Literature, Language and Cultural Studies
ISSN Print: 3079-5095
ISSN Online: 3079-5109
https://doi.org/10.63313/LLCS.9214

Constructing a Corpus-Driven Academic English Writing Model

Yining Liu. Cui Wang*
Hengxing University, Shandong Province, Qingdao 266110, China

How to cite this paper: Liu, Y., & Wang, C. (2026). Constructing a Corpus-Driven Academic English Writing Model. Literature, Language and Cultural Studies, 6(3), 76–83. ISSN Print: 3079-5095, ISSN Online: 3079-5109.
https://doi.org/10.63313/LLCS.9214
Published: 2026-09-22

Copyright © 2026 by author(s) and Erytis Publishing Limited.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Abstract

Addressing problems in academic English writing instruction such as insufficient authentic input and learners' weak textual awareness, this study constructs a corpus-driven academic English writing model grounded in corpus linguistics and data-driven learning theory. The model consists of three layers—a corpus resource layer, a feature analysis layer, and a writing application layer—and covers four stages: corpus preparation, feature discovery, writing transfer, and comparative feedback. A one-semester teaching experiment was conducted with 64 non-English-major postgraduates. Combining academic chunk statistics with writing quality ratings, the results show that the experimental group outperformed the control group in the diversity of academic chunk use, textual coherence, and academic appropriateness, indicating that the model can effectively enhance learners' academic English writing ability.

Keywords

Corpus; Academic English Writing; Writing Model; Data-Driven Learning; Academic Chunks

1. Introduction

Academic English writing is a core competency for learners to engage in international academic communication at the tertiary level. As universities internationalize, an increasing number of non-English-major postgraduates are required to produce course papers and journal submissions in English; yet their writing commonly suffers from monotonous vocabulary, rigid sentence patterns, and inadequate awareness of textual organization. Traditional instruction relies mostly on model-text analysis and grammar drills, and the linguistic material learners encounter is simplified and filtered, distancing it from authentic academic discourse and making it difficult to develop a systematic perception of academic stylistic features. Corpus linguistics offers a new path for writing instruction, while data-driven learning advocates that learners discover linguistic regularities themselves through concordance searches. In recent years, domestic scholars have explored various applications of corpora in writing teaching, but most have focused

on a single skill or local stage; studies that systematically construct a corpus-driven writing model and provide empirical validation remain limited. Accordingly, this study discusses how such a model can be constructed and whether it can effectively improve writing quality.

2. Theoretical Foundations and Related Research

(1) Theoretical Rationale for the Corpus-Driven Approach

The corpus-driven approach is rooted in respect for the empirical facts of language use. Sinclair's idiom principle holds that language users rely on a large store of prefabricated chunks to express themselves fluently, rather than assembling language from words and grammatical rules on every occasion. Analyses of academic discourse by Biber and his colleagues reveal that four-word-and-longer chunks recur frequently in academic writing, serving functions such as organizing discourse, expressing stance, and referring to research—suggesting that the cultivation of academic writing competence should attend to the holistic acquisition of multi-word units and typical sentence patterns. Data-driven learning (DDL) is the pedagogical application of corpus methods; Johns's DDL model emphasizes that learners induce rules of language use by observing concordance lines. Wei Naixing points out that corpus-based phraseological research provides a shift in vocabulary teaching from isolated lexical items to collocational networks.

(2) Evolution of Research on Academic English Writing

Research on academic English writing has evolved along a path from product-oriented to process-oriented and then to socio-culturally oriented approaches. Swales's move-analysis framework reveals the rhetorical structure of sections such as the introduction, offering an operational analytical tool for writing instruction. Hyland introduced the concept of metadiscourse into academic writing research, arguing that writers establish a negotiated relationship with readers through metadiscoursal resources. Corpus technology has enabled academic writing research to move from qualitative description toward a combination of quantitative and qualitative methods. Based on the Chinese Learner English Corpus, Yang Huizhong and his colleagues identified features of Chinese students' vocabulary, syntax, and textual organization. Ding Yanren and Qi Yan found a significant correlation between chunk use and learners' oral English and writing proficiency. Recent contrastive studies of academic chunks indicate differences between L2 learners and native speakers in the frequency and functional preferences of academic chunk use.

3. Constructing the Corpus-Driven Academic English Writing Model

(1) Basic Principles of Model Construction

The model follows three principles. First is the principle of authenticity:

instructional input is drawn from screened authentic academic texts, ensuring that the linguistic forms learners encounter conform to the expressive norms of the academic community. Second is the principle of induction: instructional activities center on guided observation and discovery, with learners constructing their understanding of academic English by searching corpora, analyzing concordance lines, and generalizing regularities. Third is the principle of personalization: learners may select the appropriate discipline-specific sub-corpus according to their research field, so that the content of learning matches their academic development needs.

(2) Hierarchical Structure of the Model

The model consists of three interrelated layers; its overall architecture is shown in Figure 1.

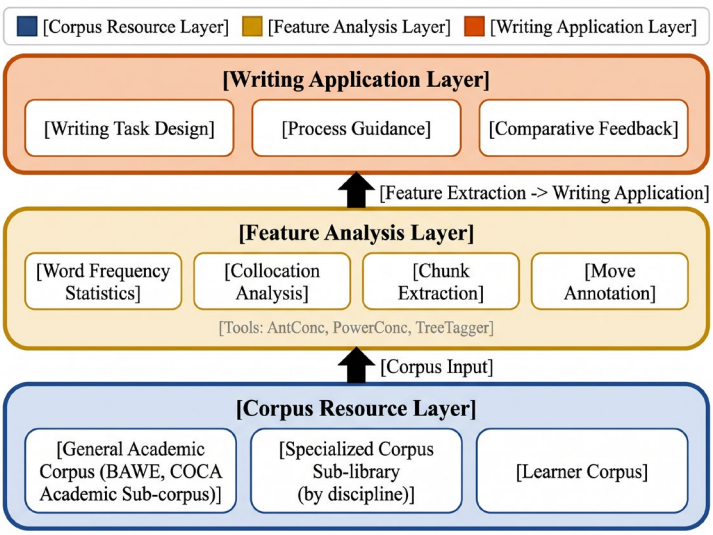

Figure 1. Architecture of the corpus-driven academic English writing model

At the bottom is the corpus resource layer, the data foundation on which the model runs, integrating three types of corpora: a general academic corpus provides cross-disciplinary foundational academic English expressions, for which the British Academic Written English corpus and the academic sub-corpus of the Corpus of Contemporary American English may be used; discipline-specific sub-corpora are built according to learners' fields, covering journal articles and thesis/dissertation abstracts across disciplines; and a learner corpus collects students' writing output for longitudinal tracking and cross-sectional comparison. All three types of corpora undergo cleaning, annotation, and part-of-speech tagging. In the middle is the feature analysis layer, which uses tools such as AntConc and PowerConc to perform four functions—word-frequency statistics, collocation analysis, chunk extraction, and move annotation—helping learners identify high-frequency academic vocabulary, analyze typical collocations, screen academic chunks, and observe the linguistic realization of moves. At the top is the writing application layer, comprising

three modules: writing task design, process guidance, and comparative feedback, which convert linguistic features into practical writing competence through progressive tasks, retrieval prompts, and multi-dimensional comparison.

(3) Operating Workflow of the Model

The model operates in four stages; the core tasks, tools, and outputs of each stage are shown in Table 1.

Table 1. Tasks and tools of each stage in the corpus-driven writing model

Stage	Core Task	Main Tools	Stage Outputs
Corpus preparation and annotation	Screen academic texts, build discipline-specific sub-corpora, complete POS tagging and move annotation	AntConc, TreeTagger, manual annotation	Searchable annotated corpus
Feature discovery and induction	Search high-frequency academic vocabulary, analyze typical collocations, extract academic chunks, observe move features	AntConc, PowerConc, Range	Word lists, collocation lists, chunk lists
Writing transfer and application	Complete writing tasks based on discovered linguistic features; actively search the corpus during writing to verify expressions	Writing platform, online corpus retrieval system	First draft of academic English writing
Comparative feedback and revision	Compare the first draft with reference corpora, identify differences in feature use, revise and refine the text	Text comparison tools, learner corpus	Revised draft and reflection journal

Note: The corpus preparation stage is teacher-led; the feature discovery stage combines teacher demonstration with student operation; and the writing transfer and comparative feedback stages are student-centered. The stages do not proceed in a strictly linear, one-way fashion; during writing, learners may return to the feature discovery stage to re-search the corpus as needed.

1. Corpus Preparation and Annotation

The corpus preparation stage is teacher-led. The teacher selects academic texts published in the past five years from SCI/SSCI journals and outstanding dissertations, cleans them, and imports them into the corpus tool. Annotation includes automatic part-of-speech tagging and manual move annotation: POS tagging is completed with TreeTagger, while move annotation follows the CARS model and is performed independently by two annotators; only when inter-annotator agreement reaches 0.85 or above is the material incorporated into the teaching corpus.

2. Feature Discovery and Induction

The feature discovery stage is the core of the model. The teacher first demonstrates corpus retrieval operations, using high-frequency academic verbs as an example to show how to observe typical collocations and semantic preferences through concordance lines. Learners then search independently around their writing tasks, using the word-list function to extract high-frequency academic vocabulary, the collocation function to analyze significant co-occurrences of node words, and n-gram functions to extract three- and four-word academic chunks, while completing an observation record sheet.

3. Writing Transfer and Application

The writing transfer stage requires learners to apply the linguistic knowledge they have induced to authentic writing tasks. Task design follows a progression from fragments to complete texts: local tasks such as abstract writing and research-method description come first, before moving on to course-paper writing.

During writing, learners are encouraged to use an online corpus retrieval system at any time to verify the appropriateness of lexical collocations and sentence patterns, while the teacher monitors progress through the platform and offers prompts.

4. Comparative Feedback and Revision

The comparative feedback stage promotes reflection and revision through multi-dimensional comparison. The teacher compares learners' first drafts with reference corpora in the same field and generates a feedback report from three perspectives: the density of academic chunk use, the coverage of high-frequency vocabulary, and the completeness of move structure. Guided by the report, learners return to the corpus to retrieve relevant expressions, complete their revisions, and write reflection journals. The feedback emphasizes learners' active participation; the teacher's role is to guide discovery rather than to correct errors directly.

4. Effectiveness Analysis of Model Application

(1) Research Design

A teaching experiment was conducted in an academic English writing course for non-English-major postgraduates at a university. The participants were 64 postgraduates from two intact classes, 32 in the experimental group and 32 in the control group; the two groups showed no significant difference in English proficiency or in a pre-test of writing. The experimental group received instruction with the corpus-driven writing model, while the control group followed the traditional approach of model-text analysis plus writing practice. The experiment lasted 16 weeks, with two class hours per week. Both the pre-test and post-test used an academic abstract writing task, scored by two teachers along three dimensions: linguistic accuracy, textual coherence, and academic appropriateness.

(2) Results and Analysis

1. Changes in Academic Chunk Use

Analysis of the post-test learner corpus shows that the experimental group displayed greater diversity and appropriateness in academic chunk use. Table 2 presents a comparison of the per-capita use frequency of four types of academic chunks between the two groups.

Table 2. Comparison of per-capita frequency of academic chunks between the experimental and control groups

Chunk Type	Experimental (per essay)	Control (per essay)	Significance
Research-oriented chunks	8.6	5.2	$p<0.01$
Text-oriented chunks	6.3	4.1	$p<0.05$
Stance-expressing chunks	4.7	2.9	$p<0.05$
Referential chunks	5.1	4.8	$p>0.05$
Total chunk frequency	24.7	17.0	$p<0.01$

Note: Chunk classification follows Biber and colleagues' functional framework. Research-oriented chunks include the results show, in this study, etc.; text-oriented chunks include it should be noted, as shown in, etc.; stance-expressing chunks include it is suggested that, etc.; and referential chunks include the purpose of this, etc. The statistics are based on 32 post-test essays from each group, using an independent-samples t-test.

As Table 2 shows, the experimental group used research-oriented, text-oriented,

and stance-expressing chunks significantly more frequently than the control group, with the greatest difference in research-oriented chunks. This is consistent with the model's design, which emphasizes training in the extraction and application of academic chunks: by observing high-frequency chunks in authentic corpora and using them deliberately in their writing, learners gradually develop the habit of organizing academic expression with prefabricated chunks. No significant between-group difference appeared for referential chunks, possibly because their forms are relatively fixed and are already addressed in traditional instruction.

2. Changes in Writing Quality Scores

Statistics on the writing test show that the experimental group's mean post-test score was 78.4 versus 71.6 for the control group; an independent-samples t-test indicated that the difference was statistically significant ($p<0.01$). Across dimensions, the experimental group improved most markedly in textual coherence (mean difference = 5.8) and academic appropriateness (mean difference = 6.1), while the difference in linguistic accuracy was relatively small (mean difference = 3.2). The two groups' comparison across dimensions of writing ability is shown in Figure 2.

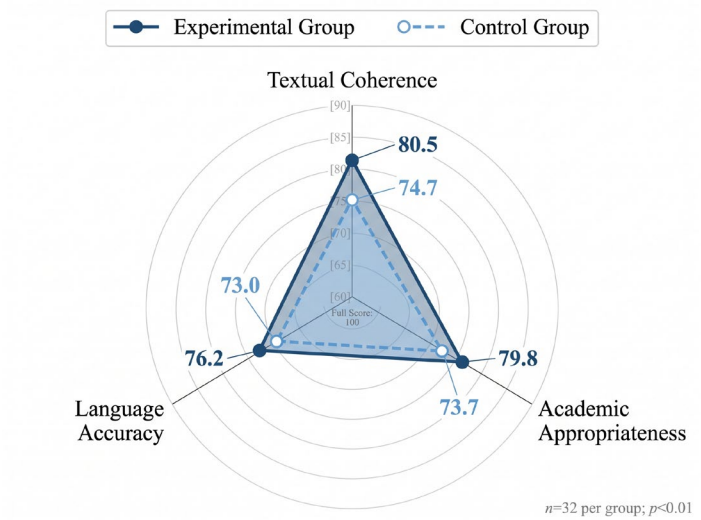

Figure 2. Comparison of scores across dimensions of writing ability between the experimental and control groups

3. Learner Perceptions and Feedback

A questionnaire administered at the end of the course showed that 87.5% of learners considered corpus retrieval helpful in solving lexical collocation problems in writing, and 78.1% reported that observing concordance lines gave them a clearer understanding of the norms of academic English expression. Some learners mentioned that using the corpus felt cumbersome at first, but with more practice they gradually learned to locate the expressions they needed quickly.

5. Discussion

The experimental results show that the corpus-driven academic English writing model has positive effects on enhancing learners' academic chunk use and writing quality, consistent with prior findings that DDL facilitates L2 writing.[16] The three layers of the model play complementary roles: the authentic input provided by the corpus resource layer addresses the lack of authentic material in traditional teaching; the feature analysis layer enables learners to induce regularities of academic expression from data; and the writing application layer promotes the transfer of linguistic knowledge into writing competence through task drive and comparative feedback. Several issues merit attention in implementation: corpus operations present a certain technical threshold, and some learners feel daunted at first, requiring the teacher to arrange adequate tool training in the first few weeks of the course; interpreting corpus retrieval results demands a certain level of linguistic-analytic ability, so the teacher should provide a structured observation framework; and building discipline-specific sub-corpora requires sustained investment, since academic writing varies across disciplines in terminology and discoursal conventions—future work could explore collaboration between discipline teachers and foreign-language teachers in building discipline-specific corpus resources. The model requires teachers to shift from transmitters of linguistic rules to builders of corpus resources, guides of the learning process, and facilitators of data analysis; universities can enhance foreign-language teachers' corpus teaching literacy through training.

6. Conclusion

Focusing on the construction of a corpus-driven academic English writing model, this study, grounded in corpus linguistics and data-driven learning theory, proposes a three-layer architecture—the corpus resource layer, feature analysis layer, and writing application layer—and four operating stages: corpus preparation and annotation, feature discovery and induction, writing transfer and application, and comparative feedback and revision. The teaching experiment shows that the model can significantly improve learners' performance in academic chunk use, textual organization, and academic appropriateness, providing an actionable practical path for reforming academic English writing instruction. The study has limitations: the experiment lasted only one semester, and the participants were non-English-major postgraduates from a single university, so generalizing the conclusions to other groups requires caution. Future research may extend the experimental period, broaden the participant pool, and examine the model's impact on the writing process by incorporating process data.

References

[1] LIU C. Construction and negotiation of hyperthematic and spatial resources in

non-English-major postgraduates' academic English writing[D]. Southeast University, 2025.

[2] TIAN L. Analysis of teaching practice and countermeasures for the academic English writing course of non-English-major undergraduates[J]. English Square, 2022, (24): 105-108.

[3] Johns T.Should you be persuaded:Two examples of data-driven learning[C]//Johns T,King P.Classroom Concordancing.Birmingham:University of Birmingham,1991:1-16.

[4] LIU H Y, LIU Q. Constructing a corpus-based autonomous learning model for academic abstract writing[J]. Journal of University of Shanghai for Science and Technology (Social Science Edition), 2020, 42(4): 301-308.

[5] CHEN B R, LI W M. Teaching research on corpus-driven academic English writing: a case study of abstract writing[J]. Modern English, 2024, (2): 52-54.

[6] WANG H. Exploring a corpus-driven academic English writing teaching model: a case study of abstract writing[J]. Foreign Language Research, 2020, (1): 49-55.

[7] Sinclair J M C.Corpus,Concordance,Collocation[M].Oxford:Oxford University Press,1991.

[8] Biber D,Johansson S,Leech G,et al.Longman Grammar of Spoken and Written English[M].London:Longman,1999.

[9] WEI N X. Corpus-based contrastive phraseology[J]. Journal of Foreign Languages, 2011, 34(4): 32-42.

[10] Swales J M.Genre Analysis:English in Academic and Research Settings[M].Cambridge:Cambridge University Press,1990.

[11] Hyland K.Metadiscourse:Exploring Interaction in Writing[M].London:Continuum,2005.

[12] YU Q L. A study of referential cohesion based on the Chinese Learners' Spoken English Corpus[D]. Xinjiang University, 2013.

[13] YANG X X. A study of the correlation between chunk use and oral English and writing proficiency[J]. English Square, 2019, (4): 58-59.

[14] XU Z. Chunks in academic writing among English L1, L2, and foreign-language university students[D]. Nanjing Normal University, 2012.

[15] Charles M.Corpora and the teaching of academic writing[J].Language Teaching,2018,51(2):217-232.