International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 25, No. 2, 2026

A Corpus-Based Comparative Analysis of L1 and L2 English writers by AntConc Software in China and the UK

Jixian Wang*, Jing Wang

Huzhou Normal University, Huzhou, Zhejiang 313000, China
***Corresponding Author:** Jixian Wang (Email: 01316@huznu.edu.cn)

Abstract: This research intends to analyse the distribution of of the most frequently used words in the writing corpus of L2 writers in China and the UK, which aims to find the different use of vocabulary. The method is to construct the corpus with the discussion section of L2 writers including about 180000 words and then to analyse the data quantitatively by using AntConc software. The result shows that all four functions with regard to stating a goal or purpose, explaining a procedure, stating results or claims, expressing self-benefits and elaborating an argument, were found in the whole corpus and each sub-corpus. There was no difference in the role that the discourse markers play between the three groups and the two groups. But there was a difference in the frequency of SM functions due to the different frequency of SM markers between the three groups and two groups. The study has implications for the education reform of L2 English learning and teaching.

Keywords: Corpus Linguistics, Academic English, L1 and L2 Writers, AntConc Software.

1. Introduction

In an era of accelerating academic globalization, the ability to produce coherent and persuasive written arguments in English has become an indispensable skill for second language (L2) learners worldwide. Lexical competence, a cornerstone of this ability, extends beyond simply knowing words to encompass the understanding of their frequency, collocation, and most importantly, their functional application within specific discourse communities. The investigation of vocabulary use in L2 writing has long been a central focus of applied linguistics, often highlighting the deviations of learner language from native-speaker norms. However, a more nuanced approach, championed by contrastive rhetoric and corpus linguistics, seeks to understand these differences not merely as errors but as potential reflections of distinct rhetorical traditions, learning contexts, and cognitive preferences [1]. This study positions itself within this paradigm, aiming to conduct a comparative analysis of the most frequently used words and their discourse functions in the academic writing of L2 learners from China and the United Kingdom.

Existing research has extensively documented aspects of L2 lexical performance, such as diversity, sophistication, and accuracy [2]. Nevertheless, a significant gap remains in the systematic, cross-national comparison of high-frequency words within a controlled, genre-specific context. High-frequency words, often function words or simple nouns and verbs, are the bedrock of any text. Their strategic use is crucial for achieving key rhetorical moves in academic writing, particularly in the "Discussion" section, where writers must skillfully articulate their research goals, explain procedures, state claims, elaborate arguments, and reflect on the significance of their findings [3]. A mere list of frequent words offers limited insight; it is the examination of the semantic and pragmatic functions these words serve that reveals how L2 writers construct meaning and engage in academic dialogue [4]. Preliminary observations suggest that while L2 learners may possess the requisite vocabulary to perform these functions, their strategic preferences-manifested in the frequency of specific discourse markers-may vary significantly based on their linguistic and educational backgrounds.

To address this gap, this research employs a data-driven, corpus-based methodology. A specialized corpus of approximately 180,000 words is constructed, comprising exclusively the "Discussion" sections of academic papers written by L2 English learners from China and the UK. This genre-focused approach ensures comparability by controlling for the rhetorical demands placed on the writers. The corpus is then analyzed quantitatively using AntConc software [5][6] a powerful tool for corpus linguistics, to generate frequency lists and identify key lexical patterns across the two sub-corpora.

Guided by the research objectives, this study is framed by the following research questions:

(1) What are the most frequent words used in the corpus as a whole, and how are they distributed across the Chinese and UK learner sub-corpora?

(2) How do the discourse functions of these high-frequency words differ between the two learner groups in terms of stating a goal or purpose, explaining a procedure, stating results or claims, and elaborating an argument?

2. Research Method

The method is to construct the corpus with the discussion section of L2 writers including about 180000 words and then to analyse the data quantitatively by using AntConc software. Altogether 21 discussion sections of PhD theses are included in this study, spanning three groups of doctoral writers, L2 writers in China (CHNL2) and L1 and L2 writers in the UK (UKL1 and UKL2). The number of included written texts is based on linguistic presentation of the literature component of PhD theses. The size of the corpus is about 180,000 words and the frequencies were identified by AntConc, which is popular for linguistic analysis.

3. Data Analysis and Discussion

3.1. Provisional Overview of the Whole Corpus

In order to give an overview, the (key) word list was checked before turning to an investigation on the specific uses of self-mentions (SM). The similar feature shown by comparing three-corpora and its high similarity in comparison to Li and Wharton's corpus demonstrate that the data I collected is both representative and reliable [7]. On the one hand, the report reveals that functional words are used similarly in the three groups of writers. For example, the keyword function of AntConc Software displays that three sub-corpora have the same key words and the resemblance of fourteen occurrences can be seen in the top 20 most frequently used keywords listed in the corpus (See Table 1). This list is automatically generated in the corpus showing a list of 20 the most frequently used words. We can see that there are the same thirteen words including *The, Of, And, In, To, That, A, Is, For, As, Their, This* and *With* in each of the sub-corpus. Among them, the top 7 words that are most frequently used are almost at the same level of use; there is only a small difference in rank.

Table 1. Top 20 most frequent words in China-UK Corpus

	CHNL2CORP			UKL1CORP			UKL2CORP		
R	Word	RF	FPW	Word	RF	FPW	Word	RF	FPW
1	The	4921	753	The	3642	565	The	3275	604
2	Of	2455	376	Of	2270	352	And	1943	389
3	And	1929	295	And	1947	302	Of	1836	339
4	In	1741	266	To	1926	299	In	1688	312
5	To	1729	265	In	1679	261	To	1457	269
6	That	918	141	A	1340	208	A	852	157
7	A	828	127	That	948	147	That	727	134
8	Is	790	121	Be	815	126	As	651	120
9	For	594	91	Is	761	118	Learning	580	107
10	Language	557	85	As	730	113	Is	574	106
11	Learners	554	85	This	628	97	For	540	100
12	English	533	82	For	583	91	Their	450	83
13	Their	532	81	It	504	78	Students	411	76
14	As	524	80	Not	442	69	English	393	73
15	Students	501	77	With	431	67	With	393	73
16	L	490	75	They	409	63	They	391	72
17	This	490	75	Their	408	63	Teachers	373	69
18	are	472	72	or	381	59	Be	369	68
19	On	454	69	teacher	369	57	This	359	66
20	With	431	66	Which	368	57	Language	347	64
20	They	428	66.5				not	339	63

Since *language*, *learners* and *English* appeared collectively in the sub-corpus of CHNL2, I further checked the cluster. *English language learners* was not found in this sub-corpus. *Language learners* was identified 36 times, *English learners* 3 times and *English language* occurred 70 times. As a result I viewed *English language* as a cluster. Then, another rank 20 of keywords was added to the CHNL2CORP, which is *They*. Similarly, *English learning* was used 5 times, *English students* 4 times, *English teachers* 31 times, *English language* 103 times in the UKL2CORP. In the same way another rank 20 was added to this sub-corpus. Therefore, I can say that there are 14 of the same words in three groups of English writers with the new addition of *They* in the CHNL2CORP.

In contrast to the 20 most frequently used words there are 15 of the same keywords between Geng and Wharton' s corpus and the CHNL2CORP , and 16 keywords between their corpus and the Corpus of UKL1 and UKL2 respectively [8]. Interestingly, the cluster *of English Language learner(s)* in my corpus is also consistent with Lee and Deakin's (2016) findings [9]. He listed *Language* and *English* as the rank of 11 and 37 respectivley in his sub-corpus of linguistic theses. In addition, the word *learner* appeared as a collocate of the keyword according to his report of the disciplinary content in the *English Language*. The high resemblance between the keyword lists might partially indicates the reliability and representatives of my samplings, which I will return to later. Given the high resemblance of keywords found in my corpus and to that of Li and Wharton's (2012) study, I would state that my data accurately represents texts in doctoral thesis writing.

In addition to a similarity of the distribution of functional words there is a difference in the use of content words between the three groups. That is, in contrast to the UKL1CORP there are more content words in the CHNL2 and UKL2 groups, which might suggest that L1 writers are more confident in using pronouns or reference expressions than L2 writers. CHNL2 and UKL2 writers have 4 of the same content keywords and 13 of the same function keywords. There is just 1 of the same content keywords and 14 of the same function keywords between L1 and L2 writers. My results indicate that fundamentally, all doctoral students apply a similar use of functional words. But there are still variations in using function and content words between L1 and L2 writers. This difference between three groups of writers my study compares shows that there might be factors contributing to this variation in terms of the context and first language background of English writers. For example, UKL1 and UKL2 writes used 17 of the same keywords out of the top 20 word list. CHNL2 and UKL2 writers also employed the amount of 17 of the same keywords if *learner* and *learning* are viewed as the same in semantics. From the overview of the corpus, the different and similar use of language is

identified between different groups of writers. In what to follow, the feature of the self-mention use will be analysed in terms of its use frequency within and between groups.

3.2. Frequency of Discourse Function in the China-UK corpus

Once I identified SM references within the corpus, I turned to examining their discourse function. Given I explored were examined SM features of L1 and L2 writers on the basis of Hyland's model of interactional metadiscourse, I chose to further adopt his classification of rhetorical roles of person markers [2]. Hyland (2002) proposes that there are five rhetorical functions, which include explaining a procedure, stating results or claims, elaborating an argument, stating a goal/purpose and expressing self-benefits [3]. However, upon investigating the corpus, I did not find any instances that function as an expression of self-benefits. This finding is consistent with Hyland's claim "this is a function which does not occur in the professional research texts" and nor in other studies that have been conducted on SM reference. As a result, I chose to focus on the initial four discourse functions of SM references that Hyland defines, and these considered through raw and mean frequencies. These are listed below in Table 2. The first two sub-categories represent low risk function, and the other two sub-categories indicate high risk function, again according to him.

Table 2. Discourse functions of SM in the China-UK corpus

Function		CHNL2 CORP	UKL1 CORP	UKL2 CORP	Total	Per (%)	FPW
Low risk function	Stating a goal/purpose	11	29	13	53	8.73	2.88
	Explaining a procedure	12	17	39	68	11.20	3.70
High risk function	Stating results /claims	38	184	182	404	66.56	21.97
	Elaborating an argument	12	45	25	82	13.51	4.46

The various roles that person markers play are different throughout the corpus; on the whole, there is a greater usage of person markers that serve the high risk function as opposed to low risk function. What the data shows is that 66.56% of SM markers are used the most frequently to state results or claims, followed by those, which elaborate an argument (13.51%), explain a procedure (11.20%) or state a goal or purpose (8.73%). The writers in all three groups used SM markers most frequently to state results or claims and this is followed by the function of elaborating an argument. In addition, there are different frequencies between the three groups in relation to each discourse functions as shown in Table 3. It may indicate that there is a different use of SM markers, which serve different discourse functions between three groups of writers.

Table 3. Mean frequency of discourse functions of SM in the corpus

Function	CHNL2CORP		UKL1CORP		UKL2CORP	
	RF	FPW	RF	FPW	RF	FPW
Stating a goal/purpose	11	1.68	29	4.50	13	2.40
Explaining a procedure	12	1.84	17	2.64	39	7.20
Stating results/claims	38	5.82	184	28.57	182	33.59
Elaborating an argument	12	1.84	45	6.99	25	4.61

With regards to the function of individual markers, it is shown that person markers that account for the discourse function varies in each group. For example, in the CHNL2CORP group, *We* is the only marker to realise all of the discourse functions of SM references. This is followed by *TR*, which covers two functions. The rest of the three markers play only one role to state results or claims in this sub-corpus: *I* is employed by UKL1 writers such that it crosses all four discourse roles in its use of SM references. *My* and *We* are used for two functions, and the rest of *Me*, *Our* and *Us* are used with one function only. For UKL2 writers, both *I* and *We* are found to again engage the four roles of discourse function described by Hyland, with *My* playing two roles and *Our*, *Me*, *Mine* and *TR* one role only. The detailed difference in the frequency of SM function between the two groups will be addressed in the following part.

4. Conclusion

The preliminary findings indicate a complex picture of convergence and divergence. Notably, all four targeted discourse functions were identified in both the overall corpus and each individual sub-corpus, suggesting that learners from both backgrounds possess the fundamental ability to execute essential rhetorical moves in academic writing. Crucially, the study found that the core role of discourse markers in facilitating these functions was consistent across groups; however, a significant difference was observed in the frequency with which specific markers were employed to realize these functions. This distinction points not to a deficit in knowledge but to a divergence in rhetorical style and preference.

The implications of this research are primarily pedagogical. By pinpointing the exact nature of lexical and functional preferences between different L2 learner groups, the findings can inform the development of more tailored, effective teaching materials and pedagogical strategies. This moves L2 writing instruction beyond a one-size-fits-all model towards a more responsive approach that acknowledges and incorporates the nuanced realities of learner language, ultimately empowering students to become more versatile and effective academic writers.

Acknowledgement

Sponsored by Zhejiang Provincial Association of Foreign Languages 2023 Special Research Project (ZWYB2023012)：Identity Construction in Academic English Writing；Teaching

Reform Research Project of Huzhou Normal University (JG202512).

References

[1] Hyland, K. (2002). Authority and invisibility: Authorial identity in academic writing. *Journal of Pragmatics*, 34(8), 1091–1112. https://doi.org/10.1016/S0378-2166(02)00035-8

[2] Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies*, 7(2), 173–192. https://doi.org/10.1177/1461445605050365

[3] Hyland, K. (2017). Metadiscourse: What is it and where is it going? *Journal of Pragmatics*, 113, 16–29. https://doi.org/10.1016/j.pragma.2017.03.007

[4] Chen, X. (2020). New proposal for metadiscourse classification based on metapragmatics. *Foreign Languages and Foreign Language Teaching*, 10(8), 50–68.

[5] Anthony, L. (2024). Addressing the challenges of data-driven learning through corpus tool design – in conversation with Laurence Anthony. In P. Crosthwaite (Ed.), *Corpora for language learning*. Routledge.

[6] Anthony, L. (2016). Introducing corpora and corpus tools into the technical writing classroom through Data-Driven Learning (DDL). In J. Flowerdew & T. Costley (Eds.), *Discipline specific writing* (pp. 162–180).

[7] Li, T., & Wharton, S. (2012). Metadiscourse repertoire of L1 Mandarin undergraduates writing in English: A cross-contextual, cross-disciplinary study. *Journal of English for Academic Purposes*, 11(4), 345–356. https://doi.org/10.1016/j.jeap.2012.07.004

[8] Geng, Y., & Wharton, S. (2016). Evaluative language in discussion sections of doctoral theses: Similarities and differences between L1 Chinese and L1 English writers. *Journal of English for Academic Purposes*, 22, 80–91. https://doi.org/10.1016/j.jeap.2016.02.004

[9] Lee, J. J., & Deakin, L. (2016). Interactions in L1 and L2 undergraduate student writing: Interactional metadiscourse in successful and less-successful argumentative essays. *Journal of Second Language Writing*, 33, 21–34. https://doi.org/10.1016/j.jslw.2016.06.002