby implying that Jane Austen's writing style changed over time even when writing in the same genre (Can and Patton 2004, Lee, Park and Seo 2006, Pennebaker and Stone 2003).

Based on the MRC Psycholinguistics Database (Coltheart 1981), the Coh-Metrix system provides imageability, concreteness, age of acquisition, familiarity, and meaningfulness scores, which are measures related to vocabulary characteristics. The measure of imageability indicates the degree to which a particular word comes to mind. The measure of concreteness indicates how specific a particular word is. The acquisition age measure indicates at what age a particular vocabulary is acquired. The familiarity measure refers to how familiar a particular word is to a reader. The meaningfulness measure refers to how meaningful a particular vocabulary is (Graesser et al. 2004). These vocabulary measurements are presented on a scale of 100-700. As shown in Table 3, statistically significant differences were found between the two works for all measures except for the familiarity measure. These results consistently indicate that the vocabularies used in *Pride and Prejudice* are more difficult to understand than *Persuasion*. In other words, more imaginary, concrete, and meaningful words were used in *Pride and Prejudice*. In addition, more vocabularies acquired at higher ages were used in *Pride and Prejudice*. These results also suggest that Jane Austen's writing style changed over time, especially at the vocabulary level, similar to the word frequency and lexical diversity measures (Can and Patton 2004).

4.4 Major Parts of Speech

The major parts of speech analysis results are presented in Table 4. This measure was calculated as an incidence score. The incidence score refers to the value of a word that occurs in a text per 1,000 words.

Table 4. Results of Major Parts of Speech

Indices	Pride and Prejudice (1813) (n = 322)	Persuasion (1817) (n = 204)	*F*	*p*	*Eta*	*power*
Nouns	195(30.7)	201(29.3)	6.180	.013*	.011	.699
Verbs	138(15.7)	135(16.7)	5.364	.021*	.010	.638
Adjectives	62.9(14.5)	70.4(15.5)	32.876	.000*	.058	1.000
Adverbs	80.6(18.1)	79.7(16.3)	.327	.568	.001	.088

Note. * $p < .05$

As shown in Table 4, more verbs were used for *Pride and Prejudice*. On the other hand, more nouns and adjectives were used in *Persuasion.* These results imply that the sentences used in *Persuasion* can be syntactically complex due to modifiers such as adjectives. In fact, this result is in line with the result of the sentence length analysis, which indicated that more longer sentences were used for *Persuasion.* These results also suggest that Jane Austen's writing style changed at the vocabulary level over a period of four years (Altintas, Can and Patton 2007, Can and Patton 2004).

© 2021 KASELL All rights reserved

4.5 Personal Pronouns

The results of pronoun analysis were presented in Table 5. The measures of pronouns were also expressed as incidence scores. As shown in Table 5, *Pride and Prejudice* contained more pronouns statistically significantly than *Persuasion.* Specifically, *Pride and Prejudice* included more pronouns of all types than *Persuasion*. If a text contains more pronouns, the difficulty of the text tends to increase because it is necessary to find the objects that they refer to. Therefore, the difficulty of understanding *Pride and Prejudice* is expected to increase due to the pronouns. These results also indicated that Austen's style changed over time (Altintas, Can and Patton 2007).

Table 5. Results of Personal Pronouns

Indices	Pride and Prejudice (1813) (n = 322)	Persuasion (1817) (n = 204)	*F*	*p*	*Eta*	*power*
First person pronouns	15(12.7)	10.5(11.6)	17.090	.000*	.031	.985
Second person pronouns	7.35(8.34)	4.43(5.98)	18.989	.000*	.034	.992
Third person pronouns	37.6(13.8)	34.3(11.6)	8.529	.004*	.016	.830
All pronouns	133(28.2)	112(29.4)	73.363	.000*	.121	1.000

Note. * $p < .05$

4.6 Connectives

The results of connective analysis were presented in Table 6. As shown in Table 6, *Pride and Prejudice* contained more causal connectives than *Persuasion*. On the other hand, *Persuasion* consisted of more additional connectives than *Pride and Prejudice*. These results can be explained in connection with the results of sentence length analysis. In other words, the sentences used in *Persuasion* were longer than those used in *Pride and Prejudice*. It can be assumed that the additional connectives used in *Persuasion* functioned as the cause of the increase in sentence length.

Table 6. Results of Connectives

Indices	Pride and Prejudice (1813) (n = 322)	Persuasion (1817) (n = 204)	*F*	*p*	*Eta*	*power*
Causal connectives	27.6(10)	25.3(8.84)	7.009	.008*	.013	.753
Additive connectives	12.9(6.18)	14.1(5.94)	4.441	.036*	.008	.557
Temporal connectives	16.4(8)	17.5(8.57)	2.361	.125	.004	.335
All connectives	101(16.9)	101(17.7)	.105	.746	.000	.062

Note. * $p < .05$

The use of causal connectives can promote the degree of causal connection between sentences. Therefore, there is a possibility that the causality of *Pride and Prejudice* will increase due to these connectives rather than *Persuasion*, resulting in lowering the difficulty of text understanding. Austen's writing style was also influenced by the use of connectives over time (Can and Patton 2004).

© 2021 KASELL All rights reserved

4.7 Reference Cohesion

The measures provided by Coh-Metrix in relation to reference cohesion included the argument repetition ratios between adjacent sentences and the argument repetition ratios for all sentences in text. Arguments refer to nouns, pronouns, and noun phrases included in text (Graesser et al. 2004).

In general, when arguments are used repeatedly between sentences in a text, the cohesion of the text tends to increase, resulting in decreasing the difficulty of text understanding (Graesser et al. 2004). *Pride and Prejudice* and *Persuasion* showed no statistically significant difference for the reference cohesion measures (see Table 7). These results suggest that Austen's writing style was maintained for the use of arguments between sentences. In other words, Austen's writing style changed at the vocabulary level, but tended to be maintained at the sentence cohesion level (Hartley, Howe and McKeachie 2001, Hartley and Jeon 2010).

Table 7. Results of Reference Cohesion

Indices	Pride and Prejudice (1813) (n = 322)	Persuasion (1817) (n = 204)	*F*	*p*	*Eta*	*power*
Argument overlap (adjacent sentences)	.54(.18)	.54(.18)	.001	.977	.000	.050
Argument overlap (all sentences)	.47(.17)	.47(.17)	.013	.908	.000	.052

Note. * $p < .05$

4.8 Syntactic Features

The measures of syntactic features of Coh-Metrix included syntactic complexity and syntactic similarity measures. The syntactic complexity measures contained the noun density score and the number of words before main verbs (Graesser et al. 2004). The noun phrase density score represents a value obtained by dividing the number of modifiers that modify the headword of a noun phrase by the total number of words of the noun phrase. Therefore, as this value of a text increases, the syntactic complexity of the text increases. The number of words before main verbs refers to the number of all words before the main verbs. Therefore, similar to the noun phrase density score, the larger this value of a text, the greater the syntactic complexity of the text (Graesser et al. 2004).

The syntactic similarity indices are used to measure how similar the syntactic structures of sentences used in a text are. The syntactic similarity measures consist of the similarity values for adjacent sentences and the similarity values for all sentences in a text. The larger the values of these measures, the more sentences with similar syntactic structures in the text (Graesser et al. 2004). As shown in Table 8, *Pride and Prejudice* and *Persuasion* did not show significant differences for the syntactic similarity measures. These results suggest that Austen used sentences with similar syntactic structures in each work. As a result of syntactic complexity analysis, there was no significant difference between the two works for the number of words before the main verbs. However, for the noun phrase density measure, there was a significant difference between *Pride and Prejudice* and *Persuasion*. These results are similar to those of adjective analysis. In other words, more

© 2021 KASELL All rights reserved

adjectives were used in *Persuasion*. The use of these adjectives, in fact, is expected to contribute to increasing the value of noun phrase density. In summary, Austen maintained her own writing style for the syntactic complexity measure at the sentence level. However, over time she tended to use more complex noun phrases in her work (i.e., *Persuasion*).

Table 8. Results of Syntactic Features

Indices	Pride and Prejudice (1813) (n = 322)	Persuasion (1817) (n = 204)	*F*	*p*	*Eta*	*power*
Syntactic complexity						
Modifiers per NP	.64(.13)	.70(.15)	19.822	.000*	.036	.994
Words before main verbs	3.86(2.83)	3.86(2.44)	.000	.991	.000	.050
Syntactic similarity						
Similarity (adjacent sentences)	.08(.03)	.08(.03)	.754	.386	.001	.140
Similarity (all sentences)	.07(.02)	.07(.02)	.935	.337	.002	.160

Note. * $p < .05$

5. Conclusion

This study analyzed Jane Austen's *Pride and Prejudice* and *Persuasion* on the basis of corpus stylistic methods (Hartley and Jeon 2010, Hoover 1999, Louw 1993, Martin 2011, Mahlberg 2007, Mahlberg 2013, Mastropierro 2017, McEnery et al. 2006, Semino and Short 2004, Stubbs 2005). Specifically, the main purpose of the present study was to analyze the effect of time on Austen's writing style using the Coh-Metrix system, a language analysis tool (Graesser et al. 2004). The Coh-Metrix indices used for the current study consisted of basic counts (the number of words, the number of sentences, average sentence length, average word length), word frequency, lexical diversity (type-token ratio), word characteristics (imageability, concreteness, age of acquisition, familiarity, meaningfulness), content words (nouns, verbs, adjectives, adverbs), personal pronouns (the first-person, the second-person, the third-person, all pronouns), connectives (causal connectives, additive connectives, temporal connectives, all connectives), standard readability indices (Flesch Reading Ease score, Flesch-Kincaid Grade Level), sentence syntax analysis (syntactic complexity, syntactic similarity), and reference cohesion (argument overlap) measures.

The main findings of this study are as follows. First, *Pride and Prejudice* consisted of longer words than *Persuasion*. Longer sentences, however, were used in *Persuasion*. Next, the results of the analysis on vocabulary characteristics showed that *Pride and Prejudice* contained more low frequency vocabularies. Also, the vocabularies included in *Pride and Prejudice* were more diverse. These results suggest that the vocabularies used in *Pride and Prejudice* can be more difficult to understand than *Persuasion*. On the other hand, the vocabulary used for *Persuasion* was more specific and easily came to mind. And also more meaningful vocabularies were used for *Persuasion*. These results suggest that easy-to-understand vocabularies were used for *Persuasion*. Finally, *Pride and Prejudice* included more vocabularies acquired at a relatively higher age than

© 2021 KASELL All rights reserved

Persuasion. These results show that Jane Austen's writing style changed over time at the vocabulary level (Can and Patton 2004, Lee, Park and Seo 2006, Pennebaker and Stone 2003). In other words, Austen tended to use short and easy-to-understand words in her novels over time.

While the major parts of speech analysis results indicated that *Persuasion* contained more nouns and adjectives than *Pride and Prejudice*, fewer verbs were used in *Persuasion*. It is possible that the use of more adjectives can lengthen the sentences used in *Persuasion*. More pronouns and causal connectives were used in *Pride and Prejudice*, whereas *Persuasion* included more additive connectives. It can be assumed that the reason long sentences are used in *Persuasion* is due to the use of additive connectives. In summary, these results also suggest that Austen's writing style varied over time.

On the other hand, the results of reference cohesion analysis indicated that there were no statistically significant differences between the two works for argument overlap ratios for both adjacent and all sentences. The results of syntactic complexity also indicated that there was no significant difference between the two works for the number of words before main verbs. The two works subsequently showed no differences for syntactic similarity measures. These results suggest that Austen's writing style remains unchanged at the sentence level (Hatch, Hill and Hayes 1993, Hartley, Branthwaite, Ganier and Heurley 2007, Pennebaker and King 1999).

In summary, the current study suggests that Jane Austen's writing style changes at the vocabulary level over time, but remains unchanged at the sentence level. These results indicate that time influences writers' writing styles, unlike previous studies reporting that writers' writing style is not significantly affected by time (Hartley, Branthwaite, Ganier and Heurley 2007, Pennebaker and King 1999). In addition, the results of this study suggest that the corpus stylistic approach can be usefully applied to analyze literary texts. In other words, as successfully used in the present study, it is expected that the corpus stylistic method can contribute to in-depth analysis of complex and heterogeneous literary texts based on various linguistic features.

The results of the current study present pedagogical implications as well. For example, it is important to grasp the characteristics of genres and writers' writing styles in English literature classes. The characteristics of genres and the writing styles of authors can be defined by the linguistic features reflected in literary works. The results of this study can provide students with objective genre characteristics and writers' writing style characteristics based on various linguistic measures, thereby promoting their deep interpretation of literary texts. Learners also can develop essays that reflect genre's characteristics by applying the corpus stylistic methods applied to this study in English writing classes. The Coh-Metrix system applied in this study also encourages students to analyze various types of literary works by themselves, thereby enhancing their understanding of the literary works. In summary, the corpus stylistic methodology applied in this study and its findings will be useful for understanding English literary works and teaching literary theories in English literary classes.

In addition to the implications of the current study, this study also recognizes some limitations. In this study, specific literary works of a specific author, Jane Austen, were used to analyze the changes of writers' writing styles over time. But at the same time, more diverse writers and literary texts are required to increase the degree of generalization of research findings. Stylistic changes were also analyzed based on literary texts written with a time interval of four years in the current study. However, a future research can analyze the stylistic changes over time in more detail by analyzing literary texts written at various time intervals. For the present study, the

© 2021 KASELL All rights reserved

Coh-Metrix system was used to analyze Austen's writing style. Coh-Metrix is useful for analyzing stylistic characteristics and changes because it provides a wide range of linguistic measures. Despite this usefulness, the Coh-Metrix system only provides average scores for measures, thus raising the need for a qualitative study. In other words, it is necessary to analyze example sentences in which the measures are used by concordancing programs such as WordSmith Tools and TextSTAT. Finally, future studies can apply the corpus stylistic methods used in this study in order to analyze literary texts' linguistic characteristics, writers' stylistic differences, genre differences, and differences in translations.

References

Altintas, K., F. Can and J. M. Patton. 2007. Language change quantification using time-separated parallel translations. *Literary and Linguistic Computing* 22(4), 375-393.

Bennett, G. R. 2010. *Using Corpora in the Language Learning Classroom: Corpus Liinguistics for Teachers*. Ann Arbor, Michigan: University of Michigan.

Can, F. and J. M. Patton. 2004. Change of writing style with time. *Computers and the Humanities* 38(1), 61-82.

Coltheart, M. 1981. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology* 33(4), 497-505.

Enkvist, N. E. 1964. On defining style: An essay on applied linguistics. In J. Spencer ed., *Linguistics and Style*, 1-56. Oxford: Oxford University Press.

Enkvist, N. E. 1985. Text and discourse linguistics, rhetoric and stylistics. In T. A. van Dijk, ed., *Discourse and Literature: New Approaches to the Analysis of Literary Genres*, 11-38. Amsterdam: John Benjamins.

Fowler, R. 1971. *The Languages of Literature: Some Linguistic Contributions to Criticism*. London: Routledge.

Gibbons, A. and S. Whiteley. 2018. *Contemporary Stylistics*. Edinburgh: Edinburgh University.

Goldstein, E. B. 2015. *Cognitive Psychology: Connecting Mind, Research, and Everyday Experience*. Boston: Cengage.

Graesser, A. C., M. Jeon, Z. Cai and D. S. McNamara. 2008. Automatic analyses of language, discourse, and situation models. In J. Auracher and W. van Peer, eds., *New Beginnings in Literary Studies*, 72-88. Cambridge: Cambridge Scholars Publishing.

Graesser, A. C., M. Jeon, Y. Yan and Z. Cai. 2007. Discourse cohesion in text and tutorial dialogue. *Information Design Journal* 15(3), 199-213.

Graesser, A. C., D. S. McNamara, M. M. Louwerse and Z. Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers* 36(2), 193-202.

Hardy, D. E. 2003. *Narrating Knowledge in Flannery O'Connor's Fiction*. Columbia, South Carolina:

© 2021 KASELL All rights reserved

University of South Carolina Press.

Hartley, J., A. Branthwaite, F. Ganier and L. Heurley. 2007. Lost in translation: Contributions of editors to the meanings of texts. *Journal of Information Science* 33(5), 551-565.

Hartley, J., H. J. A. Howe and W. J. McKeachie. 2001. Writing through time: Longitudinal studies of the effects of new technology on writing. *British Journal of Educational Technology* 32, 141-151.

Hartley, J. and Jeon, M. 2010. The effects of genres and voices on writing styles in Alistair Cooke's & Jeremy Clarkson's writings. *The Journal of Linguistic Science* 52(1), 243-262.

Hatch, J. A., C. A. Hill and J. R. Hayes. 1993. When the messenger is the message: Readers' impressions of writers' personalities. *Written Communication* 10(4), 569-597.

Hoover, D. L. 1999. *Language and Style in The Inheritors*. Lanham: University Press of America, Inc.

Hough, G. 1972. *Style and Stylistics*. London: Routledge.

Jeon, M. 2014. Analyzing the cohesion of English text and discourse with automated computer tools. *Journal of Pan-Pacific Association of Applied Linguistics* 18(2), 123-133.

Jurafsky, D. and J. H. Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall.

Lee, C., J. Park and Y. Seo. 2006. An analysis of linguistic styles by inferred age in TV dramas. *Psychological Reports* 99(2), 351-356.

Leech, G. and M. H. Short. 1981. *Style in Fiction: a Linguistic Introduction to English Fictional Prose*. London: Longman.

Lindquist, H. 2009. *Corpus Linguistics and the Description of English*. Edinburgh: Edinburgh University Press.

Louw, B. 1993. Irony in the text or insincerity in the writer? In M. Baker and G. Francis, eds., *Text and Technology: In Honour of John Sinclair*, 157-176. Amsterdam: John Benjamins.

Mahlberg, M. 2007. Corpus stylistics: bridging the gap between linguistic and literary studies. In M. Hoey and M. Mahlberg, eds., *Text, Discourse, and Corpora*, 219-246. London: Continuum.

Mahlberg, M. 2013. *Corpus Stylistics and Dickens's Fiction*. New York: Routledge.

Martin, G. A. 2011. Comparison and other "Modes of Order" in the plays of Bernard Shaw. *International Journal of English Studies* 12(2), 151-169.

Mastropierro, L. 2017. *Corpus Stylistics in Heart of Darkness and its Italian Translations*. London: Bloomsbury.

McEnery, T., R. Xiao and Y. Tono. 2006. *Corpus-based Language Studies*. London: Routledge.

McEnery, T. and A. Wilson. 2007. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Meyer, C. F. 2002. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Partington, A. 1998. *Patterns and Meanings*. Amsterdam and Philadelphia: Benjamins.

Pennebaker, J. W. and L. A. King. 1999. Linguistic styles: Language use as an individual difference.

© 2021 KASELL All rights reserved

Journal of Personality and Social Psychology 7(6), 1296-1310.

Pennebaker, J. W. and L. D. Stone. 2003. Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology* 85(2), 291-301.

Scott, M. 2004. *WordSmith Tools*. Oxford: Oxford University Press.

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

Spencer, J. and M. Gregory. 1964. An approach to the study of style. In J. Spencer, ed., *Linguistics and Style*, 57-105. Oxford: Oxford University Press.

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

Studer, P. 2014. *Historical Corpus Stylistics: Media, Technology and Change*. New York: Bloomsbury.

Ullmann, S. 1964. *Language and Style*. Oxford: Basil Blackwell.

van Peer, W. 1989. Quantitative studies of literature: a critique and an outlook. *Computers and the Humanities* 23(4), 301-307.

Verdonk, P. 2002. *Stylistics*. Oxford: Oxford University Press.

Wynne, M. 2006. Stylistics: corpus approaches. In K. Brown, ed., *The Encyclopedia of Language and Linguistics*, 223-226. Oxford: Elsevier.

Examples in: English
Applicable Languages: English
Applicable Level: Tertiary

© 2021 KASELL All rights reserved