A Corpus-Based Comparative Study on Syntactic Complexity in University Students' EFL Writing in Southwestern China: A Case of Pu'er University

Yang Yang[1], Ngee Thai Yap[1] & Afida Mohamad Ali[1]

[1] Faculty of Modern Languages and Communication, Universiti Putra Malaysia, Malaysia

Correspondence: Ngee Thai Yap, Department of English, Faculty of Modern Languages and Communication, Universiti Putra Malaysia, 43400, Serdang, Selangor, Malaysia.

Received: September 8, 2022 Accepted: October 10, 2022 Online Published: October 11, 2022

doi:10.5430/wjel.v12n8p172 URL: https://doi.org/10.5430/wjel.v12n8p172

Abstract

Syntactic complexity is the variety and sophistication degree of the syntactic structures conveyed in written production. The syntactic complexity of general Chinese university students' EFL writing has been studied previously, but the performance of university students in educationally underdeveloped Southwestern China remains unclear. Taking Pu'er University as a case, this study collected 400 EFL compositions from 100 university students in Southwestern China and compared them with 200 writing samples from the *Louvain Corpus of Native English Essays*. Scores of 11 syntactic complexity indices were calculated using the *L2 Syntactic Complexity Analyzer*. The independent samples t-test was conducted to investigate whether and the extent to which the two groups differed on syntactic complexity indices. The results showed that university EFL students in Southwestern China produce a similar length of linguistic units when compared to native English writers. However, the amount of subordination in EFL writing is significantly less than that of native English writers. For the amount of coordination, the university EFL students produced a lower proportion of coordinate phrases than that of native writers, but the proportion of coordinate sentences is not significantly different between the two groups. Finally, for degree of phrasal sophistication, university EFL students in Southwestern China produce significantly fewer complex nominals than native writers do. The results imply that university students in Southwestern China should write more subordinated sentences and complex nominals, such as nominal clauses, infinitives, or gerunds, in their future EFL writing, instead of writing long sentences just heavily relying on simple coordination.

Keywords: corpus-based study, comparative study, syntactic complexity, EFL writing, Southwestern China

1. Introduction

Several linguistic features are highly correlated with writing quality, such as word frequency, lexical diversity, and coherence (McNamara et al., 2010). In addition to these features, syntactic complexity (SC) is an important indicator of the quality of English as a foreign language (EFL) writing (Bi & Jiang, 2020; Casal & Lee, 2019; Laufer & Nation, 1995; McNamara et al., 2010; Malvern & Richards, 2012). For example, Bi and Jiang (2020) reported that SC can significantly predict the quality of young adolescent EFL learners' writing. Besides, Casal and Lee (2019) claimed that the global and phrasal measures of SC can significantly distinguish between different grade tiers of the second language (L2) writing.

SC can be understood as the variety and sophistication degree of the syntactic structures conveyed in written production (Bulté & Housen, 2014; Lu, 2011; Ortega, 2003). The importance of SC in EFL writing research and pedagogy has long been recognized. Over the past two decades, a large variety of studies have examined the relationship of SC in EFL writing to language proficiency (e.g., Ai & Lu, 2013; Lu, 2011; Norrby & Håkansson, 2007; Ortega, 2000, 2003; Vyatkina, 2013; Wolfe-Quintero et al., 1998) or the quality of EFL writing (e.g., Taguchi et al., 2013; Yang et al., 2015). A comparison of SC in native English writers' language production with that of non-native writers is also an important strand of studies in this area (e.g., Ai & Lu, 2013; Foster & Tavakoli, 2009; Lu & Ai, 2015).

When it comes to the context of China, comparative studies on the difference in SC between native speakers and Chinese EFL students have been conducted by taking China as a whole (e.g., Ai & Lu, 2013), but the difference between university students in Southwestern China, an educationally under-developed area, and the native speakers remained unknown. To address this research gap, the present study has collected compositions to build a learner corpus by taking Pu'er University as a case. Pu'er University is located in Yunnan Province, which is in the southwestern part of China. Because of the unbalanced development of education in China, especially the underdeveloped education in Southwestern China (Ren, 2019; Shi, 2018), the EFL proficiency of university students in this area is regarded as lower than that in other parts of China.

This study takes English writing samples produced by students enrolled in English programs of study at Pu'er University to investigate the difference in SC in English writing between this group of EFL learners and native English writers. The research questions of this study are as follows:

1) Are there significant differences between the writing produced by university students in Southwestern China and that of native English

writers in terms of SC?

2) To what extent and in which aspects are there differences in the writing produced by Chinese university EFL students and that of native English writers?

2. Literature Review

Syntactic complexity, also called linguistic complexity or syntactic maturity, can be understood as the variety and degree of sophistication of the syntactic structures used in written production (Bulté& Housen, 2014; Lu, 2011; Ortega, 2003). For example, a) at the global level, the mean length of sentences, b) at the clausal level, the amount of coordination and subordination, and c) at the phrasal level, the number of complex nominals have been taken as measures of SC of a written product. It has received broad attention from second or foreign language writing development researchers who have been searching for valid and reliable developmental measures that can be used to impartially gauge second or foreign language learners' overall proficiency or developmental level in their target language (e.g., Larsen-Freeman, 1978, 2009; Lu 2011; Norris & Ortega, 2009; Ortega, 2003; Wolfe-Quintero et al., 1998).

An important and major focus in such research ~~attempts~~ is the relationship between language proficiency and SC in second or foreign language writing. Some researchers conducted longitudinal research to find out this relationship by examining changes in SC over a certain period (e.g., Casanave, 1994; Norrby & Håkansson, 2007; Ortega, 2000; Stockwell, 2005; Stockwell & Harrington, 2003). For instance, Stockwell and Harrington (2003) examined email exchanges for 5 weeks between native Japanese students and Japanese as foreign language learners and reported a credible increase in syntactic development reflected in several measures as well as in qualitative ratings given by native speakers.

Some other researchers (e.g., Ferris, 1994; Larsen-Freeman, 1978; Lu, 2010) have conducted cross-sectional studies to determine the degree to which various measures of SC correlate with language proficiency. For example, Ferris (1994) conducted a correlation analysis for 160 English as second language (ESL) texts produced by a group of low proficiency students and a group of high proficiency students and reported that some measures, such as number of words, word length, use of synonyms/antonyms and passive structures significantly predict ESL learners' writing proficiency. With factor analysis, it was reported that variables of words per sentence, relative clauses, coordination, and prepositional phrases covary with each other in terms of sentence complexity. Lu (2010) analyzed college-level second language writing data from the *Written English Corpus of Chinese Learners* (WECCL; Wen et al., 2005), and presented findings of a corpus-based assessment of 14 measurements of SC as objective indices of EFL writers' language development. The 14 indices, such as mean length of T-unit, dependent clauses per T-unit, coordinate phrases per clause, and complex nominal per clause, are from the following five dimensions: length of production unit, sentence complexity, amount of subordination, amount of coordination, and degree of phrasal sophistication.

Though these longitudinal studies (e.g., Casanave, 1994; Norrby & Håkansson, 2007; Ortega, 2000; Stockwell, 2005; Stockwell & Harrington, 2003) and cross-sectional studies (e.g., Ferris, 1994; Larsen-Freeman, 1978; Lu, 2010) differed in terms of data size, specific ways to measure SC, as well as the operationalization of language proficiency (e.g., standardized test scores, holistic ratings, or using program level), most of these studies reported that SC is an important and reliable predictor for language development and language proficiency. More importantly, these researchers have proposed various reliable indices to measure SC.

Another important strand in this field is studies examining the relationship of the various task-, context-, and learner-related variables with SC and the second or foreign language development or proficiency (e.g., Ellis & Yuan, 2004; Lu, 2011; Ortega 2003; Sotillo, 2000; Way et al., 2000). Sotillo (2000) investigated the SC in ESL learners' target language production via computer-mediated communication (CMC) and the result showed that the degree of SC is different due to the different modes of CMC, such as synchronous communication and asynchronous communication. Way et al. (2000) investigated the effects of three different writing tasks (descriptive, narrative, and expository) and three different writing prompts (bare, vocabulary, and prose model) on target language writing and found out that the writing tasks and prompts can influence the SC of language production. By analyzing 42 Chinese learners' written narratives elicited by picture composition, Ellis & Yuan (2004) reported that planning conditions (pre-task planning, unpressured online planning, and no planning) have effects on the second language learners' writing and the pre-task planning can result in greater syntactic variety. Using a computational system designed to automatically measure SC with 14 indices, Lu (2011) also reported the impact of genre and timing conditions on SC measures proposed in his previous second language writing development studies (Lu, 2010). These past studies have shown that SC is correlated with various variables.

Besides the relationship between SC and foreign language development or proficiency, there has also been significant attention on comparing SC in native speakers' language production with that of non-native speakers. For instance, Foster and Tavakoli (2009) reported the applicability and validity of using the writing production of native English speakers as the baseline to investigate ESL or EFL learners' performance. Taking subordination and mean length of utterance as indices of SC, they analyzed the effect of task features on SC in native speakers' language production and compared the result with that obtained from their previous study on non-native speakers' language production (Tavakoli & Foster, 2008).

When it comes to the context of China, there remains a paucity of comparative studies on Chinese students' EFL writing with a few exceptions. Ai and Lu (2013) conducted a comparative study investigating SC in university students' writing of native speakers and non-native speakers. Analyzing essays from the *Written English Corpus of Chinese Learners* Version 2.0 (WECCL 2.0; Wen et al., 2008) and essays from the *Louvain Corpus of Native English Essays* (LOCNESS; Granger, 1998), Ai and Lu used 10 SC indices to investigate

whether and to what degree the native and non-native English speakers differ in their writing from four dimensions: length of production unit, amount of subordination, amount of coordination, and degree of phrasal sophistication. However, the Chinese EFL learner corpus used in Ai and Lu's (2013) study, WECCL 2.0, may no longer be regarded as representative of Chinese EFL learners. On the one hand, the corpus is not up-to-date since it was built in 2008. On the other hand, EFL writing proficiencies vary greatly in different regions of China. WECCL 2.0 collects English writing samples from 34 different universities in China, but only one of them is from Southwestern China. Southwestern China is regarded as a less developed area compared with other parts of China both in terms of the economy and education (Ren, 2019; Shi, 2018). It is reported that the EFL proficiency of university students from different regions or parts of China is different (Bao, 2013; Xia et al., 2019) and that from Western China is lower than that of Eastern China because of the language learning environment (Zhang et al., 2014). Therefore, taking Pu'er University as a case, this study seeks to investigate the level of SC in EFL writing of university students in Southwestern China in comparison with that of native English writers to fill the aforementioned gaps.

3. Methodology

3.1 Data Collection

Four hundred compositions written by 100 Chinese university students were collected. These students were second-year English program students at Pu'er University. The 113 students were from three classes of the English Writing course. Compositions of 13 students who cannot finish all four writing tasks were deleted from the corpus to keep the consistency of the writing production. They are assigned to complete four English writing tasks in four different weeks, the first, third, fifth, and seventh weeks of a semester. The genres of writing include letter writing, narration, causes and effects, and argumentation (see Appendix A, writing task instructions). The writing tasks are adopted from the textbook of the writing course (Wang, 2013). Two example compositions by students of Pu'er University are attached in Appendix B. The data have been reviewed and published on Mendeley Data (Yang et al., 2022).

The *Louvain Corpus of Native English Essays* (LOCNESS; Granger, 1998) was adopted in this study as a reference corpus and as a baseline to be compared with Chinese university students' English writing. LOCNESS contains three parts: British pupils' A-level essays (60,209 words), British university students' essays (95,695 words), and American university students' essays (168,400 words). Two hundred essays were randomly sampled from the latter two parts of LOCNESS. The essays were argumentative and literary essays written by American and British university students. In this study, the dataset for both the American and British writers is taken together to represent the native English writers. Yang and Geng (2021) have earlier shown that there are no significant differences between these two samples in terms of syntactic complexity. Table 1 is the summary of the dataset.

Table 1. Summary of data

	Chinese students' writing	LOCNESS
Number of compositions	400	200
Mean length of compositions	174.08	793.22
Standard deviation of length	45.01	390.81
Total	69,632	158,643

It can be found that the mean length of compositions in the LOCNESS corpus is much larger than that of Chinese students' writing. This will not affect the analysis and result of the present study since all the indices of the SC are calculated as the mean length of production unit or ratios of frequency of one syntactic structure to that of another in complete texts (Ai & Lu, 2013; Lu & Ai, 2015; Wang & Slater, 2016).

3.2 Data Analysis

A large variety of SC measures in second or foreign language writing have been proposed by many researchers in the literature. Measures employed in the present study are adapted from Lu's (2010) model, which uses 14 indices covering 1) length of production unit, 2) amount of subordination, 3) amount of coordination, 4) degree of phrasal sophistication and overall sentence complexity to gauge SC of English writing. To relieve the work strength of manual analysis, Lu (2010) designed a computational system for automatic measurement of SC, the *L2 Syntactic Complexity Analyzer* to compute the 14 measures mentioned above. Eleven of 14 measures were applied in this study, as summarized in Table 2. Three measures, clauses per T-unit, complex T-units per T-unit, and clauses per sentence, were excluded in the present study because they were weak candidates for developmental measures (Lu, 2011). By using the *L2 Syntactic Complexity Analyzer*, a zip file containing the writing samples can be uploaded, and the results of the aforementioned measures were obtained in a CSV file, which was then imported into spreadsheets and statistical packages for further analysis.

Table 2. SC measures investigated and formulae

Measures	Code	Formula
Length of production unit		
Mean length of sentence	MLS	number of words/number of sentences
Mean length of clause	MLC	number of words/number of clauses
Mean length of T-unit	MLT	number of words/number of T-units
Amount of subordination		
Dependent clauses per clause	DC/C	number of dependent clauses/number of clauses
Dependent clauses per T-unit	DC/T	number of dependent clauses/number of T-units
Amount of coordination		
Coordinate phrases per clause	CP/C	number of coordinate phrases/number of clauses
Coordinate phrases per T-unit	CP/T	number of coordinate phrases/number of T-units
T-units per sentence	T/S	number of T-units/number of sentences
Degree of phrasal sophistication		
Complex nominals per clause	CN/C	number of complex nominal/number of clauses
Complex nominals per T-unit	CN/T	number of complex nominal/number of T-units
Verb phrases per T-unit	VP/T	number of verb phrases/number of T-units

Adapted from Ai and Lu (2013).

It should be noted that consistent and unambiguous definitions of the production units and syntactic structures involved in gauging one or more of the SC measures are supposed to be given and used. The following are definitions of seven linguistic units involved in the present study.

1) Sentence: a set of words that is complete in itself, typically containing a subject and predicate, that in writing usually begins with a capital letter and concludes with appropriate end punctuation, such as a period, exclamation mark, question mark, and occasionally elliptical marks or closing quotation marks.

2) Clause: a group of words that contains a subject and predicate (finite verb), including independent, adverbial, adjective, and noun clauses. Non-finite verb phrases do not belong to this category but are counted as verb phrases (Hunt, 1965; Polio 1997).

3) Dependent clause: a group of words that has both a subject and a verb but (unlike an independent clause) cannot stand alone as a sentence, including adverbial, adjective, and noun clauses (Cooper 1976; Hunt 1965; Kameen 1979).

4) T-unit: a linguistic unit coined by Hunt (1965, 1970). It is defined as the "shortest grammatically allowable sentences into which (writing can be split) or minimally terminable unit" (Hunt, 1965, p. 21). Often, but not always, a T-unit is a sentence. Young (1995) gave some examples of what a T-unit is and is not:

> "The following elements were counted as one T-unit: a single clause, a matrix plus subordinate clause, two or more phrases in apposition, and fragments of clauses produced by ellipsis. Co-ordinate clauses were counted as two t-units. Elements not counted as t-units include backchannel cues such as *mhm* and *yeah*, and discourse boundary markers such as *okay*, *thanks*, or *good*." (Young, 1995, p. 38)

5) Coordinate phrase: a coordinate adjective, adverb, noun, or verb phrase.

6) Complex nominal: (1) a noun phrase with one or more adjective, possessive, prepositional phrase, adjective clause, participle, or appositive; (2) a nominal clause; or (3) an infinitive or gerund in subject position (Cooper 1976).

7) Verb phrase: a finite or non-finite verb phrase.

After the values of the aforementioned 11 measures of SC were calculated by using the *L2 Syntactic Complexity Analyzer*, the independent samples t-test was conducted to compare the SC between Chinese university students' writing and LOCNESS.

Since 11 independent samples t-tests were conducted simultaneously, Bonferroni correction was applied to avoid the spurious positives caused by the multiple comparisons problem. A stricter significance threshold, 0.0045, which is 0.05 divided by 11, was taken in the present study. Before conducting a t-test, internal consistency reliability among the values of 11 SC measures was analyzed. Based on the result, Cronbach's alpha of 0.7 shows that the instrument is reliably acceptable.

Kolmogorov-Smirnov normality test was conducted to test the distribution of values of 11 SC measures of both groups. The result showed that the values of most of the measures were not normally distributed, except for DC/C for both groups (p = 0.20 for LOCNESS and 0.18 for another group respectively) and CN/C for LOCNESS (p = 0.76). However, according to the Central Limits Theorem, the independent samples t-test is valid for large samples from non-normal distributions since the sample size of the present study is as large as 600 (see Table 1).

4. Results and Discussion

Group statistics and results of the independent samples t-test are shown below in Table 3. As shown in Table 3, none of the mean values of the aforementioned 11 SC measures of Chinese university students' EFL writing in Southwestern China is larger than that of native English writers. Except for the values of T/S, which are approximately equal between the two groups, all the other 10 values of Chinese university students' EFL writing are slightly or dramatically smaller than that of native writers. Besides, all the standard deviations of the

11 average means of Chinese university students' EFL writing in Southwestern China are larger than that of native speakers, which means that the values of SC measures are more dispersed among Chinese university students' EFL writings. It can be generally concluded that the university students' EFL writing in Southwestern China is syntactically less complicated than that of native speakers since most values of SC indices of the EFL group are smaller than that of native English writers. Besides, the syntactic proficiencies are more diverse among university students' EFL writing in Southwestern China.

Table 3. Group statistics and results of independent samples t-tests

Measures	Code	Mean (SD)		t	Sig.
		EFL writing	LOCNESS		
Length of production unit					
Mean length of sentence	MLS	20.48 (18.87)	20.66 (4.85)	-0.18	.858
Mean length of clause	MLC	10.15 (8.89)	10.28 (1.73)	-0.27	.785
Mean length of T-unit	MLT	18.03 (16.42)	18.19 (4.00)	-0.19	.852
Amount of subordination					
Dependent clauses per clause	DC/C	0.33 (0.13)	0.40 (0.09)	-8.05	.000*
Dependent clauses per T-unit	DC/T	0.63 (0.45)	0.74 (0.28)	-3.91	.000*
Amount of coordination					
Coordinate phrases per clause	CP/C	0.16 (0.17)	0.25 (0.10)	-6.70	.000*
Coordinate phrases per T-unit	CP/T	0.28 (0.26)	0.44 (0.17)	-9.14	.000*
T-units per sentence	T/S	1.14 (0.26)	1.14 (0.11)	0.03	.974
Degree of phrasal sophistication					
Complex nominals per clause	CN/C	0.96 (0.59)	1.25 (0.34)	-7.61	.000*
Complex nominals per T-unit	CN/T	1.65 (1.00)	2.21 (0.65)	-8.16	.000*
Verb phrases per T-unit	VP/T	2.30 (0.78)	2.38 (0.42)	-1.76	.079

* The difference is significant at the level of 0.0045 after Bonferroni Correction

4.1 Length of Production Unit

In terms of length of production unit, based on the results of the independent samples t-test, the significant values of mean length of sentences, clauses, and T-units are greater than alpha at 0.0045 level of significance, so there is no sufficient evidence to reject the null hypothesis that there is no difference between the EFL group and the native English writers in terms of indices of length of production unit. It can be concluded that the mean length of sentences, clauses, and T-units of writing are not significantly different between Chinese university EFL students and native speakers. This result is not consistent with what Foster and Tavakoli (2009) reported where they found that non-native speakers produced shorter utterances than native speakers do in oral narratives. Neither is it consistent with the result reported by Ai and Lu (2013), who obtained the conclusion that the length of production units of non-native speakers is shorter than that of native speakers in English writing. The result of the present study may be caused by the teaching focus on sentence-combining skills that were included in the English Writing courses for English program students at Pu'er University where students were encouraged to write "long sentences".

4.2 Amounts of Subordination

Concerning amounts of subordination, measured by dependent clauses per clause and dependent clauses per T-unit, Chinese university students in Southwestern China use significantly (p = 0.000) less proportion of dependent clauses in English writing than native speakers do. This result is consistent with what Foster and Tavakoli (2009) found and what Ai and Lu (2013) found where the non-native speakers produced fewer dependent clauses than native speakers do in oral narratives and written production.

4.3 Amounts of Coordination

For amounts of coordination, significant differences are found in CP/C and CP/T, but not found in T/S between the two groups. This suggests that Chinese university EFL students averagely produce significantly (p = 0.000) fewer coordinate phrases than native speakers do, but they do produce approximately the same proportion of coordinate sentences as native speakers do in English writing. However, in Ai and Lu's (2013) research, the result of non-native speakers differed from native speakers only on coordinate phrases per T-units but not on coordinate phrases per sentence and the amount of sentential coordination. Similar values, often approximating 1.1, of T-units per sentence among EFL learners with different proficiency levels (Ai & Lu, 2013), as well as among writers with diverse first-language backgrounds (Lu & Ai, 2015) were reported in most works of literature. This may be because a T-unit is often, though not always, a sentence as well as the fact that combining simple sentences into a compound sentence with coordinate clauses and conjunction is a simple skill that most English language users and learners can easily grasp. This result also suggests that T/S may be not an effective indicator of SC to differentiate different proficiency levels of English language users or learners because its value always remains steady. Some example sentences from Pu'er university students' EFL writing are provided below.

1) At that time, my father had a motorcycle, and he then took us to go to the county. (from the writing of task 2)

2) Therefore, my aunt, my cousin and I decided to visit the West Lake on account of the distance between HangZhou and JiaXing was very close and it was just a fantastic time for enjoying the sight of beautiful lotus. (from the writing of task 2)

The two example sentences are coordinated sentences connected by coordinating conjunction "and", but example sentence 2 is a run-on sentence. The statistical result of Table 3 and the example sentences show that university students in Southwestern China can produce "long" enough sentences in their EFL writing. However, they do so just merely by relying on combining simple sentences into coordinated sentences but without subordination. Furthermore, they may produce grammatically wrong sentences, such as run-on sentences.

3) Although a part-time job after class can only earn a little money, but money is can reduce the family burden of students or let them have enough money to spend, to enjoy my college life. (from the writing of task 4)

Example sentence 3 starts with subordinating conjunction "although", so on the surface, it may look like the EFL learner knows how to write a subordinate compound sentence. However, the learner also uses the coordinating conjunction "but" in the same sentence, which is a common error among EFL learners in China who use both a subordinating and coordinating conjunction in the same sentence (Chen, 2013; Mei, 2014). This is piece of evidence that they have not understood subordination well and this is due to their thinking mode of Chinese writing and negative transfer of their mother tongue (Zhang, 2022).

4.4 Degree of Phrasal Sophistication

Finally, in terms of the degree of phrasal sophistication, significant differences ($p = 0.000$) in mean values of complex nominals are found between the two groups, but the mean values of verb phrases are not significantly different ($p = 0.079$). The result is largely in line with what Ai and Lu (2013) and Lu (2011) found. This result suggests that Chinese university EFL students use fewer complex nominals than native speakers do, but they produce a similar proportion of verb phrases to native speakers.

5. Conclusion and Implications

To conclude, the English writing of Chinese university students in Southwestern China is syntactically less complicated than that of native English writers. The two groups differ in amounts of subordination, amounts of coordination, and degree of phrasal sophistication but not in the length of production units. These results are largely consistent with most of the former studies (e.g., Ai & Lu, 2013; Foster & Tavakoli, 2009; Lu, 2011, Lu & Ai, 2015) on SC, but the similar length of production units between non-native writers and native English writers is an interesting different finding compared with the previous literature. Chinese university EFL students in Southwestern China can produce a similar length of sentences, clauses, and T-units, but they use less subordination and phrasal coordination in their EFL writing. Besides, they use fewer complex nominals than native English writers do. This may suggest that Chinese university EFL students in Southwestern China can write "long sentences" but these "long sentences" may be just simple sentential combinations without much subordination, or these sentences may be grammatically wrong, such as run-on sentences.

Thus, university students in Southwestern China should try to write more subordinated sentences and more complex nominals, such as nominal clauses, infinitives, or gerunds, in their future EFL writing, instead of writing long sentences just heavily relying on simple coordinated sentences. In addition, in future English writing teaching, more attention should be given to the structure of subordination.

The measures of SC study in the present study are adapted from Ai and Lu's (2013) model and the values of these measures are calculated by using the *L2 Syntactic Complexity Analyzer*. Three measures are excluded in this study because of the proven weak representativeness for developmental measures. For a similar reason, T-units per sentence (T/S) can be excluded in future research, especially for research aiming to differentiate developmental levels because its value always remains steady.

There are also limitations to the present study. As mentioned before, there are studies on the difference in SC between native speakers and non-native speakers in China as a whole, but the difference between students in Southwestern China, an educationally under-developed area, and the native speakers are unknown. To address this gap, this study collected compositions from Pu'er University to build the learner corpus to represent Southwestern China, but the representativeness may not be perfectly valid. For future research, writing samples can be collected from EFL students not only from Yunnan Province but also from the other two Southwestern provinces of China, Guizhou and Sichuan Province, to compare the difference between their EFL writing and that of native speakers. Even though the study may not be representative enough, the results have proven that SC is indeed a good predictor of writing quality. Besides, the SC proficiency of university students' EFL writing in Southwestern China is different from that of both native English writers and average Chinese EFL writers.

Acknowledgements

We would like to express our gratitude to Dr. Helen Tan (ORCID: 0000-0002-8418-6169). This paper and the research behind it would not have been possible without her valuable suggestions and revision of this paper. We would also like to give thanks to the editorial team and anonymous reviewers. Their feedback has played a crucial role in improving the quality of this paper.

References

Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data*. (pp. 249-264). Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/scl.59.15ai

Bao, Y. (2013). How to Treat Students' Regional Differences in College English Teaching[J]. *Science & Technology Information, 19*, 209. https://doi.org/10.3969/j.issn.1001-9960.2013.19.159

Bi, P., & Jiang, J. (2020). Syntactic complexity in assessing young adolescent EFL learners' writings: Syntactic elaboration and diversity. *System*, 102248. https://doi.org/10.1016/j.system.2020.102248

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing, 26*, 42-65. https://doi.org/10.1016/j.jslw.2014.09.005

Casal, J. E., & Lee, J. J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing, 44*, 51-62. https://doi.org/10.1016/j.jslw.2019.03.005

Casanave, C. P. (1994). Language development in students' journals. *Journal of Second Language Writing, 3*(3), 179-201. https://doi.org/10.1016/1060-3743(94)90016-7

Chen, H. (2013). An Analysis of the Reasons for the Wrong Use of Adverbial Clauses in English Writing. *Journal of Teaching and Management, 39*(12), 103-104. Retrieved from https://kns.cnki.net/kcms/detail/detail.aspx?FileName=JXGL201312038&DbName=CJFQ2013

Cooper, T. C. (1976). Measuring written syntactic patterns of second language learners of German. *The Journal of Educational Research, 69*(5), 176-183. https://doi.org/10.1080/00220671.1976.10884868

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26*, 59-84. https://doi.org/10.1017/S0272263104026130

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *Tesol Quarterly, 28*(2), 414-420. https://doi.org/10.2307/3587446

Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency and lexical diversity. *Language Learning, 59*(4), 866-896. https://doi.org/10.1111/j.1467-9922.2009.00528.x

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In Granger, S. (Ed.) *Learner English on Computer* (pp. 3-18). London & New York: Addison Wesley Longman. https://doi.org/10.4324/9781315841342-1

Hunt, K. W. (1965). *Grammatical Structures Written at Three Grade Levels*. Champaign: National Council of Teachers of English.

Hunt, K. W. (1970). Syntactic maturity in schoolchildren and adults. *Monographs of the society for research in child development, 35*(1), iii-67. https://doi.org/10.2307/1165818

Kameen, P. (1979). Syntactic skill and ESL writing quality. In C. Yorio, K. Perkins. & J. Schachter (Eds.), *On TESOL' 79, The Learner in Focus* (pp. 343-364). Washington, DC: TESOL.

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly, 12*, 439-448. https://doi.org/10.2307/3586142

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 579-589. https://doi.org/10.1093/applin/amp043

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics, 16*(3), 307-322. https://doi.org/10.1093/applin/16.3.307

Lu, X (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474-496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36-62. https://doi.org/10.5054/tq.2011.240859

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing, 29*, 16-27. https://doi.org/10.1016/j.jslw.2015.06.003

Malvern, D., & Richards, B. (2012). Measures of lexical richness. *The encyclopedia of applied linguistics*. Wiley Online Library. https://doi.org/10.1002/9781405198431.wbeal0755

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written communication, 27*(1), 57-86. https://doi.org/10.1177/0741088309351547

Mei, N. (2014). Error Analysis of Adverbial Clauses in College Students' English Writing. *Examination Weekly, 16*(6), 92-93. https://doi.org/10.3969/j.issn.1673-8918.2014.06.085

Norrby, C., & Håkansson, G. (2007). The interaction of complexity and grammatical processability: The case of Swedish as a foreign language. *IRAL-International Review of Applied Linguistics in Language Teaching, 45*(1), 45-68. https://doi.org/10.1515/IRAL.2007.002

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555-578. https://doi.org/10.1093/applin/amp044

Ortega, L. (2000). *Understanding* syntactic complexity*: The measurement of change in the syntax of instructed L2 Spanish learners* [Doctoral dissertation]. University of Hawaii.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492-518. https://doi.org/10.1093/applin/24.4.492

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language learning, 47*(1), 101-143. https://doi.org/10.1111/0023-8333.31997003

Ren, F. H. (2019). Research on the Development Strategy of Higher Vocational Education in Remote Areas of Southwest China: A SWOT Model Based Analysis]. *Knowledge Economy*, *7*, 78. https://doi.org/10.3969/j.issn.1007-3825.2019.05.078.

Shi, Y. C. (2018). "Internet+" Education Transformation for Educational Poverty Alleviation in Ethnic Regions Southwest China. *China Educational Technology*, *4*, 26-34. https://doi.org/10.3969/j.issn.1006-9860.2018.04.005.

Sotillo, S. M. (2000). Discourse functions and syntactic complexity in synchronous and asynchronous communication. *Language Learning and Technology, 4*(1), 82-119. https://doi.org/10125/25088

Stockwell, G. (2005). Syntactic and lexical development in NNS-NNS asynchronous CMC. *The JALT CALL Journal, 1*(3), 33-49. https://doi.org/10.29140/jaltcall.v1n3.16

Stockwell, G., & Harrington, M. (2003). The incidental development of L2 proficiency in NS-NNS email interactions. *CALICO Journal, 20*(2), 337-359. https://doi.org/10.1558/cj.v20i2.337-359

Taguchi, N., Crawford, B., & Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly, 47*(2), 420-430. https://doi.org/10.1002/tesq.91

Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning, 58*(2), 439-473. https://doi.org/10.1111/j.1467-9922.2011.00642.x

Vyatkina, N. (2013). Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus. *The Modern Language Journal, 97*(S1), 11-30. https://doi.org/10.1111/j.1540-4781.2012.01421.x

Wang, S., & Slater, T. (2016). Syntactic complexity of EFL Chinese students' writing. *English Language and Literature Studies, 6*(1), 81-86. https://doi.org/10.5539/ells.v6n1p81

Wang, X. (2013). *English Writing* (Vol. 1). Shanghai Foreign Language Education Press.

Way, D. P., Joiner, E. G., & Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of French. *The Modern Language Journal, 84*(2), 171-184. https://doi.org/10.1111/0026-7902.00060

Wen, Q., Liang, M., & Yan, X. (2008). *Spoken and Written English Corpus of Chinese Learners (Version 2.0)*. Beijing: Foreign Language Teaching and Research Press.

Wen, Q., Wang, L., & Liang, M. (2005). *Spoken and written English corpus of Chinese learners*. Beijing, China: Foreign Language Teaching and Research Press.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. Honolulu HI: University of Hawaii Press.

Xia, Y. W., Kong, L. Y., Pan, Y. R., & Min, T. (2019). A Survey on the Current Situation of English Teaching Connectivity in College and Middle School. *English Campus, 26*, 30.
Retrieved from https://kns.cnki.net/kcms/detail/detail.aspx?FileName=XYYY201926027&DbName=CJFN2019

Yang, W., Lu, X., & Weigle, S. A. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing, 28*, 53-67.
https://doi.org/10.1016/j.jslw.2015.02.002

Yang, Y., & Geng, H. (2021). A Quantitative Comparison of Syntactic Complexity Between British and American University Students' ENL Writing. *Proceedings of the 4th International Conference of Languages, Education and Tourism (ICLET) 2021*, 29-43. https://conference.iium.edu.my/iclet2021/

Yang, Y., Yap, N. T., & Ali, A. M. (2022). 400 Compositions by EFL Learners of a University in Southwestern China (CEFLLUSC), *Mendeley Data*, V1. https://doi.org/10.17632/jxm2bwkc33.1

Young, R. (1995). Conversational Styles in Language Proficiency Interviews. *Language Learning, 45*(1), 3-42.
https://doi.org/10.1111/j.1467-1770.1995.tb00961.x

Zhang, J., Zhang, Y. X., & Shi, J. H. (2014). An Analysis of the Difference of English Cloze Mistakes between Senior High School Students and Junior High School Students in Western China. *Modern Female, 38*(11), 318.
Retrieved from https://kns.cnki.net/kcms/detail/detail.aspx?FileName=XDFN201411250&DbName=CJFQ2014

Zhang, P. (2022). A Research on the Improper Use of Contrastive Discourse Markers in CET 4 & CET 6 Writing and Analysis of the Causes. *Modern Linguistics, 10*(1), 105-112. https://doi.org/10.12677/ML.2022.101014

Appendix A

Writing Task Instructions

Writing Task 1, Letters

Your aunt sent you a sweater you need. Write a thank-you letter to her. You should write more than 150 words.

Writing Task 2, Narration

Write a memorable experience in your life. You should write about 150 words.

Writing Task 3, Causes & Effects

What are the reasons why the government of a developing country may want to send students abroad (to a developed country) to study?

Writing Task 4, Argumentation

Do you think college students should or should not do part-time jobs in their spare time? Write a paragraph to state your opinion.

Appendix B

Two Example Compositions by Students of Pu'er University

An Example Composition of Writing Task 2

This memorable and funny event happened when i was 7.At that time, i had short hair and looked like a little boy.

In that morning, my mother and i went to market to buy some fresh vegetables and fruits. Then my mother absorbedly was bargainning with the vendor who seells apples about price and she seems to forget that i near her. Suddenly, i felt a shadow behind me. It covered my mouth with its hands and took me away rapidly. I tried to struggle to free myself but his hands were so powerful! I tried to shouting but i could not emit any sound. Almost all the people around me did not notice me and this man. I only watched as i went farther and farther away from my mother.

Until a nobody's corner, this man said:"Listen!Boy! If you can keep quiet,i will let you go." I nodded heavily. After he let me go, i said:"Who are you?Strange sir,please call me pretty girl and you'd better let me back!" "What?! Are you a girl?"The man said surprisedly. "Yes ,i am." I replied. "Shit! Go! Go! Don't let me see you again!" The man shouted vehemently ,and then he left.

Finally, My mother and police found me.After they had listened my experience,they not only felt angry,but also felt funny.

An Example Composition of Writing Task 4

A part-time job is very popular in the university. This behavior is right or wrong? Some people think that it can bring many benefits to the students, while others think part-time job is not worth the time consuming. In my opinion, every coin has two sides. There are several reasons to support my point of view.

On the one hand, part-time jobs can earn some money, make more new friends, and get some work experience. Although a part-time job after class can only earn a little money, but money is can reduce the family burden of students or let them have enough money to spend, to enjoy my college life. In addition, in different places to do a part-time job can come into contact with all sorts of people, it definitely can make more new friends. Moreover, students can get some work experience from their part-time jobs, this is from the school curriculum not to middle school. This experience will help their career after graduation.

On the other hand, can take up part-time students time and not conducive to the establishment of team spirit. Everyone's energy is limited. If students have part-time jobs to take up their time, they will reduce the time to study. The results will affect their learning. If you have any time to do part-time students once, they will be less and less time to communicate with classmates. This is bad for students to form a team spirit. In general, a part-time job has both advantages and disadvantages. Each student should consider their own situation to choose whether to do a part-time job. If they can balance all of this, part-time job is acceptable. College students should not do a part-time job, everyone will hold different attitude toward life. Now, a part-time job is very popular among college students. Some people think part-time job is beneficial to the student, because it can relieve the economic burden and help to improve yourself. But others think part-time job is not conducive to the development of students. As far as I am concerned, I agree with the latter. There are several reasons to support my point of view.