

A CORPUS-BASED MULTIDIMENSIONAL ANALYSIS OF HEMINGWAY'S WRITING STYLE

Jianfeng Lian ^{1*}, Cholthicha Sudmuk², Angvarrah Lieungnapar³, Behrad Aghaei⁴,
Suwaree Yordchim⁵

^{1*,2,5}*Department of Linguistics, Graduate School, Suan Sunandha Rajabhat University, Bangkok, Thailand*

³*Department of Humanities, Faculty of Humanities and Social Science, Suan Sunandha Rajabhat University, Bangkok, Thailand*

⁴*Department of Middle East Studies, University of Michigan, Ann Arbor, Michigan, USA*
E-mail: s64584952007@ssru.ac.th^{1*}, E-mail: cholthicha.su@ssru.ac.th², angvarrah.li@ssru.ac.th³,
aghaeib@umich.edu⁴, suwaree.yo@ssru.ac.th⁵

ABSTRACT

This study has employed a corpus-based multidimensional analysis proposed by Biber (2013) to compare Hemingway's Usage of linguistic features in textual dimensions with those of the other Lost Generation writers and determine his writing style from his linguistic choices. The results found that 14 linguistic features are significantly different, among them, the most significant are: more to be as the main verb, shorter words, fewer prepositions, and fewer past participial WHIZ deletions, indicating that Hemingway's linguistic choices include purposes with less information, but more interaction. These choices can be identified as his descriptive writing style, which is straightforward and colloquial.

Keywords: Corpus-Based Multidimensional Analysis, Hemingway's Writing Style

INTRODUCTION

Ernest Hemingway, one of the Lost Generation Writers (Yaoxin, 2022), is popularly acknowledged as the foremost prose stylist for the influence that he has exerted on contemporary style (Noble Foundation, 2018). According to Wyatt (2017), Hemingway's works are marked by vulnerability and deep feelings of the awkwardness and embarrassment of everyday life. His style of writing is simple and natural on the surface, but deliberate and artificial (Xie, 2008). Even though his distinctive style is occasionally commented on and controversial, it was probably the most widely imitated in the 20th century (Renjing, 2014).

Writing style refers to the different linguistic choices that an author uses in the text. The "linguistic choices" means the selection of words to create a specific impact or tone in the writing, which can be identified as the writing style of the author (Isti'annah, 2013). Regarding Hemingway's writing style, most traditional scholars emphasize his language features such as clusters, word length, sentence length, and collocation. For example, Zhiqin (2010) conducted a corpus-based stylistic analysis of Hemingway's *Cat in the Rain* novels on phonology and lexical-grammatical features, and McIntyre and Walker (2019) conducted a case study on Hemingway's authorial style. The data of their research are purposively

selected from some of Hemingway's novels, and none of them study the relativity of textual features and specific linguistic features, which can determine the author's writing style (Leech, 2007).

This study employs the corpus-based technique, which utilizes a large and principled collection of texts as the basis for analysis, and the Multidimensional (MD) Analysis (Biber, 2013), which identifies co-occurrence patterns, namely textual dimensions, and linguistic features based on factor analysis, to compare Hemingway's usage of linguistic features in textual dimensions with those of the other Lost Generation writers and determine his writing style from his linguistic choices.

RESEARCH OBJECTIVES

1. To compare Hemingway's usage of linguistic features in textual dimensions with those of the other Lost Generation writers.
2. To determine Hemingway's writing style from his linguistic choices in textual dimensions.

METHODOLOGY

1. Research Method

This study has employed a Corpus-Based Approach by utilizing the MD Analysis proposed by Biber (2013).

2. Research Process

1) Collect data from Hemingway's and the other Lost Generation writers' novels; 2) The Antfile converter software (Windows (Installer) (2.0.2)) is used to transform novels PDF into "txt" files. All the "txt" files are cleaned manually, to cut the acknowledgments and advertisements to get clean data, only the titles and contents are left in the "txt" files; 3) The software MD Analyzer Tagger (MAT 1.3.3) is used to tag all the linguistic features and calculate the z-score of 6 dimensions and 48 linguistic features in the texts; 4) The SPSS (IBM SPSS Statistics 22) is used to conduct variance analysis of the above z-scores to find the significantly different dimensions and linguistic features in a study corpus and a reference corpus; and 5) Determine Hemingway's writing style from his linguistic choices in textual dimensions.

3. Data Collection

There are two groups of data in the study:

- 1) Hemingway's 9 novels are collected as a study corpus. In total, Hemingway's corpus consists of 747383 tokens as shown in Table 1.

Table 1: Hemingway's novels collected for analysis.

Name of Novels	Tokens	Publishing year
The Torrents of Spring	21709	(1926)
The Sun Also Rises	68970	(1926)
A Farewell to Arms	108756	(1929)
To Have and Have Not	58808	(1937)
For Whom the Bell Tolls	147465	(1940)
Across the River and into the Trees	69696	(1950)
The Old Man and the Sea	26736	(1952)
Islands in the Stream	176960	(1970)
The Garden of Eden	68283	(1986)

2) The other Lost Generation writers' 9 novels from random sampling as a reference corpus. The representatives of the other Lost Generation writers are F. Scott Fitzgerald, John Dos Passos, William Faulkner, James Thurber, and E. E. Cummings (Yaoxin, 2022). In total, the tokens in the reference corpus (778827 tokens) are almost equal to the study corpus (747383 tokens), as shown in Table 2.

Table 2: The other Lost Generation writers' novels collected for analysis.

Authors	Name of the novels	Tokens	Publishing year
James Thurber	Is Sex Necessary?	23920	(1929)
	The 13 Clocks	12215	(1950)
F. Scott Fitzgerald	The Great Gatsby	50133	(1925)
	The Love of the Last Tycoon	46826	(1941)
John Dos Passos	The 42nd Parallel	133460	(1930)
	The Big Money	180462	(1936)
E.E. Cummings	The Enormous Room	91463	(1922)
William Faulkner	The Sound and the Fury	133939	(1929)
	Absalom, Absalom!	106409	(1936)

4. Data Analysis

Biber (1988) classified 67 linguistic features into 7 dimensions by factor analysis, but later, he dropped dimension 7 and 9 linguistic features due to weak interpretation. After years of research and testing, Biber (2013) cut another 10 linguistic features out because they had weak factor loadings on the relevant dimensions, and their functional interpretation was unclear. This study follows Biber's framework (2013) in which 6 dimensions and 48 linguistic features as research variables to conduct a Contrastive Analysis by the software Multidimensional Analyzer Tagger (MAT 1.3.3).

1) Textual Dimensions

The 6 textual dimensions consist of (1) Informational versus Involved Production, (2) Narrative versus Non-Narrative Concerns, (3) Explicit versus Situation-Dependent Reference, (4) Overt Expression of Persuasion, (5) Abstract versus Non-Abstract Information, and (6) Online Informational Elaboration.

2) Linguistics Features

The 48 linguistic features consist of 26 linguistic features in dimension 1, 4 linguistic features in dimension 2, 7 linguistic features in dimension 3, 5 linguistic features in dimension 4, 5 linguistic features in dimension 5, and 1 linguistic feature in dimension 6. The 48 linguistic features and 6 textual dimensions are shown in Table 3.

Table 3: The 48 linguistic features and 6 textual dimensions

Dimension	Linguistic Features
Dimension 1	Involved Production features: private verbs, that-deletions, contractions, present tenses, 2 nd person pronouns, do as pro-verb, demonstrative pronouns, emphatics, 1st person pronouns, pronoun <i>it</i> , <i>be</i> as the main verb, causative subordinations, discourse particles, indefinite pronouns, hedges, amplifiers, direct wh- questions, possibility modals, non-phrasal coordination, wh clauses, stranded prepositions.
	Informational Production features: nouns, average word length, prepositions, type/token ratio, attributives adjectives.
Dimension 2	past tenses, 3rd person pronouns, perfect aspect, and public verbs.
Dimension 3	Explicit Reference features: wh relative clauses on object position, pied-piping constructions, wh relative clauses on subject position, nominalizations.

	Situation-dependent reference features: time adverbials, place adverbials, and general adverbs.
Dimension 4	predicative modals, suasive verbs, conditional subordinations, necessity modals, and split auxiliaries.
Dimension 5	conjuncts, agentless passives, by passives, past participial WHIZ deletions, and other adverbial subordinators.
Dimension 6	<i>That</i> clause as verb complements.

RESULTS

By applying the Corpus-Based Approach with the MD analysis to compare Hemingway's usage of linguistic features in textual dimensions with those of the other Lost Generation writers and determine his writing style from his linguistic choices, the findings are summarized as follows:

1. The comparison between Hemingway's usage of linguistic features in textual dimensions with those of the Other Lost Generation writers

The independent sampling T-test, statistics show both similarities and differences in linguistic features in textual dimensions between Hemingway and the other Lost Generation writers as follows:

1)Textual Dimensions

The Similarities

Hemingway's novels and other Lost Generation writers' novels have some similarities in the textual dimensions, which are Dimension 2, they focus on narrative concern, Dimension 4, they don't express their ideas to persuade the addressee, Dimension 5, they don't use highly abstract and technical information in their novels, and Dimension 6, they present the information in a relatively loose, fragmented manner.

The Differences

There are two textual dimensions, 1 and 3, that have some differences. In dimension 1, Hemingway's novels have a positive mean value ($M=6.844$), while other Lost Generation writers' novels have a negative one ($M=-1.0556$), and the mean difference is 7.9000. In dimension 3, the p-value ($p=0.001$) illustrates that Hemingway's novels have negative mean values ($M=-1.8322$), while other Lost Generation writers' novels have positive mean values ($M=1.7089$), and the mean difference is -3.54111.

The similarities and differences of linguistic features in textual dimensions between Hemingway and the other Lost Generation writers are shown in Table 4.

Table 4: The similarities and differences of linguistic features in textual dimensions between Hemingway and the other Lost Generation writers

Dimensions	Hemingway	Other LG	Mean difference	t value	P value
	Mean value	Mean value			
Dimension 1	6.844	-1.0556	7.9	3.301	0.005
Dimension 2	4.6356	4.3678	0.26778	0.315	0.757
Dimension 3	-1.8322	1.7089	-3.54111	-4.248	0.001
Dimension 4	-0.681	-0.5456	-0.13556	-0.0238	0.815
Dimension 5	-1.6422	-0.8333	-0.80889	-2.058	0.056
Dimension 6	-1.6533	-0.9544	-0.69889	-1.947	0.069

2) Linguistic Features

The Similarities

There are 34 out of 48 linguistic features that are not significantly different, which accounts for 70.8%, as shown in Table 5.

Table 5: 34 linguistic features in Hemingway's and other Lost Generation writers' novels are not significantly different

Textual Dimension	Linguistic Feature
Dimension 1	nouns, private verbs, subordinator that deletion, contractions, present tense verbs, do as pro-verb, demonstrative pronouns, emphatics, 1 st person pronouns, causative subordination, discourse participles, indefinite pronouns, hedges, amplifiers, non-phrasal coordination, stranded prepositions, and wh-clause.
Dimension 2	past tense verbs, third-person pronouns, perfect aspect verbs, and public verbs.
Dimension 3	wh-relative clauses on object positions, time adverbials, place adverbs, and adverbs.
Dimension 4	predictive modals, suasive verbs, conditional subordination, necessity modals, and split auxiliaries.
Dimension 5	conjuncts, agentless passives, by-passives, and other adverbial subordinators.

The Differences

There are 14 out of 48 linguistic features that are significantly different, which are illustrated in Table 6.

Table 6: The significantly different linguistic features between Hemingway and other Lost Generation writers.

Dimensions	Linguistic features	Hemingway	other LG	Mean Difference	P Value
		mean	mean	(MD)	
Dimension 1	Word length	-1.1556	-0.8156	-0.76	0.003
	Prepositions	-1.2611	-0.6011	-0.66	0.003
	Type-token ratio	-0.9074	-0.6011	-1.31222	0.017
	Attributive adjectives	-0.9367	-0.4144	-0.5222	0.042
	2 nd person pronouns	0.7556	0.2056	0.55	0.029
	Pronoun it	1.1467	0.44	0.7.667	0.006
	Be as the main verb	-0.6067	-1.2278	0.62111	0.001
	Direct Wh-questions	3.5833	1.6844	1.90889	0.008
Dimension 3	Possibility modals	0.2978	-0.1156	0.41333	0.047
	Pie piping constructions	-0.5178	0.1656	-0.68333	0.031
	Wh relative clauses on subject position	-0.6111	-0.15	-0.46111	0.026
Dimension 5	Nominations	-1.1167	0.72	-0.39667	0.008
	Past participial WHIZ deletions	-0.6224	-0.4944	-0.13	0.001
Dimension 6	That clauses as verb complements	-0.7144	-0.32	-0.39444	0.004

Table 6 shows the most significant different linguistic features: Hemingway uses more be as the main verb (MD=0.62111), shorter word length (MD=-0.7600), and fewer prepositions (MD=-0.6600) and past participial WHIZ deletions (MD=-0.39444).

2. Hemingway's writing style is determined by linguistic choices in textual dimensions.

1) Hemingway's linguistic choices in the textual dimension

As the results of the MD analysis shown in 5.1, two of Hemingway's textual dimensions, 1 and 3, have different linguistic features' usage from the other Lost Generation writers. Dimension 1, Involved versus Informational production, shows that he represents his novels with more interactional, affective, and involved purposes. On the other hand, in Dimension 3, Explicit versus Situation-Dependent reference, he uses fewer relative clauses on object positions, time adverbials, place adverbs, and adverbs. It shows that he likes to use nonspecific Explicit versus situation-dependent reference, and nonspecific time and place for locative and temporal reference.

2) Hemingway's writing style is determined by his linguistic choices.

The MD analysis results of the difference in linguistic features indicate that there are three most significant different linguistic features in Hemingway's linguistic feature usage: be as the main verb, shorter word length, fewer prepositions, and fewer past participial WHIZ deletions, which can identify his writing style as follows:

The verb to be is the main verb used to indicate the existence of an entity (person, object, abstraction) or relate an entity to its qualities or characteristics. This verb means to exist, to occur, or to have the characteristics of something. Hemingway's choice of the main verb to be is part of his descriptive writing style, which is used to portray imagery to create a clear picture in the readers' minds.

Hemingway's choice of short word length shows that his writing style is simple, involving purposes with less information since short words are easy to read and understand, leaving no room for misunderstanding. Hemingway's straightforward writing style represents his clarity, conciseness, and substance.

Prepositions are linguistic units that show a relationship in space or time or a logical relationship between two or more people, places, or things. Hemingway's choice of using fewer prepositions or prepositional phrases makes his novels more interactive and verbal to readers, which means Hemingway prefers to employ a colloquial style of writing to help the reader understand his ideas.

Past participial WHIZ deletions are used to present propositions with reduced emphasis on the agent, demote the agent to object position, or eliding the agent altogether, and used to give a prominent position to the patient of the verb, the entity acted upon, which marks abstract information rather than a concrete referent. Hemingway tends to use fewer past participial WHIZ deletions to achieve precise representation, and to make the readers understand his concrete information easily.

In sum, Hemingway's linguistic choices involve purposes with less information, but more interaction. These choices can be identified as his descriptive writing style, which is straightforward and conversational.

CONCLUSION AND DISCUSSION

This study has employed a corpus-based approach with the MD Analysis proposed by Biber (2013) to compare Hemingway's linguistic features in textual dimensions with those of the other Lost Generation writers to determine his writing style from his linguistic choices. By using short word length, more utilization of to be as the main verbs, fewer prepositions, and more locative and temporal references, Hemingway presents his involving purpose with less information, which can be identified as his straightforward and colloquial style.

Hemingway's specific linguistic choices found in this study match the findings of previous studies where the data collection was not as comprehensive as the present study. For example, Wangqin (2011) pointed out Hemingway tends to use mainly 4-letter words to directly express personal feelings and employ many conversations to help readers participate in the stories, Zhang Renxia (2010) also stated that Hemingway likes to use shorter words and less variety of vocabulary to make readers understand his ideas, and McIntyre and Walker (2019) emphasized that Hemingway tends to use shorter sentences and less complex sentences to make his works more conversational and straightforward.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to Suan Sunandha Rajabhat University for invaluable help throughout this research.

REFERENCES

- Arina Isti'annah. (2013, October 18). Linguistic Features for More Understanding Literary Work. The 1st Literary Studies Conference. Yogyakarta, Indonesia.
- Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press, Cambridge.
- Biber, D. (1995). *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge University Press, Cambridge.
- Chang Yaoxin. (2022). *A Survey of American Literature*. Nankai University Press.
- Leech. (2007). *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. Pearson Longman. New York.
- McIntyre & Walker. (2019). *Corpus Stylistics: Theory and Practice*. Edinburgh University Press.
- Susan Conrad & Biber, D. (2013). *Variation in English: Multi-Dimensional Studies*. Routledge.
- Wangqin. (2011). A Corpus-based Study on Stylistics of Hemingway's Novels. *Journal of Taizhou University*, Vol. 33 (2011), 37-39.
- Wyatt, Davis. (2017). *Hemingway, Style, and the Art of Emotion*. University of Maryland. College Park.
- Xie Yaochen. (2008). Hemingway's Language Style and Writing Techniques in *The Old Man and the Sea*. *English Language Teaching*, Vol.1 (2008), 156-158.
- Yang Renjing. (2014). *Essays on Ernest Hemingway*. Yilin Press.
- Zhang Renxia. (2010). Application of Concordance Method to Literary Criticism: A Case Study of a Farewell to Arms. *Journal of Guangdong University of Foreign Studies*, Vol. 21 (2010), 34-38.
- Zhang Zhiqin. (2010). The Interpretation of a Novel by Hemingway in Terms of Literary Stylistics. *Language, Society, and Culture*, Vol.30 (2010), 155-61.
- https://www.noble.org/wp-content/uploads/2023/04/2018_annual-report.pdf