ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 14, No. 1, pp. 37-47, January 2023
DOI: https://doi.org/10.17507/jltr.1401.05

A Multi-Dimensional Analysis of English Writings by Chinese EFL Learners

Xiaoyun Li
Department of Theoretical Linguistics, Doctoral School of Linguistics, University of Szeged, Hungary

***Abstract*—The exponential growth in the population of Chinese EFL learners has fueled the study of Chinese EFL learner writing. A survey of relevant literature indicates that the majority of studies are confined to the exploration of individual linguistic features, with a few exceptions which employ a broader perspective that might involve multiple features. This work aims to investigate the English writings by Chinese EFL learners via Multi-Dimensional (MD) analysis, a corpus-based approach that combines both microscopic (i.e., individual linguistic features) and macroscopic perspectives (i.e., textual dimensions). A comparison between writings by Chinese EFL learners and native English speakers shows that the former are high on involvement, informativeness, and referential explicitness while the latter exhibit superiority on word-choosing, information integration, narrativity, and persuasiveness. Regarding their specific use of 67 MD linguistic features, the two writer groups also show certain significant but interesting differences. Analysis of Chinese EFL learner corpora from different English education levels indicates that writings by learners from higher levels are lower on involvement, but are higher on informativeness, narrativity, referential explicitness, and persuasiveness. This trend is manifested by their decreasing use of involvement features, but increasing use of features marking the latter four aspects.**

***Index Terms*—multi-dimensional analysis, Chinese EEL learners, native English speakers, English writing**

I. Introduction

Writing is commonly regarded as an outcome, a finished product that can tell “us about language” rather than “about meaning-making” (Hyland, 2016, p. 4). Among the measures designed to evaluate learners’ language proficiency, writing stands out for being an indispensable role in language proficiency tests since it involves numerous elements of language, such as vocabulary, grammar, and style. In addition, compared with speaking, writing provides more convenient and direct access to real language materials for researchers or teachers attempting to explain how languages are acquired by learners. Conducting a writing study, therefore, has long been a centerpiece in the language acquisition study.

As one of the largest learner groups of English as a foreign language (EFL) in the world, Chinese EFL learners, along with their writings in English, have garnered enormous attention from researchers. Of the various approaches taken, Corpus Linguistics (CL) has gained in popularity owing to its empirical nature and heavy reliance on “both quantitative and qualitative techniques” (Biber et al., 1998, p. 4).

Nevertheless, scrutiny of the relevant literature reveals that most of the studies have adopted a microscopic angle, or more specifically, focused on individual or a limited number of linguistic features, whereas analysis of textual features, which might involve a broader range of linguistic features, is sparse. Admittedly, a microscopic angle may allow researchers to better manipulate their research (in terms of scope, feasibility, space, etc.) and thus achieve an in-depth investigation. This approach, however, can be problematic as it treats linguistic features in isolation, whereas in real language production, linguistic features, more often than not, co-occur due to their internal connections (Biber, 1995).

This study is dedicated to exploring the English writings by Chinese EFL learners via Biber’s Multi-Dimensional (MD) analysis, an approach that the present study believes to be able to mitigate the paucity mentioned above since it combines macroscopic (i.e., textual dimension) and microscopic (individual linguistic feature) perspectives (Friginal & Hardy, 2014). Following previous studies, this study firstly intends to know the differences between Chinese EFL learners and native English speakers in English writing. It is hoped that by comparing their dimension scores and detailed use of linguistic features, the macroscopic and microscopic differences between them can be determined. In addition, this study also wants to know the differences in writing between Chinese EFL learners from different levels of English education. To this end, comparisons in terms of dimension scores and individual linguistic feature usage between the three groups of Chinese EFL learners, viz, Chinese English majors, Chinese non-English major undergraduates, and high school students are to be performed.

II. Previous Studies on Chinese EFL Learner Writing

The past decades after the opening-up of China have witnessed a remarkable increase in the population of Chinese EFL learners and a consequent surge of research interests in “Chinese English”. To date, the Chinese EFL learner writing study has covered various aspects of writing and has been fruitful.

Similar to the EFL writing studies from other language backgrounds, plenty of Chinese EFL learner writing studies endeavor to investigate the influence of the Chinese language and culture on the writings of Chinese EFL learners with the aim of providing pedagogical implications for the improvement of writing instruction. A noteworthy study conducted by Zhang (2013) claims that the transfer of Chinese in Chinese EFL learners' writings in terms of lexicon, grammar, and grammatical metaphors can be attributed to the similarities and differences between Chinese and English in conceptualization and categorization. He further concludes that the transfers of Chinese are systematic, regular, and unique as compared with Austrian EFL learners whose writings are similar to native speakers. Among the rich literature of Chinese EFL learner writing study, the exploration of the unique features of Chinese EFL learner writing has occupied the central place. At the lexical level, the characteristics of Chinese EFL learner writing on the utilization of linguistic features such as phrasal verbs (Chen, 2013), adverbial conjuncts (Xu & Liu, 2012), collocation use (Chen, 2019), lexical bundles (Bychkovska & Lee, 2017), and cohesive device (Yang & Sun, 2012) have been explored exhaustively. At the syntactic level, Chinese EFL learners and native speakers exhibit different preferences regarding the use of inanimate subjects (Ji & Liang, 2015), non-finite clauses (Fang, 2014) and syntactic complexity (Ai & Lu, 2013), among others. With regard to the discourse features, Wen et al. (2003) confirm that Chinese EFL learners display an apparent oral style in writing. Liang (2008) points out that Chinese EFL learners' overuse of surface features such as pronouns and connectives does not contribute much to the coherence of their English writings. Much effort has also been made to examine the frequent errors made by Chinese EFL learners in compositions. It is shown that errors tend to cluster around the misuse of tense, determiners, prepositions, collocations, etc. (Ong, 2011; Liu, 2012; Darus & Subramaniam, 2009; Liu & Lu, 2020; Yang et al., 2013).

A prominent problem pertinent to Chinese EFL learner writing research is that most of it is limited to a microscopic perspective (i.e., focusing on individual linguistic features) while in texts linguistic features are more likely to work together (Biber et al, 2002), some researchers thus turn to macroscopic analysis to overcome the issues brought by microscopic analysis (e.g., ignoring the connections between individual linguistic features). Against this background, MD analysis, an approach that combines microscopic and macroscopic analyses and is originally proposed for analyzing register variation, is introduced into the study of Chinese EFL learner writing. Ma (2002) discovers that Chinese and American undergraduates show significant differences on nine out of 66 linguistic features examined in Biber's (1988) MD analysis. She states that writings of Chinese undergraduates appear to be more informational and formal, while American undergraduates show more features of involvement, personal affection, and postpositional elaboration of nominal information. Pan (2012) uses MD analysis to compare the writings by Chinese non-English major undergraduates and graduates with those by native English speakers (i.e., LOCNESS or The Louvain Corpus of Native English Essays). The results indicate that writings by the two groups of Chinese EFL learners are highly interactive and persuasive while those by their American counterparts are high in informative, narrative, explicit and abstract dimensions. It is revealed that there are different degrees of deviations between the native and non-native speakers when using function-related language features. Besides, internal variations are also observed inside the Chinese student group she examined.

Despite using the same MD analysis method and focusing on the same learner group, this study sets itself apart from these previous two studies in two main aspects. One is that the Chinese EFL learner group the current study targets is expanded. Unlike the two mentioned studies which pay special heed to college students, this study, in addition to college students (non-English major college students), also includes Chinese EFL learners who are at a lower English education level (i.e., high school students), as well as English majors who are trained for career pathways or academic purposes. The inclusion of the latter two Chinese learner groups is made to observe their writing developments or changes across different levels of English education. The second aspect lies in that the corpora selected in the present study are more balanced (in terms of genre, length, corpus size, etc.) and thus more representative. The corpora the two mentioned studies selected seem not comparable enough. Their results, hence, may face validity issues. Taking Pan (2012) as an example, the reference corpus and the Chinese learner corpora it employed vary greatly as far as essay length (800 words in LOCNESS versus 213 words in Chinese learner corpus) and genre (academic essays in LOCNESS versus argumentative writings in Chinese learner corpus) are concerned. This study, as we will see in the following section, is conducted on language data (both native and non-native data) that are more strictly controlled.

III. Methodology

A. *Multi-Dimensional (MD) Analysis*

MD analysis derives from Biber's (1988) exploration of the variation between written and spoken registers. Based on the previous linguistic finding that some features co-occur in a relatively fixed pattern in a particular register to perform certain functions, Biber (1988) proposed the notion of MD analysis and conducted an MD analysis on the written and spoken registers extracted from the Lancaster-Oslo-Bergen Corpus (LOB) and London-Lund Corpus via factor analysis. In his study, Biber identified the following six dimensions of variation of spoken and written registers.

Dimension1: Involved versus informational production.
Dimension2: Narrative versus non-narrative concerns.
Dimension3: Explicit versus situation-dependent reference.
Dimension4: Overt expression of persuasion.

Dimension5: Abstract versus non-abstract.

Dimension6: Online production.

Every dimension (or factor) is determined and named through a functional interpretation of the linguistic features it contains. Of special note is that no single dimension is capable of profiling a register as some registers might be different from each other on one dimension but highly similar on another. *Official documents* and *broadcasts*, for example, resemble each other on Dimension 2 for both of them are not strictly narrative, but differ on Dimension 3 since the former is explicitly reference-concerned while the latter is situation-dependent. In the present study, the first five dimensions were applied.

To date, there have been generally two ways to conduct an MD analysis. One is to carry out a new factor analysis (also called full MD analysis) as Biber did in the late 1980s. Researchers firstly calculate the normalized frequencies of numerous linguistic features and then conduct a factor analysis on those frequency values to reach several new factors (or dimensions in MD analysis). New factors are then interpreted and named by researchers through qualitative interpretations of corresponding features. This way is widely used for identifying the co-occurrence patterns of specialized or newly born registers, for example, research article abstracts (Cao & Xiao, 2013), non-western languages including Portuguese (Sardinha et.al, 2014) and Chinese (Zhang, 2012), and Pop Songs (Dutra, 2014), among others. The other approach is called additive MD analysis. This approach applies Biber's (1988) dimensions directly and hence omits the factor extraction. Studies using additive MD analysis calculate the dimension scores of a text based on the mean frequencies for variables (i.e., linguistic features) provided in Biber (1988) [1] and then contrast the obtained dimension scores to the scores of registers in Biber's (1988) dimension scale to evaluate how the text under discussion is related to the wide variety of discourses investigated in Biber (1988). In discussing the application of additive MD analysis, Nini (2019, p. 91) argues that it sheds light on "the identity", "the location" and "the linguistic peculiarities" of a corpus.

The present study applied the latter method as it is comparatively less demanding for researchers to conduct an MD analysis in comparison with a full MD analysis due to the omissions of factor analysis and interpretation. Another benefit brought by this method is that other discourses that are investigated in the same way, along with the discourses in Biber's (1988) dimension scale, can provide rich reference information for a more fine-grained analysis. What makes this approach more feasible is a highly reliable tool called MAT (Multi-dimensional Analysis Tagger), a versatile piece of software which is designed by Nini (2013) specifically for conducting such an MD analysis. With this tool, researchers can be saved from tedious MD part-of-speech (POS) tagging and dimension score calculation. In this study, the newest version of MAT, MAT 1.3.2, was adopted.

B. Selection of Corpus Data

The language data explored in this study were sampled from three well-known existing corpora, namely, TECCL (Ten-thousand English Compositions of Chinese Learners), WECCL (Written English Corpus of Chinese Learners), and the Written Essay Module of ICNALE (International Corpus Network of Asian Learners of English).

TECCL, as its name implies, contains about 10,000 writings that are written by Chinese EFL learners at different education levels (Xue, 2015). According to the project initiator, Xu (2016), this corpus is representative in that it collects updated materials (writings completed between 2011 to 2015), features a wide range of topics or prompts (over 1000 different topics), corresponds to the actual proportion of universities of China, covers by far the widest spread of Chinese EFL learners (learners from mainland China, Hong Kong, and Taiwan), and involves diversified writing tasks(in class, after class, timed, and untimed). In this study, the Chinese High School Student (CHSS) writing corpus and the Chinese Non-English Major Undergraduate (CNEMU) writing corpus, were separately established by arbitrarily choosing corresponding argumentative writings from TECCL. The detailed information on these two sub-corpora can be seen in Table 1.

WECCL 2.0 is a sub-corpus of SWEECCL (Spoken and Written English Corpus of Chinese Learners) (Wen et al., 2005). This corpus gathers English writings by English majors in different grades from nine different Chinese universities. It takes into consideration the possible factors that might exert influences on learners' writings and controls context variables including time (timed vs. untimed), writing types (argumentative, narrative, and expository writing), length of writing (200 to 800 words), writing proficiency (Grades 1- 4) and student enrolling year. In total, 3578 writings were included. This study established a Chinese English Major (CEM) writing corpus by randomly extracting 50 timed and 50 untimed argumentative writings from WECCL2.0 (See Table 1). In addition, the selection threshold regarding the essay length was set between 200 and 300 words.

ICNALE is a corpus constructed for conducting contrastive interlanguage analyses on Asian learners of English (Ishikawa, 2013). This corpus is described as reliable as it encompasses the language production by Asian learners of English that are strictly controlled in terms of topics (2 topics), time (20 to 40 minutes), length (200 – 300 words), and reference use, etc (Ishikawa, 2013). The Written Essay Module of ICNALE contains writings by Asian learners of English from 12 countries or regions, as well as by native English speakers from 5 English-speaking countries. Currently, it contains approximately 5600 writings, totaling 1.3 million words. In this study, a reference corpus called

[1] For a more detailed description, please see Van Vooy (2008, p. 276).

Native English Speaker writing corpus (NES) was built by randomly selecting 100 writings by experienced L1 English writers from ICNALE (See table 1).

TABLE 1
CORPORA

Corpus	*Number of Writings*	*Number of Words*
Chinese High School Student writing corpus (CHSS)	100	15375
Chinese Non-English Major Undergraduate writing corpus (CNEMU)	100	21986
Chinese English Major writing corpus (CEM)	100	26360
Native English Speaker writing corpus (NES)	100	22060
Total	400	85781

IV. RESULTS AND INTERPRETATION

A. Dimension 1

In Biber's 1988 MD analysis, Dimension 1 is the most fundamental parameter accounting for the variation across spoken and written registers. As a result, it contains the most linguistic features compared with other dimensions, with twenty-four positive loading features and five negative loading features included. As its name suggests, this dimension covers a continuum ranging from a focus on involved production to a focus on informational production, with the former occupying the positive pole and the latter taking the negative pole. The positive pole of Dimension 1 represents discourses that are "interactional, affective, involved purposes, associated with strict real-time production and comprehension constraints" (Biber, 1995, p. 115), for example, *telephone conversation*. Discourses at this pole often contain high frequencies of typical spoken language features that bear positive loadings, such as personal pronouns, present tense, and private verbs. As for the discourses distributed at the negative pole, they are highly informational and contain high frequencies of informational features that are with negative loadings, including nouns, adjectives, and prepositions. A typical example is academic papers.

In Table 2, all the corpora obtain positive dimension scores and therefore exhibit an involved and interpersonal focus. Among the three groups of Chinese learners of English, CHSS receives the highest score (8.15) and thus demonstrates a distinctive involved writing style. It is close to the dimension score of *email* (8.7), an internet genre (including both business messages and personal exchanges) investigated by Sardinha (2014). The other two Chinese learner corpora, CNMCS and CEM, however, score considerably lower (4.14 & 4.10). In Biber's (1988) scale, they resemble *romantic fiction* (4.3) – a genre that can be categorized into written registers. It is surprising to see that ENS (3.6) scores marginally less than their CNMCS and CEM, indicating that experienced L1 writers do not avoid involvement in writing.

Table 2 also presents the normalized (per 1000 words) counts of linguistic features of each corpus on Dimension 1, through which we can reach a microscopic understanding of the four corpora. Overall, native English speakers and Chinese EFL learners show different preferences towards the common linguistic features loaded on Dimension 1, including PRIV, THATD, CONT, VPRT, SPP2, PROD, EMPH, FPP1, CAUS, POMD, ANDC, STPR, NN, PIN, TTR, and JJ.

As regards the positive loading features, NES has a lower frequency of use of features that are considered as informal from the perspective of formal writing, including THATD, CONT, and SPP2. Besides this, native speakers seem to refrain from using the features that mark the involvement of the author/speaker, namely, PRIV and FPP1. Native English speakers also make less use of features that are frequently found in conversations or interactive discourse according to Biber (1988), including VPRT, PROD, EMPH, POMD, and ANDC. There is, however, one feature that is more frequently used by native English speakers than by Chinese learners of English: CAUS (because). Rather than the affect from spoken register/genre, this may be due to the transfer of Chinese, as Chinese relies "on semantic or logical comprehension rather than connectives in the juxtaposition of syntactic units" (Tse, 2010, p. 351). In other words, owing to the transfer from Chinese, Chinese learners are apt to omit the causal connector *because* in writing though it has a Chinese equivalent that is commonly used.

For the negative loading features which are used to mark informational production, it is surprising to see that NES surpasses its Chinese counterparts in terms of PIN and TTR counts while lagging behind them on NN and JJ, two features that form the basis of informativeness. The type/token ratio is associated with precise word choice (since it marks vocabulary diversity) and prepositions are an important device for information packing (Biber, 1988). Therefore, it can be concluded that writings by native English speakers feature high precision in word choice and high integration of information. In MD analysis, nouns and adjectives mark information richness. The Chinese learner corpora thus illustrate a higher degree of informativeness than the native English data, though their low type/token ratio reflects that the writings contained are short on vocabulary diversity. In addition, the low frequency of prepositions in the three written English corpora of Chinese learners seems to imply that the information in their writings is organized in a loose way. It should be noted, however, that the underuse of prepositions is perceived as universal to non-native English learners (Gilquin & Granger, 2011) since prepositions are considered to be "among the most difficult forms" that non-native speakers "have to master in learning the English language" (O' Dowd, 1998, p. 6). Overall, writings by Chinese learners of English can be characterized by high informativeness, high vocabulary repetition, and loose information structures.

There are differences between the groups of Chinese EFL learners too. From CHSS to CEM, a pronounced decreasing tendency can be observed on the counts of several positive loading features including CONT, SPP2, EMPH, FPP1, PIT, BEMA, and POMD, indicating that involvement decreases as learners' English education level advances. As for negative loading features, we can see a trend that is in stark contrast to the above tendency. Table 2 shows that there is an increasing trend from CHSS to CEM in terms of AWL, PIN, and TTR, illustrating that Chinese learners' writings appear to become more informational as their English education level rises.

TABLE 2
DIMENSION 1

	NES	CEM	CNEMU	CHSS
Dimension Score	3.6	4.10	4.14	8.15
Loading features (Positive)				
Private verbs (PRIV)	19.58	23.7	17.4	21.1
Subordinator that deletion (THATD)	2.8	7.6	2.9	4.2
Contractions (CONT)	5.8	5.8	7.8	9.7
Present tense verb (VPRT)	74.6	79.1	77.2	80.1
2nd person pronouns (SPP2)	1.9	8.8	8.1	14.5
Do as pro-verb (PROD)	2.1	2.4	2.4	3.1
Analytic negation (XX0)	12.7	12.4	12.8	11.1
Demonstrative pronouns (DEMP)	7.0	4.2	3.4	4.1
Emphatics (EMPH)	11.5	15.5	16.1	20.6
1st person pronouns (FPP1)	27.6	35.8	45.6	52.4
Pronoun it (PIT)	16.4	12.9	15.1	17.4
Be as main verb (BEMA)	24.2	21.5	25.2	26.3
Causative adverbial (CAUS)	3.8	2.3	2.2	2.1
Discourse particles (DPAR)	0.3	0.1	0.3	0.5
Indefinite pronouns (INPR)	0.5	0.5	0.4	1.0
Hedges (HDG)	0.1	0.8	0.4	0.3
Amplifiers (AMP)	3.0	2.8	2.5	2.7
Sentence relatives (SERE)	0.1	0.9	1.2	0.9
Direct WH-questions (WHQU)	0.5	0.6	1.2	1.2
Possibility modals (POMD)	9.8	13.8	15.4	17.9
Independent clause coordination (ANDC)	3.8	5.8	5.2	4.1
Wh-clauses (WHCL)	0.4	1.2	0.5	0.5
Stranded preposition（STPR）	1.5	1.0	1.1	1.3
Loading Features (Negative)				
Nouns (NN)	180.6	201.9	205.7	207
Average Word Length (AWL)	4.6	4.7	4.6	4.4
Prepositions (PIN)	96.9	86.6	84	81.5
Type/token ratio (TTR)	131.0	126.2	119.0	96.3
Attributive adjectives (JJ)	61.6	64.9	72.3	64.5

From the perspective of language exposure, this is not necessarily surprising. Alhusban & Vijayakumar (2021), in their exploration of lexical bundles in Saudi EFL student writings, argue that language exposure and context knowledge (for example, register knowledge) play a determining role in learners' use of lexical resources. Being learners at higher English education levels, Chinese college students and English majors are more likely to outdo Chinese high school students with regard to exposure to written English given that writing is indispensable in their English learning. It is likely that they might acquire relatively more register/genre knowledge and accordingly choose the usage appropriate to the context (in this case, formal writing). Similarly, Gilquin and Paquot (2008) link the frequent "spoken like" features, especially involvement features, in novice L1 and L2 writings with writers' limited acquisition of the rules of academic writing and "lack of knowledge of more formal alternatives" (p. 56) and point out that developmental factor plays a key role. Moreover, learners' English proficiency may provide another explanation as advanced learners apparently have more linguistic resources at their disposal compared with less advanced learners and thus have fewer difficulties in word-finding and grammar. This superiority not only enables advanced learners to avoid those involved features that are commonly used and early acquired (for example, pronouns) without damaging the writing quality, but also provides a basis for them to include more complicated and more diversified vocabulary in writing.

B. Dimension 2

Differing from Dimension 1, Dimension 2 in Biber (1988) deals solely with narrative concerns (Biber, 1988). The features it contains hence are all with positive loadings and are serving narrative purposes. Discourses with high scores in this dimension, such as *fiction*, feature an abundance of narrative features, thus reflecting a narrative concern. Those with low scores, for instance, *expositive writing*, are often characterized as non-narrative due to the sparsity of narrative features.

TABLE 3
DIMENSION 2

	NES	CEM	CNEMU	CHSS
Dimension Score	-1.58	-2.25	-2.5	-3.45
Loading Features (Positive)				
Past tense verbs (VBD)	8.6	7.6	9.8	14.5
3rd personal pronouns (TPP3)	30.8	31.1	19.8	19.5
Perfect aspects (PEAS)	4.7	3.2	3.1	2.3
Public verb (PUBV)	5.7	4.4	4.5	4.9
Synthetic negation (SYNE)	1.7	2.4	2.4	1.4
Past participial clauses (PASTP)	0.1	0.3	0.1	0.1

As can be seen from Table 3, all the four corpora are located at the negative pole of Dimension 2, indicating that they are mutually similar in having a low narrative concern. This is not surprising considering that all the writings investigated are, by their nature, argumentative instead of narrative. Nevertheless, what is interesting about the distribution of the four corpora is that the native speaker corpus scores higher than all Chinese learner corpora. In Biber's (1988) dimension scale, it is close to *press reviews* (-1.6), a genre that is particularly opinionated. This is in line with Pan (2012). It therefore can be inferred that Chinese learners behave poorly in employing narrative devices to increase the persuasiveness of their writings. Moreover, a gently rising narrative tendency can be observed inside the three Chinese learner groups. CHSS's score (-3.45) is the lowest and shows a minor difference to *broadcasts* (*-3.3*) - the most non-narrative genre discovered by Biber (1988); CNEMU obtains a higher score (-2.5) and thus resembles the genre of *professional letters* (-2.6); CEM is the most narrative Chinese learner corpus (-2.25) and is close to the genre of *academic prose* (-2.2) in Biber's scale. This seems to suggest that learners with higher proficiency are more capable of utilizing narrative devices to achieve a persuasive purpose.

The normalized frequencies presented in Table 3 explain the distributions of the four corpora. On the counts of PEAS and PUBV, NES is clearly higher than the three Chinese learner corpora while on another major feature, TPP3, it is negligibly lower than CEM but much higher than CNEMU and CHSS. The explanation for the low frequency of PEAS in Chinese learner corpora might be the complexity of the tense-aspect system in English. In English, aspects are often interrelated with tenses, which poses great difficulty to Chinese learners whose native language is non-inflected and tenseless (Chou & Wu, 2007). As a result, Chinese learners might hold a cautious attitude in using the perfect aspects.

For the high frequency of PUBV in NES, this study tends to attribute it to native speakers' inclination to use public verbs, especially *agree*. The concordances of PUBV reveal that more than half of the PUBVs in ENS are used to state central claims (e.g., *I agree that*, *I say*, etc.) whereas in Chinese learner corpora, similar claims are expressed with private verbs (e.g., *I think*, *I believe*, etc.). Interesting findings can also be observed between the three Chinese learner corpora. CEM and CNEMU show a minor difference in terms of their dimension scores but vary significantly on the counts of VBDs and TPP3: CNEMU has a higher rate of VBD use while CEM has more frequent use of TPP3. The reason might be CME's low frequency of FPP1 but high frequency of VPRT (see Table 2) in comparison to CNEMU. For CHSS, although it contains the greatest amount of VBDs among the four investigated corpora, its rare TPP3, PEAS, and SYNE significantly reduce the narrativity and thus contribute to a salient non-narrative style. The reason, except for CHSS's frequent use of FPP1 (see Table 2), might be again the complexity of PEAS and SYNE (compared to the analytic negation "*not*").

C. Dimension 3

Dimension 3 in Biber (1988) is associated with reference making in a text. The two poles of it respectively represent two converse manners of marking referents: explicit versus situation-dependent reference. Accordingly, the linguistic features contained in Dimension 3 can be divided into negative and positive groups as well according to their factor loadings. Positive linguistic features in this dimension contribute to the referential explicitness of a text, whereas features with negative loadings serve to "mark the physical and temporal situation" (Biber, 1988, p. 144). Discourses with high scores on this dimension are explicit in referents, for example, *official documents*. For those discourses that are with low scores, context-dependent referents appear frequently. A typical situation-dependent form of discourse is *broadcasts*, in which situations and time are attached with great importance.

As can be seen from Table 4, all the corpora examined in the present study are placed on the positive side of Dimension 3 and thus display a shared tendency towards explicit references. NES is the least referentially elaborated, and in Biber's (1988) dimension scale, it scores close to the genre of *religion* (3.7). The remaining three Chinese learner corpora, with CHSS and CEM respectively representing the least and the most referentially explicit corpora, exhibit a moderately developmental tendency towards explicit reference. The genre in Biber's scale that is the closest to CEM and CNEMU is *press reviews* (4.4); for CHSS, the nearest genre is *religion* (3.7).

TABLE 4
DIMENSION 3

	NES	CEM	CNEMU	CHSS
Dimension Score	3.17	4.85	4.46	3.73
Loading Features (Positive)				
WH relative clauses on object position (WHOBJ)	0.5	0.1	0.4	0.4
Pied-piping relative clauses (PIRE)	0.5	0.3	0.5	0
WH-relative clauses on subject position (WHSUB)	2.8	2.8	2.3	1.8
Phrasal coordination (PHC)	7.8	9.9	9.4	10.3
Nominalizations (NOMZ)	20.1	32.2	28.6	20.9
Loading Features (Negative)				
Time adverbials (TIME)	2.6	4.0	3.0	3.8
Place adverbials (PLACE)	3.3	2.8	2.6	2.3
Total adverbs (RB)	47.3	41.2	39.8	41.9

Table 4 presents the normalized frequencies of the features of the four corpora on Dimension 3. For the major positive loading features, NES is outnumbered by its Chinese counterparts on PHC and NOMZ, reflecting a comparatively less concern over explicit and elaborated references. As for the negative loading features, NES has slightly fewer TIMEs but noticeably more RBs than the three non-native corpora, hence illustrating a relatively heavier reliance on physical situation references. CHSS, CNEMU, and CEM differ considerably in the usage of NOMZ, but are quite consistent in their usage of the remaining features. Overall, native English speakers favor situation-dependent references in writing while Chinese learners of English prefer an elaborated and explicit manner in marking referents. This finding echoes the difference in informativeness between Chinese learners and native English speakers that is confirmed in the discussion of Dimension 1, as Biber (1988), in his interpretation of Dimension 3, points out that "referentially explicit discourse" also tends to be "informational" (Biber, 1988, p. 110). For the rising tendency among the three Chinese learner corpora, this study is inclined to ascribe it to their differences in the use of features marking informativeness, in particular nouns (NOMZ, though special, can still be classified into the noun category).

D. Dimension 4

Dimension 4 is interpreted as "Overt expression of persuasion" as it concerns the presence of persuasive features in a text. In Biber (1988), this dimension distinguishes between persuasive and less-persuasive discourses. Discourses with high scores on dimension 4 are common in having extensive persuasive features, for example, *professional letters* and *editorials*. By contrast, discourses with low scores on Dimension 4 are not persuasive in nature and show a rarity of persuasive features. A typical example is the register *broadcasts*, which, as claimed by Biber (1988), is extremely non-persuasive since it is "a simple reportage of events" and does not "involve opinion or argumentation at all" (Biber, 1988, p. 151).

All the four corpora on Dimension 4 (see Table 5) are well above zero, exhibiting an overt persuasive tone. Even CNEMU, the corpus which receives the lowest score among the four corpora, surpasses *professional letters* (3.5) - the most persuasive genre in Biber's (1988) dimensional scale. Again, this is expected, given that it is mostly argumentative writing that is targeted in the present study and that the primary purpose of argumentative writing is to persuade readers. To observe the dimension scores of the four corpora more closely, a clear distinction can be found between the native data and the three Chinese learner corpora. NES obtains by far the highest score (8.96) and therefore can be seen as highly persuasive in tone. Although CEM scores second to NES and still can be viewed as highly persuasive, it is much less persuasive than NES and resembles more closely the remaining two Chinese learner corpora. CHSS has a score slightly higher than that of CNEMU and consequently can also be described as persuasive as the genre *professional letters*.

TABLE 5
DIMENSION 4

	NES	CEM	CNEMU	MSS
Dimension Score	8.96	5.46	3.98	4.43
Loading Features (Positive)				
Infinitives (TO)	30.4	26.1	24.8	24.9
Predictive modals (PRMD)	11.1	8.6	8.3	8.6
Suasive verbs (SUAV)	6.0	2.8	2.4	2.7
Conditional adverbial subordinators (COND)	6.6	4.2	3.8	4.0
Necessity modals (NEMD)	5.8	8.4	7.4	8.4
Split auxiliaries (SPAU)	6.6	3.5	3.3	3.1

Table 5 shows that there are differences between Chinese learners of English and their native counterparts concerning the usage of persuasive features. The three Chinese learner corpora are higher in the use of NEMD, but significantly lower than the native speaker corpus on the rest of the linguistic features. It may suggest that Chinese EFL learners, in addition to their limited use of persuasive devices, tend to increase the persuasiveness of their writings by means of emphasizing the necessity of their arguments. A close examination of the linguistic forms of NEMD in the four corpora reflects that Chinese EFL learners use *should* and *must* far more frequently than native speakers do. Given that *should* and *must* indicate an obligatory or even authoritative stance (Biber et al., 1999; Zheng, 2010), the high frequencies of NEMD of the three Chinese learner corpora seem to signal that the Chinese learners are "forcing" readers to accept their claims or arguments. Regarding the overuse of the NEMD of the Chinese learner corpora, researchers have provided various explanations. One is that Chinese learners of English might incline to use early learned modals such as *should*, due to their consideration of reducing the possibility of making mistakes (Gu, 2014). Another major explanation offered by Liang (2008) and Zheng (2010) concerns the genre appropriateness of modals. It is argued that Chinese EFL learners fail (or at least partly fail) to understand the underlying generic features of necessity modals.

E. Dimension 5

Dimension 5 relates to the distinction between "abstract" and "non-abstract" focuses. It is used to distinguish between discourses with a highly abstract and technical informational focus and those with a non-abstract focus (Biber, 1988). Like Dimensions 2 and 4, this dimension includes only linguistic features with positive loadings. Features that mark the abstract information cluster in great numbers in some highly informational and abstract discourses such as *academic prose*, while appear infrequently in discourses that are non-abstract and less informational, such as *conversations*.

As can be seen from Figure 1, the scores of the four corpora are distributed towards the positive end of Dimension 5 with a tiny score fluctuation (2.5 to 3.07), which means that they have a consistent abstract concern. In Biber's (1988) dimension scale, the four corpora fall between *official documents* (4.7) and *religion* (1.4). The dimension score that is close to them in related literature is provided by Van-Rooy (2008) in his MD analysis of the Tawana learner English corpus (3.5) which consists of argumentative student writings. Taken together, it seems plausible to conclude that argumentative writing is generally having a relatively high abstract focus.

TABLE 6
DIMENSION 5

	NES	CEM	CNEMU	CHSS
Dimension Score	3.0	2.89	3.07	2.5
Loading Features (Positive)				
Conjuncts (CONJ)	4.1	5.9	6.8	7.1
Agentless passives (PASS)	8.7	8.5	7.1	5.8
Past participial clauses (PASTP)	0.1	0.3	0.1	0.1
By-passives (BYPA)	0.6	0.8	1.0	0.6
Past participial WHIZ deletion relatives (WZPAST)	1.3	0.6	0.8	0.3
Other adverbial subordinators (OSUB)	3.1	2.2	1.9	2.1

Despite the fact that the four corpora share a similar degree of abstract focus, they vary markedly in their use of the features loaded on Dimension 5, especially between the native and non-native corpora. In Table 6, Chinese EFL learners appear to overuse CONJ in English writing while NES has higher frequencies of WZPAST and OSUB, which suggests that experienced L1 writers are more flexible in their use of linguistic features that contribute to the increase of abstractness. The overuse of conjuncts in Chinese EFL learner writings is also observed by Xu and Liu (2012) in their comparison between Chinese learners and native English speakers on the utilization of conjuncts. The primary reason, according to them, is that Chinese learners use conjuncts not for the sake of cohesion, but to abide by certain grammatical rules or assessment standards that encourage the use of conjuncts.

V. CONCLUSION

This paper aims to conduct a comprehensive investigation into the writings of Chinese EFL learners. To achieve this aim, MD analysis, a corpus-based approach that combines macroscopic and microscopic analysis, was adopted. The MD analysis results reveal a couple of interesting facts regarding the differences between Chinese learner corpora and native English data.

On "Involved versus informational production" dimension, Chinese EFL learners show higher involvement and informativeness than their native English counterparts in English writing. A microscopic examination reveals that Chinese EFL learners favour the use of involvement features, nouns, and adjectives, and thus display an involved and at the same time informational writing style. Native speakers, by contrast, show a low involved writing style manifested by sparse use of involvement features. Besides, they surpass Chinese learners on type/token ratio and the use of prepositions, and therefore their writings can also be characterized by high precision of word choice and high integration of information.

As for the second dimension, "Narrative versus non-narrative concerns", the three groups of Chinese learners demonstrate a greater tendency towards non-narrative concerns than native English speakers because of their limited use of narrative devices, especially the perfect aspect and public verbs.

On Dimension 3, "Explicit versus situation-dependent reference", both Chinese learners and native English speakers tend to make frequent use of explicit references in writing, with the difference being that Chinese learners show a slightly higher degree. A close look at their use of linguistic features shows that phrasal coordination and nominalization occur more in writings by Chinese learners than in those by native English speakers. Native English speakers, on the contrary, make relatively more use of situation-dependent features such as place adverbials and other adverbs.

On Dimension 4, which involves the persuasiveness of a text, Chinese EFL learners score markedly low on persuasiveness in writing compared with their native English counterparts. They are found to overuse necessity modals while underuse other persuasive devices such as infinitives, predicative modals, suasive verbs, conditional adverbial subordinators, and split auxiliaries, if native English speakers' corresponding uses are taken as the benchmark.

On Dimension 5, "Abstract versus non-abstract information", the two groups show a similar focus on abstract information, but again they exhibit different preferences in their use of the features included in this dimension. Chinese EFL learners are shown to overuse conjuncts in writing while native speakers make a somewhat greater use of agentless passives, past participial WHIZ deletion relatives, and adverbial subordinators.

Concerning the differences between varied Chinese EFL learner groups, some interesting findings are also worth mentioning. On the first dimension, it is revealed that writings by learners from higher English education levels tend to be less involved but more informational in style. This is manifested by their greater use of informational features but less use of involvement features. On Dimensions 2, 3, and 4, an ascending tendency is found. Writings by Chinese EFL learners at higher education levels become more informational, more narrative, more explicit in making references, and more persuasive, and the driving force is their increasing use of informational devices, narrative devices, devices marking referential explicitness, and persuasive devices.

The above findings may provide some pedagogical implications for L2 writing. First, raising learners' register awareness should be emphasized in learners' writing practices. The overly involved writing style of Chinese learners and some informal expressions (e.g., contractions) indicate that they lack an understanding of the basic differences between speaking and writing and cannot write as the register required. The excessive use of involvement features, in addition to causing high vocabulary repetition which might impact the evaluation of writings, also makes writings "drowning" in subjectivity and spoken like. While the present study proposes that improving learners' language proficiency and increasing the exposure to written language can contribute to the lowering of involvement in learner writings, it seems to be more efficient and less time-consuming to directly raise their register awareness, the lack of which lies at the root of the issue.

Second, some involvement features, though closely related to spoken registers and subjectivity, can perform important rhetorical functions and accordingly should not be entirely avoided in writing. To take first-person pronouns as an example, Hyland (2001, 2002a, 2002b) argues that first-person pronouns possess important rhetorical functions including emphasizing the author's contribution and guiding readers, and consequently points out that avoiding personal pronouns, which is encouraged in some writing manuals or instructions, might do a "disservice" to L2 learners. From Table 3, we can see that experienced L1 writers, also utilized a certain amount of involvement features in their writings. Therefore, to prevent EFL learners from one extreme (i.e., high involvement) to the other (i.e., high detachment), the special functions performed by involvement features should be properly imparted to EFL learners.

Lastly, Chinese EFL learners' underuse and overuse of specific linguistic features may help highlight the aspects on which their writings can be further improved. For example, on Dimension 4, non-native writers could draw on native English writers' usage of persuasive devices to increase the persuasiveness of their writings.

REFERENCES

[1] Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. *Automatic treatment and analysis of learner corpus data*, 249-264.

[2] Alhusban, H. A., & Vijayakumar, C. (2021). Lexical Bundles in Saudi EFL Student Writing: A Study of Learner Corpus. *TESOL International Journal*, *16*(4). 2, 7-31.

[3] Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

[4] Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

[5] Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language, Structure and Use*. Cambridge University Press.

[6] Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL quarterly*, *36*(1), 9-48.

[7] Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English* (Vol. 2). Longman.

[8] Bychkovska, T., & Lee, J. J. (2017). At the same time: Lexical bundles in L1 and L2 university student argumentative writing. *Journal of English for Academic Purposes*, *30*, 38-52.

[9] Cao Y., & Xiao, R. (2013). A Multi-Dimensional Contrastive Study of English Abstracts by Native and Non-native Writers. *Corpora*, *8*(2), 209 - 234.

[10] Chen, M. (2013). Overuse or underuse: A corpus study of English phrasal verb use by Chinese, British and American university students. *International Journal of Corpus Linguistics*, *18*(3), 418-442.

[11] Chen, W. (2019). Profiling collocations in EFL writing of Chinese tertiary learners. *RELC Journal*, *50*(1), 53-70.

[12] Chou, M. C., &Wu, K. H. (2007). The Temporal System of Interlanguage of College EFL Learners in Taiwan. *Hwa Kang Journal of English Language & Literature*, *13*, 29 – 57.

[13] Darus, S., & Subramaniam K. (2009). Error analysis of written English essays of secondary school students in Malaysia: A case study. *European Journal of Social Sciences*, *8*, 483-495.

[14] Dutra, P. B. (2014). Multi-Dimensional analysis of pop songs. In T. B. Sardinha & M. V. Pinto (Eds.), *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber (Studies in Corpus Linguistics (SCL) 60)* (pp. 149 – 175). John Benjamins Publishing Company.

[15] Fang, X. C. (2014). *A corpus-based study on lexical and grammatical features in Chinese EFL learners' use of verbal "Non-finite Clauses"*. [Unpublished PhD dissertation]. Shanghai International Studies University.

[16] Friginal, E., & Hardy, J. A. (2014). Conducting multi-dimensional analysis using SPSS. In T. B. Sardinha, & M. V. Pinto (Eds.), *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber (Studies in Corpus Linguistics (SCL) 60)* (pp. 297–316). John Benjamins Publishing Company.

[17] Gilquin, G. & Paquot, M. (2008): Too chatty: Learner academic writing and register variation. *English Text Construction, 1.1,* 41-61.

[18] Gilquin, G., & Granger, S. (2011). From EFL to ESL: Evidence from the International Corpus of Learner English. In J. Mukherjee (Ed.), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap* (pp. 55-78). John Benjamins Publishing Company.

[19] Gu, T. (2014). A corpus-based study on the performance of the suggestion speech act by Chinese EFL learners. *International Journal of English Linguistics*, *4*(1), 103-111.

[20] Hyland, K. (2001). Humble servants of the discipline? Self-mention in research articles. *English for specific purposes*, *20*(3), 207-226.

[21] Hyland, K. (2002a). Authority and invisibility: Authorial identity in academic writing. *Journal of pragmatics*, *34*(8), 1091-1112.

[22] Hyland, K. (2002b). Options of identity in academic writing. *ELT journal*, *56*(4), 351-358.

[23] Hyland, K. (2016). *Teaching and Researching Writing* (3rd ed). Routledge.

[24] Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner corpus studies in Asia and the world*, *1*, 91-118.

[25] Ji, J., & Liang, M.C. (2015). Xue xi zhe ying yu yi lun wen zhong zhu yu sheng ming du yan jiu [Subject Animacy in Chinese EFL Learner's Argumentative Writing]. *Wai yu dian hua jiao xue*, *2,* 52-58.

[26] Liang, M. C. (2008). zhong guo da xue sheng ying yu bi yu zhong de qing tai xu lie yanjiu [A corpus-based study of modal sequences in Chinese tertiary EFL learners' written production]. *wai yu jiao xue yu yan jiu*, *40*(1), 51-58.

[27] Liu, J. (2012). CLEC-based study of tense errors in Chinese EFL learners' writings. *World Journal of English Language*, *2*(4), 11-23.

[28] Liu, Y., & Lu, X. (2020). Chinese EFL learners' misconceptions of noun countability and article use. *System*, *90*(10222), 1-12.

[29] Ma, G. H. (2002). Zhong mei da xue sheng ying yu zuo wen yu yan te zheng fen xi [Contrastive analysis of linguistic features between EFL and ENL essays]. *wai yu jiao xue yu yan jiu*, *34*(5), 345-349.

[30] Nini, A. (2014). *Multidimensional Analysis Tagger 1.2 - Manual*. Retrieved May 4 from http://sites.google.com/site/multidimensionaltagger.

[31] Nini, A. (2019). Multidimensional Analysis Tagger. In T. B. Sardinha & M. V. Pinto (Eds.). *Multidimensional Analysis: Research Methods and Current Issues* (pp. 67-93). Bloomsbury Academic.

[32] O'Dowd, E. (1998). *Prepositions and particles in English: A discourse-functional account*. Oxford University Press.

[33] Ong, J. (2011). Investigating the Use of Cohesive Devices by Chinese EFL Learners. *The Asian EFL Journal Quarterly*, *13*(3), 42-65.

[34] Pan, F. (2012). Zhong guo fei ying yu zhuan ye ben ke sheng he yan jiu sheng shu mian yu ti de duo te zheng duo wei du diao cha [MF and MD Analysis of Written Texts Produced by Chinese non- English Major Undergraduates and Graduates]. *Wai yu jiao xue yu yan jiu (bimonthly)*, *44*(2), 220-232.

[35] Sardinha, T. B. (2014). 25 years latter: Comparing Internet and pre-Internet registers. In T. B. Sardinha & M. V. Pinto (Eds.), *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber (Studies in Corpus Linguistics (SCL) 60)* (pp. 3– 33). John Benjamins Publishing Company.

[36] Sardinha, T. B., Kauffmann, C., & Acunzo, C. M. (2014). A multi-dimensional analysis of register variation in Brazilian Portuguese. *Corpora*, *9*(2), 239-271.

[37] Tse, Y. K. (2010). Parataxis and hypotaxis in the Chinese language. *International Journal of Arts and Sciences*, *3*, 351–359

[38] Van Rooy, B. (2008). A multidimensional analysis of student writing in Black South African English. *English World-Wide*, *29*(3), 268-305.

[39] Wen, Q. F., Ding, Y. R., & Wang, W. (2003). Zhong guo da xue sheng ying yu shu mian yu zhong de kou yu hua qing xiang: gao shui ping ying yu xue xi zhe yu liao dui bi fen xi [Features of oral style in English compositions of advanced Chinese EFL learners: An exploratory study by contrastive learner corpus analysis]. *Wai yu jiao xue yu yan jiu*, *4*, 268-274.

[40] Wen, Q. F, Wang, L. F., & Liang, M. C. (2005). *Spoken and written English corpus of Chinese learners*. Foreign Language Teaching and Research Press.

[41] Xu, J. J. (2016). Zhong guo xue sheng wan pian ying yu zuo wen yu liao ku jie shao [An Introduction to Ten-thousand English Compositions of Chinese Learners (The TECCL corpus)], *yu liao ku yu yan xue*, *3*(2), 108-112.

[42] Xu, Y. T., & Liu, Y. H. (2012). The Use of Adverbial Conjuncts of Chinese EFL Learners and Native Speakers–Corpus-based Study. *Theory and Practice in Language Studies*, *2*(11), 2316-2321.

[43] Xue, X. Z. (2015). *Ten-thousand English Compositions of Chinese Learners (The TECCL corpus), Version 1.1.* The National Research Centre for Foreign Language Education, Beijing Foreign Studies University.

[44] Yang, L., Ma, A. P., & Cao, Y. (2013). Lexical negative transfer analysis and pedagogic suggestions of native language in Chinese EFL writing. *The proceedings of the 2013 Conference on Education Technology and Management Science (ICETMS 2013)* (pp.669 – 672). Atlantis Press.

[45] Yang, W. X., & Sun, Y. (2012). The Use of Cohesive Devices in Argumentative Writing by Chinese EFL Learners at Different Proficiency Levels. *Linguistics and Education*, *23*. 31-48.

[46] Zhang, H. P. (2013). *A Corpus-Based Study of Conceptual Transfer in Chinese Learners' English*. [Unpublished PhD dissertation]. Northeast Normal University.

[47] Zheng, Q. (2010). Modality and Generic Features in Chinese EFL *Writings. Chinese Journal of Applied Linguistics*, *33*(5), 40-51.

[48] Zhang, Z. S. (2012). A corpus study of variation in written Chinese. *Corpus Linguistics and Linguistic Theory*, *8*(1), 209-240.

Xiaoyun Li is currently a PhD student from the Department of Theoretical Linguistics at the University of Szeged, Hungary. He received his M.Sc. degree in Foreign Linguistics and Applied Linguistics from the Xi'an Polytechnic University, China, in 2017. His current research interests include corpus linguistics, learner language study, discourse analysis, English for Academic Purpose (EAP) study, and Natural Language Processing (NLP).