Using Corpus Concordance Software for Low-Effort Multidimensional Analysis of Linguistic Style in Chinese Secondary EFL Textbooks

Mingxuan Zhang [1,*], Chong Liu [2], and Jayakaran Mukundan[1]

[1] School of Education, Faculty of Social Sciences and Leisure Management, Taylor's University, Selangor, Malaysia
[2] School of Arts, Weihai Vocational College, Weihai, China
Email: 0365972@sd.taylors.edu.my (M.X.Z.); 17865579677@163.com (C.L.); jayakaranmukundan@yahoo.com (J.M.)
*Corresponding author

***Abstract*—This study aims to investigate the linguistic style of English language textbooks by employing a "*low-effort*" Multidimensional Analysis (MDA) approach using corpus concordance software. Biber's MDA approach, developed to reveal linguistic registers, genres, and styles, is a powerful, robust yet complex method that demands substantial technical expertise and considerable time investment from researchers. To address these challenges, this study adopts a simplified analytical approach utilizing the *Wordlist* and *Keywords* functions in the corpus concordance software—WordSmith Tools. Eight Chinese secondary school English as a Foreign Language (EFL) textbooks were selected as the research sample, forming the foundation for constructing a self-built corpus, the Chinese Secondary School EFL Textbooks Corpus (CSTC). The findings reveal that the linguistic style of the sampled textbooks closely aligns with the characteristics of the written register. This study provides a practical reference of this "*low-effort*" method for future research that requires preliminary and general evaluations of linguistic style.**

***Keywords*—corpus concordance software, EFL textbooks, WordSmith Tools, linguistic style**

I. Introduction

Despite the increasing prevalence of electronic materials, English Language (EL) textbooks remain a core resource in English language teaching [1]. They provide learners with essential language input [2] and are especially crucial in non-native English-speaking regions (e.g., English as a foreign language in China) [3]. In these contexts, learners often have limited opportunities for real-life English interaction [4], making EL textbooks the primary and most critical source of exposure to formal English [5–7].

Some previous studies have investigated the linguistic style of textbooks or highlighted the necessity of defining their linguistic characteristics to support subsequent research. For example, Le Foll [8] focused on register variation in English as a Foreign Language (EFL) textbooks.

The Multidimensional Analysis (MDA) approach [9] was originally developed to analyze registers and investigate variation within the English language. It is widely used to identify registers, genres, or text types. However, this traditional MDA approach demands considerable effort, technical expertise, and specialized tools, such as the Multidimensional Analysis Tagger [10].

As an alternative, Tribble [11] proposed a "*low-effort*" MDA approach that utilizes the *Wordlist* and *Keywords* functions in WordSmith Tools [12]. This approach provides a quick and straightforward way of evaluating genres based on MDA dimensions, demonstrating effectiveness in capturing key genre characteristics identified through the MDA approach [13].

In studies where analyzing register variation is not the primary research objective, such as those examining the impact of register effects on learner writing in English (e.g., [14]), it is essential to first identify the relevant registers. This preliminary identification serves as a prerequisite for further analysis, ensuring the validity of subsequent findings. In such cases, the "*low-effort*" MDA approach proves particularly suitable for efficiently identifying register styles.

II. Literature Review

Corpus linguistics involves the empirical analysis of extensive, electronically stored, and representative collections of texts, commonly known as corpora [15]. This methodological approach offers researchers a fresh and insightful way to explore language patterns [16] and has been gaining increasing prominence in English education research.

Corpus linguistics can be integrated into teacher education through approaches such as Corpus-Based Language Pedagogy (CBLP) and Corpus-Based Reflective Practice (CBRP) [17]. A recent systematic review highlights the growing integration of corpus linguistics and Data-Driven Learning (DDL) in language classrooms. This integration enables educators to enhance

Manuscript received February 28, 2025; accepted July 4, 2025; published October 22, 2025.

teaching practices and foster language acquisition by effectively utilizing authentic language data [18].

In the context of English Language (EL) textbooks, corpus linguistics can also be applied to textbook development [19]. Numerous studies have investigated this approach in EL textbooks using corpus data derived from these materials [15]. A literature review spanning the past two decades (2005–2024) identified 42 relevant articles from the Web of Science (WOS) Core Collection database.

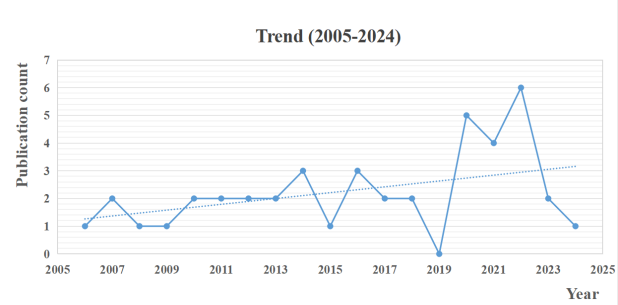

Fig. 1. Annual publication count over time (2004–2024).

As shown in Fig. 1, the overall trend over the past two decades demonstrates a fluctuating increase in publications, whereas the output in recent years has remained relatively small. From the perspective of research domains, the topics of these studies primarily include phraseological units (24%) (e.g., [20]), English textbook development (21%) (e.g., [21]), pragmatic features (19%) (e.g., [22]), and vocabulary (17%) (e.g., [7]). Among these topics, phraseological units have consistently constituted a primary focus. Regarding the selected textbook samples, 60% of the studies (25 articles) focus on the EFL context, with 29% (12 articles) specifically targeting Chinese EFL textbooks. This indicates that EFL textbooks, especially Chinese EFL textbooks, remain a central focus in corpus-related textbook research.

However, over the past three years (2022–2024), only one study [23] has briefly addressed the linguistic style of Chinese secondary school EFL textbooks using corpus concordance software. The study suggests that the linguistic style of these textbooks aligns with the Spoken British National Corpus Spoken Version 2014 (BNCS 2014). Nevertheless, it lacks a detailed analysis process.

In China, EFL textbooks play a crucial role, particularly in the elementary educational stage (primary and secondary schools) [24]. This is because students in higher educational stages (higher vocational colleges and universities) are capable of independent and autonomous learning, which reduces the reliance on textbooks by both teachers and students [25]. Therefore, future research should focus on EL textbooks used at the elementary educational stage.

This study aims to utilize corpus concordance software to demonstrate the "*low-effort*" MDA approach for analyzing linguistic style in Chinese secondary EFL textbooks, thereby providing a practical reference for future educators on how to apply this approach and offering insights into the linguistic styles presented in these textbooks to support further research. The research addresses the following question: To which register does the linguistic style of Chinese secondary school EFL textbooks most closely align?

III. Methods and Samples

The present study adopts a corpus linguistics approach to examine the linguistic style in EL textbooks. The first step involves constructing a self-compiled EL textbook corpus, comprising eight secondary school English as a Foreign Language (EFL) textbooks as samples. These textbooks represent the latest editions of Chinese secondary school EFL textbooks, published by the People's Education Press (PEP) in 2019. Textbook samples include five textbooks for junior secondary school and three for senior secondary school (Compulsory Volumes One to Three). The corpus construction process involves several steps:

(1) Collection and processing of samples: The electronic versions (PDF files) of the sample textbooks were collected and processed using optical scanning to convert them into editable text files (.txt).
(2) Data cleaning: These text files were manually cleaned to remove non-text elements such as images.
(3) Data verification: These files were thoroughly manually checked to identify and correct spelling errors and missing words.
(4) Corpus creation: The finalized text files were imported into corpus concordance software—WordSmith Tools [26], completing the creation of the self-compiled corpus.

The details of the resulting target corpus are presented in Table I.

TABLE I. Composition of Target Textbook Corpus

Target textbook corpus	Textbook samples	Size in tokens (running words)	Size in types (distinct words)
Chinese Secondary School EFL Textbooks Corpus (CSTC)	Five junior secondary school EFL textbooks	107,760	4,221
	Three senior secondary school EFL textbooks	84,166	6,415
Total	Eight EFL textbooks	191,926	7,768

The British National Corpus (BNC) has been selected as the reference corpus to assess the linguistic style of the textbooks, as it represents naturally occurring language and offers register generalizability. Several versions of the BNC have been released.

The original BNC from the 1990s, known as BNC 1994, comprises two subsets: the BNC Sampler and the BNC Baby. The BNC Sampler mirrors the text variety of the complete BNC, selecting 1 million words from each register, making it particularly suitable for studies requiring balanced written and spoken texts. The BNC Baby, on the other hand, is a sample covering four domains—fiction, newspapers, academic writing, and

spontaneous conversation—with each represented by 1 million words. The most recent version, the BNC 2014, serves as the successor to the BNC 1994, aiming to provide a comparable corpus that reflects language changes over the past two decades through modern data collection methods. The BNC 2014 is divided into two parts: Spoken BNC 2014, which is publicly available for download, and Written BNC 2014, accessible since late 2021 via the proprietary LancsBox X software. However, this software has limited search functionalities, making it less suitable for large-scale research. Table II presents a comparative overview across versions.

TABLE II. COMPARISON OF DIFFERENT VERSIONS OF THE BNC

Versions	Release date	Register composition	Words (Approximate)
BNC 1994	Initial release (1995) World Edition (2001) XML Edition (2007)	90% written register 10% spoken register	100 million
BNC Sampler	1997	50% written register 50% spoken register	2 million
BNC Baby	2007	25% fiction 25% newspapers 25% academic writing 25% spontaneous conversation	4 million
Spoken BNC 2014	2017	100% spoken register	11.5 million
Written BNC 2014	2021 (via LancsBox X)	100% written register	100 million

The Written BNC 2014 corpus is unavailable for download and is thus not considered a reference corpus. Considering the size of the target self-constructed corpus (CSTC) (191,926 tokens) and the comparison in Table I, the smaller-sized BNC Sampler (representing a corpus that spans between spoken and written registers), the smaller-sized BNC Baby (removing the 25% spontaneous conversation and representing a written register corpus), and the recently released Spoken BNC 2014 (representing a spoken register corpus) are selected as three reference corpora to serve as benchmarks for comparing the linguistic style of target textbooks.

The corpus concordance software utilized in the present study is WordSmith Tools [26], developed by Mike Scott (available at http://www.lexically.net). The *Wordlist* and *Keywords* functions in this software will be employed to conduct a "*low-effort*" MDA approach, also referred to as "*WordSmith-style keyword analysis*", a term coined by Xiao [27].

IV. RESULT AND DISCUSSION

Based on the works of [11, 13, 28], this study adopts the following steps. The first step is to generate wordlists for all the corpora using the *Wordlist* function in WordSmith Tools. Although the sizes of these corpora vary, the corpus used to create the reference wordlist is relatively unimportant [11]. This claim is further confirmed by a baseline test [13]. In Wordsmith Tools, the default sorting of the wordlist is based on frequency. However, key keywords are more useful than keywords that are sorted solely by frequency, as they exclude those keywords that occur frequently in only a limited number of texts within a specific genre [13]. Xiao employs "*cover %*" (or "*text%*") as the primary sorting criterion, which reflects the percentage of texts in the target corpus in which the word appears, thereby indicating the importance or prevalence of that word within the corpus. Another sorting criterion is "*frequency %*", which represents the percentage of the word's occurrence relative to the total frequency within the target corpus. Using these two sorting criteria, this study identified the top ten key keywords in the target EFL textbook corpus and three alternative reference corpora, prioritizing "*cover %*" followed by "*frequency %*". The detailed results are presented in Table III.

TABLE III. TOP TEN KEY KEYWORDS IN THE TARGET TEXTBOOK CORPUS (CSTC) AND THREE VERSIONS OF REFERENCE CORPORA (BNC)

Number	Target Textbook Corpus (CSTC)	Three Reference Corpora		
		BNC Sampler	BNC Baby (written register version)	Spoken BNC 2014
1	the	the	the	I
2	to	and	of	the
3	and	of	to	and
4	a	to	and	you
5	you	a	a	it
6	in	in	in	a
7	I	it	is	to
8	of	that	that	that
9	is	is	was	like
10	what	for	it	of

As shown in Table III, all three reference corpora share seven key keywords with the target textbook corpus, yet some differences also emerge. For instance, "*I*" and "*you*" appear in both the textbook corpus and Spoken BNC 2014 but are absent from the other two reference corpora, while "*in*" and "*is*" appear in the textbook corpus and the other two reference corpora but are absent from Spoken BNC 2014. These differences suggest that the linguistic style of Chinese EFL textbooks may incorporate both written and spoken registers.

Following this, the second step involves using the *Keywords* function in Wordsmith Tools to generate keyword lists based on the extracted wordlists for the target EFL textbook corpus and all the alternative reference corpora, enabling a "*Significant Consistency Analysis*" [28, p. 58]. The generated keyword lists display Log_L, representing the application of the Log-Likelihood formula for the Log-Likelihood Ratio Test. This value, also called "*Keyness*" [28, p. 59], is critical for our comparison. It measures whether a word in the target corpus occurs with a significantly different frequency compared to the reference corpus. The higher the Log-Likelihood value (LL value), the more significantly the frequency of a word in the target corpus differs, either higher or lower, from that in the reference corpus (i.e., indicating significant usage differences). By ranking the LL values from highest to lowest (i.e., sorting by Keyness), positive keywords are obtained, which

reflect key characteristics of the target corpus. Conversely, sorting from lowest to highest yields negative keywords. LL value is typically combined with *p*-value to determine statistical significance. A low *p*-value (typically < 0.05 or < 0.01) suggests that the frequency difference is significant and unlikely to be the result of random occurrence.

Due to the differing sizes of the three reference corpora, it is not possible to directly compare the number of positive and negative keywords. Therefore, this study employs a normalization method to compare the proportions of positive keywords, mitigating the influence of corpora size. A high proportion of positive keywords indicates that the target corpus exhibits distinctiveness of certain specific words, suggesting that the target corpus is more unique in its lexical usage compared to the reference corpora, with more significant differences. Through calculations, the proportions of positive keywords are as follows: BNC Sampler at 49.15%, BNC Baby at 41.84%, and Spoken BNC 2014 at 72.74%. The proportions of positive keywords in BNC Sampler and BNC Baby are significantly lower than in Spoken BNC 2014, indicating that while Chinese EFL textbooks incorporate features of both spoken and written language, they lean more towards written register.

In conclusion, based on the above comparison of the shared top ten keywords and the "*Significant Consistency Analysis*" across the three keyword lists, the linguistic style of Chinese secondary school EFL textbooks aligns more closely with the written register. Among the three versions of the BNC reference corpora, it shows the greatest similarity to the smaller-sized BNC Baby corpus, which specifically excludes spoken components and represents a 3-million-word written register corpus (i.e., the BNC Baby written register version).

V. Conclusion

By utilizing the *Wordlist* and *Keywords* functions in WordSmith Tools, this study demonstrates the applicability of corpus concordance software in implementing a "*low-effort*" Multidimensional Analysis (MDA) approach to investigate the linguistic style of Chinese secondary EFL textbooks. The findings indicate that the register of the sampled textbooks closely aligns with the written register, as evidenced by comparisons with the reference corpus, BNC Baby (written register version).

These findings hold multiple significant implications:

First, this study provides future educators with valuable insights into the linguistic styles present in Chinese secondary EFL textbooks, laying a foundation for further research in related fields, such as the improvement and optimization of textbook design.

Second, it offers a corpus-based methodological reference for studies requiring preliminary or rapid evaluations of linguistic styles. This approach lowers technical and time barriers, enabling a wider range of researchers to undertake this low-effort linguistic style-related studies.

Additionally, it introduces new perspectives and practical strategies for selecting comparable reference corpora in corpus-based analyses, facilitating more targeted and reliable research outcomes.

Future research could involve comparative analyses by selecting more closely aligned reference corpora that match the stylistic characteristics of the textbooks under study. For instance, selecting comparable native English textbook corpora with similar linguistic features could generate objective data to identify differences in lexical usage, including variations in vocabulary distribution, grammatical structures, and functional expressions. Such insights would contribute to optimizing textbook design and enhancing the alignment of textbook language with authentic, real-world English usage.

Furthermore, the research scope could be expanded to include textbooks from different versions, publishers, or educational stages, examining how variations in linguistic styles influence learners' English language proficiency development.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Mingxuan Zhang was responsible for corpus data collection, data interpretation, and manuscript preparation; Chong Liu contributed to the development of the self-compiled corpus, assisted in the peer review of the corpus, and provided technical support for software applications utilized in the study; Jayakaran Mukundan was responsible for reviewing the manuscript and providing supervision; all authors have read and approved the final version of the manuscript.

Acknowledgment

We would like to express our appreciation for the valuable contributions of all our team members in completing this study.

References

[1] B. Tomlinson, *Developing Materials for Language Teaching*, London: Bloomsbury Publishing, 2013.
[2] B. Tomlinson, *Materials Development in Language Teaching*, 2nd ed., Cambridge, UK: Cambridge University Press, 2011.
[3] M. Zhang, J. Mukundan, L. Khojasteh, and J. A. Jain, "Lexical bundles in English language textbooks: A systematic review," *World Journal of English Language*, vol. 15, no. 6, p. 100, 2025. https://doi.org/10.5430/wjel.v15n6p100
[4] Q. Xu, X. Q. Dong, and Y. Yuan, "The impact of task motivation on Chinese English learners' audiovisual retelling output," *Modern Foreign Languages*, no. 5, pp. 645–658, 2022.
[5] A. Alzahrani, "The structure and function of lexical bundles in communicative Saudi high school EFL textbooks," *International Journal of Applied Linguistics and English Literature*, vol. 9, no. 5, pp. 1–10, 2020.
[6] F. Meunier, "Formulaic language and language teaching," *Annual Review of Applied Linguistics*, vol. 32, pp. 111–129, 2012.
[7] L. Yang and A. Coxhead, "A corpus-based study of vocabulary in the new concept English textbook series," *RELC Journal*, vol. 53, no. 3, pp. 597–611, 2022.
[8] E. Le Foll, "Register variation in school EFL textbooks," *Register Studies*, vol. 3, no. 2, pp. 207–246, 2021.

[9] D. Biber, *Variation Across Speech and Writing*, Cambridge University Press, 1988.
[10] A. Nini, "The multi-dimensional analysis tagger," *Multi-dimensional Analysis: Research Methods and Current Issues*, pp. 67–94, 2019.
[11] C. Tribble, "Writing difficult texts," doctoral dissertation, University of Lancaster, 1999.
[12] M. Scott and C. Tribble, *Textual Patterns: Key Words and Corpus Analysis in Language Education*, Amsterdam: John Benjamins, 2006, vol. 22.
[13] Z. Xiao and A. McEnery, "Two approaches to genre analysis: Three genres in modern American English," *Journal of English Linguistics*, vol. 33, no. 1, pp. 62–82, 2005.
[14] T. Larsson, M. Pacquot, and D. Biber, "On the importance of register in learner writing," *Corpus-Based Approaches to Register Variation*, vol. 103, p. 235, 2021.
[15] T. McEnery and G. Brookes, "Corpus linguistics and the social sciences," *Corpus Linguistics and Linguistic Theory*, 2024. https://doi.org/10.1515/cllt-2024-0036
[16] S. Hunston, *Corpora in Applied Linguistics*, Cambridge, UK: Cambridge University Press, 2022.
[17] F. Farr and A. Leńko-Szymańska, "Corpora in English language teacher education: Research, integration, and resources," *TESOL Quarterly*, vol. 58, no. 3, pp. 1181–1192, 2024.
[18] A. Lusta, Ö. Demirel, and B. Mohammadzadeh, "Language corpus and Data Driven Learning (DDL) in language classrooms: A systematic review," *Heliyon*, 2023.
[19] M. Barlow, "Corpora for theory and practice," *International Journal of Corpus Linguistics*, vol. 1, no. 1, pp. 1–37, 1996.
[20] H. Hoang and P. Crosthwaite, "A comparative analysis of multiword units in the reading and listening input of English textbooks," *System*, vol. 121, 103224, 2024.
[21] H. Wang, "Online corpus construction of English text collection, data cleaning, and similarity analysis," *Mobile Information Systems*, vol. 2022, no. 1, 3105790, 2022.
[22] L. Zhang, Y. Zhang, and R. Cao, "Can we stop cleaning the house and make some food, Mum? A critical investigation of gender representation in China's English textbooks," *Linguistics and Education*, vol. 69, 101058, 2022.
[23] L. X. Li, "Improving modal verb treatment in English textbooks used by Chinese learners: A corpus-based approach," *SAGE Open*, vol. 12, no. 1, 21582440221079918, 2022.
[24] Q. Wang, H. Y. Guo, Z. H. Chen, and X. F. Qian, *Evaluation Study of High School English Curriculum Standard Experimental Textbooks*, Nanning: Guangxi Education Press, 2020.
[25] Y. Hui, "A twenty-year study of English textbooks at home and abroad," *Teaching Research*, vol. 43, no. 4, pp. 38–49, 2020.
[26] M. Scott, *WordSmith Tools (Version 7.0)*, [Computer software], Liverpool, UK: Lexical Analysis Software, 2016.
[27] R. Xiao, "Multidimensional analysis and the study of world Englishes," *World Englishes*, vol. 28, no. 4, pp. 421–450, 2009.
[28] M. Scott, "Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs," in *Small Corpus Studies and ELT: Theory and Practice*, John Benjamins Publishing, 2001, pp. 53–68.