Corpus-based Development of Syntactic Complexity in EFL Writing

Weilu Wang [1], *Manfu* Duan [2,*] and *Hongmei* Zhang [1]

[1] Foreign Languages College, Inner Mongolia University, Hohhot, P.R.C
[2] Divisions of International Cooperation & Exchange, Inner Mongolia University, Hohhot, PRC

Abstract. Syntactic complexity has long been taken as a significant factor in determining writing quality by EFL learners. And researchers in recent years have made a lot of efforts to devise and verify a wide range of factors or indicators for the purpose of measuring syntactic complexity or predicting EFL writing quality. The present study discussed bases itself on a self-built learner corpus with data collected over three years, with the aim of determining the applicable indices for predicting beginner writing quality. Based on previous research, the present study takes into account such factors as unit length, verb-VAC complexity, and clausal complexity. The results of pairwise comparisons by year indicate that there are significant differences for some indices but not for others. In terms of unit length, the three indicators of mean length of sentence, mean length of T-uni, and mean length of clause can serve as the main descriptive variables for the development of the language of beginners; for clausal complexity, seven indices: coordinate phrases per T-unit, verb phrases per T-unit, clauses per T-unit, coordinate phrases per clause, complex nominals per T-unit, complex nominals per clause and T-units per sentence, are the reliable indicators for beginner writing development. But indices for noun phrase complexity and verb-VAC complexity show no significant difference in the Kruskal-Wallis tests. The findings provide proof for the conclusion that knowledge and skills of modification, coordination, and subordination form the real difficulties for EFL beginners. It provides implications for coursebook design and classroom teaching where beginners are supposed to focus more on breeding awareness and skills in these aspects.

1 Introduction

As an important indicator of writing proficiency, syntactic complexity has been the focus of research in the field of second language acquisition in recent years. The complexity of language is closely related to the complexity of thinking, and complex language forms correspond to complex expressions of thought and vice versa. As young adults, the thinking ability of college students approximates that of an adult, and they need language forms corresponding to their thinking to express themselves fully and freely in writing. Therefore, the characteristics and development of EFL learner writing have become one of the most

* Corresponding author: duanmanfu@imu.edu.cn

concerning issues in English teaching. This will help to clarify the learning difficulties of learners and provide an important reference for foreign language teaching, syllabus design and textbook writing, and even classroom practice.

For different types of language users, there are differences in their actual language performance, and researchers have designed various indicators to measure this ability, which mainly boils down to three aspects namely accuracy, fluency, and complexity. Complexity is further divided into lexical complexity, syntactic complexity, and phonological complexity. In recent years, syntactic complexity has attracted widespread attention from language researchers and educational researchers, who have tried to design and verify a wide range of syntactic measurements, including verb-argument-construction complexity from the perspective of construction grammar ([1-4]) and some conventional global indicators: average T-unit length ([5, 6]), average subordinate clauses ([6]), average coordinating clauses ([6]), and noun-phrase complexity derived from the phrase structure grammar ([7-9]), but there are also studies showing that neither the average of clauses nor the complexity of noun phrases is a valid variable for predicting learners' writing performance ([10]). This controversial issue requires further data validation, and more evidence is needed to unveil such questions as which indicators can effectively mark the development of EFL learner writing, especially for beginners.

The syntactic complexity of EFL learner writing certainly requires some predictive indicators, but it is more than necessary to comprehensively describe its grammatical structure and language use to highlight the characteristics and development of learner language ([11]). By analyzing the use of that clause in learner language, Man's study ([12]) finds that learner language development is an organic, dynamic process. Indeed, learner performance in writing is developing over time. There are laws guiding such development, but the laws are not linear in deed, rather learner language is a self-contained system, which goes in its own right. And the actual use of language by EFL learners is affected by a variety of factors, among which cultural background, language proficiency, task type, and difficulty, more or less exert their influence. Research by Lu & Ai ([13]) shows that learners from different language backgrounds are using language at various syntactic complexities in English writing. Brunner ([14]) compares written English in a corpus by learners from the UK, Kenya, and Singapore and finds significant differences in the complexity of noun-phrases use. Nassseri ([15]) and Staples & Reppen ([16]) respectively investigate language differences in research papers by graduate students from different cultural backgrounds, and they find that learners were more similar in vocabulary use but significantly different in clause and phrase complexity. These studies demonstrate the peculiarities between learner groups, but learner language use undergoes development for sure, and such development is never unpredictable. The research discussed here is motivated by the need to work out the laws and rules guiding the development of learner language. It is based on the self-built learner corpus of English writing by beginners, and mainly focuses on the following questions: a) what are the indicators that predict the development of syntactic complexity of beginner writing? b) In what way is beginner writing developing over the years?

2 Method

The data for this study is from a self-built English writing corpus of beginners, which is collected during a period of three years of a college English program for beginners. The corpus consists of 500 English essays with a total capacity of about 7,000 words.

In this study, the natural language processing tool TAASSC (Syntactic Complexity Automatic Analysis Tool) was used to calculate the values for the indices of syntactic complexity by year. TAASSC was developed by Christopher Kyle, professor of linguistics

at the University of Oregon, and Scott Crossley, professor of applied linguistics at Georgia State University, to analyze the syntactic complexity of texts. The tool contains both conventional global metrics (such as average T-unit length) and phrase-level metrics (e.g., average adjectives of noun phrases and the average number of clauses), as well as some novel fine-grained indices like frequency of verb argument construction. The results of TAASSC calculation were analyzed in SPSS (Ver. 21) for significance. Based on previous research, this study tries to examine such indicators for syntactic complexity as noun phrase complexity, verb argument construction complexity, unit length, and clausal complexity.

3 Results and discussion

3.1 Noun phrase complexity

Based on the research by Kyle ([17]), noun phrase complexity can be measured by indices like noun phrase elaboration, complexity, and variation of nouns as modifiers, determiners in noun phrases, and possessives in noun phrases, of which the descriptive statistics are shown in table 1. According to Kyle ([17]), noun phrase elaboration mainly captures dependents, prepositions, adjectives, determiners, and verbal modifiers of nominals. The index of nouns as modifiers primarily captures nouns as nominal modifiers, direct object, and nominal subject modifiers. It also captures variation in the number of modifiers per nominal as prepositional objects, direct objects, and nominal subjects. The index of determiners captures the use of determiners in noun phrases as objects of the preposition, direct objects, and nominal subjects. The last index of possessives primarily captures the use of possessives in nominal subjects, direct objects, and prepositional objects.

As the data collected are in abnormal distribution, Kruskal-Wallis tests are employed for differences among year groups. The results indicate that there is no significant difference for the four indices (Sig. > 0.05), which serves as proof that indices for noun phrase complexity don't predict writing quality for beginners. It also means that the variety and complexity of noun-phrase use form the space for future development. It seems that the effective use of combinations of modifiers or determiners preceding a noun may take a fairly long time to develop. In beginner language, a bare countable noun is the most commonly used as either a subject or an object. In the case of one of the most frequently used words: "people", although there are about 587 hits of either subject or object in the corpus, those going with specific modifying words are comparatively rare: "many people" (23 hits), "some people" (22 hits), "more people" (18 hits), "old people" (8 hits), "young people" (6 hits), "other people" (6 hits), and "lots of people" (1 hit). Therefore, the rather limited mastery of rich, varying, and specific modifying words delimits beginners' use of noun phrases. And their syntactic knowledge deprives them of using complex modifying structures like non-infinitive verbs, prepositional phrases, and relative clauses.

Table 1. Descriptive statistics, results of Kruskal-Wallis tests and pairwise comparisons for noun phrase complexity (by year)

Variables	Year	N	Mean	Std. Deviation	Kruskal-Wallis test (Sig.)
np_elaboration	1st Year	136	.00000	5.076904	.967
	2nd Year	165	.00000	5.794375	
	3rd Year	199	.00000	5.831414	
	Total	500	.00000	5.612566	
nouns_as_modifiers	1st Year	136	.00000	3.071651	.988
	2nd Year	165	.00000	3.091648	
	3rd Year	199	.00000	3.119858	
	Total	500	.00000	3.091304	
determiners	1st Year	136	.00000	3.043116	.861
	2nd Year	165	.00000	3.093616	
	3rd Year	199	.00000	3.097435	
	Total	500	.00000	3.075328	
possessives	1st Year	136	.00000	2.248428	.974
	2nd Year	165	.00000	2.144256	
	3rd Year	199	.00000	2.217277	
	Total	500	.00000	2.197630	

3.2 Verb-VAC complexity

Another indicator for syntactic development is the complexity of verb argument constructions (VAC for short), which is composed of indices like verb-VAC frequency, VAC frequency, association strength, frequency, and diversity and frequency, as shown in Table 2. According to Kyle ([17]), verb-VAC frequency captures verb and verb–VAC frequency. The index of VAC frequency captures the frequency of VACs, the incidence of direct objects, and the incidence of direct object dependents. Association strength captures the main verb lemma–VAC association strength. The index of diversity and frequency captures the diversity of VACs, main verb lemmas, and main verb lemma–VAC combinations. It also captures the main verb lemma–VAC frequency. The index of frequency captures VAC, main verb lemma, and main verb lemma–VAC combination frequency.

Table 2. Descriptive statistics and results of Kruskal-Wallis tests for verb-VAC complexity (by year)

Variables	Year	N	Mean	Std. Deviation	Kruskal-Wallis test (Sig.)
verb_vac_frequency	1st Year	136	.00000	4.994866	.984
	2nd Year	165	.00000	5.084341	
	3rd Year	199	.00000	4.768725	
	Total	500	.00000	4.926356	
vac_frequency	1st Year	136	.00000	2.664405	.983
	2nd Year	165	.00000	2.625313	
	3rd Year	199	.00000	2.567961	
	Total	500	.00000	2.608142	
association_strength	1st Year	136	.00000	2.456159	.314
	2nd Year	165	.00000	2.369969	
	3rd Year	199	.00000	2.530135	
	Total	500	.00000	2.453200	
frequency	1st Year	136	.00000	1.919116	.990
	2nd Year	165	.00000	2.107946	
	3rd Year	199	.00000	1.813133	
	Total	500	.00000	1.939385	
diversity_and_ frequency	1st Year	136	.00000	1.957720	.996
	2nd Year	165	.00000	2.046564	
	3rd Year	199	.00000	2.074720	
	Total	500	.00000	2.030132	

The results of Kruskal-Wallis tests show that there is no significant difference for all the indices under VAC complexity, which further implies that the indices for the complexity of verbs and VACs do not predict beginner writing quality and they are not applicable for the development of beginner language. A possible explanation is beginners' lack of knowledge of the uses of most commonly used verb lemmas, such words as "take", "do", "have", "make", "put", and others. The English language is not beginner-friendly, as the simple, short, and daily words are loaded with multiple uses and functions, which poses a lot of difficulty for EFL beginners. In the case of a common word such as "have", we recognize more than 10 types of VACs, while in the learner corpus most types are rarely noticed and most of the sentences (88.5%) belong to the type "nominal subject-v-direct object". It is generally believed that constructions are embodiments of stereotypes, which has nothing to do with language proficiency, but it seems that, in constructing the right constructions, grammatical and syntactic knowledge of some complex combinations takes more time to settle in for EFL learners. Beginners have to familiarize themselves with the uses of coordinating conjunctions, prepositions, adverbial modifiers, and adverbial clausal modifiers, so that they may have confidence in producing the various constructions to achieve complexity in their use of VACs.

3.3 Unit length

Based on the study by Lu (2010), three indices of unit length for syntactic complexity are taken in the present study, namely mean length of sentence (MLS), mean length of T-unit

(MLT), and mean length of clause (MLC). As shown in Table 4, the mean values for the three indices are growing over the years, which indicates that beginners are really making progress in this regard. And the values for standard deviation are also growing for the three indices over the three years, which displays that individual differences are widening greatly.

The results of Kruskal-Wallis tests (Table 3) display that there are significant differences for the three indices (Sig.=0.000). Pairwise comparisons were employed to determine the differences by year, the result of which illustrate that there are significant differences for all the three year groups of both MLS and MLC (Sig.<0.05). For MLT, there are significant differences for the 1st-3rd year group and the 2nd-3rd year group (Sig.=0.000), while for the 1st-2nd year group, there is no significant difference (Sig.=0.329).

The results of pairwise comparisons also demonstrate that the three indices of unit length (mean length of sentence, mean length of T-unit, and mean length of clause) are reliable indicators for predicting the development of beginner writing. It can be concluded that, with the passage of time, beginners are writing longer compositions, longer sentences, and longer T-units. As the data are from guided writing in exams, the overall length of students' compositions is mainly determined by what is stated in the directions. Over the three years, the requirement for the minimum words for writing increased from 100 in the first year to 150 words in the third. In order to achieve this purpose, a learner should be working on his skills in paragraph development; he has to learn to transfer or just translate what he has already known of the languages he is familiar, into new situations. In this process, what he really needs is concrete knowledge of the core vocabulary, the most commonly used 3000 words, to write about the things and events around him under the sun. When knowledge of the core vocabulary is combined with basic syntactic rules, a beginner is equipped with the necessary know-how to express himself literally in a new language. In the case of longer T-units and longer clauses, they are mainly realized by having modifiers of various kinds or having parallel constructions, or both. Syntactically, a beginner's job is to extend his "I read many books" into "I could do my homework and read many books" or "I could do my homework and read many books in the library when I did not have classes".

Table 3. Descriptive statistics, results of Kruskal-Wallis tests, and pairwise comparisons for unit length (by year)

Variables	Year	N	Mean	Std. Deviation
MLS	1st Year	136	10.1347	3.85345
	2nd Year	165	11.8007	4.67945
	3rd Year	199	14.4848	5.87135
	Total	500	12.4158	5.30500
MLT	1st Year	136	10.60802	3.504779
	2nd Year	165	11.27925	3.923416
	3rd Year	199	14.61826	5.377148
	Total	500	12.42560	4.807003
MLC	1st Year	136	7.40349	1.519288
	2nd Year	165	7.87149	1.556686
	3rd Year	199	9.39443	2.088471
	Total	500	8.35032	1.976487

3.4 Clausal complexity

According to Lu ([18]), the present study takes eleven indices for clausal complexity, including clauses per sentence (C_S), verb phrases per T-unit (VP_T), clauses per T-unit (C_T), dependent clauses per clause (DC_C), dependent clauses per T-unit (DC_T), T-units per sentence (T_S), complex T-unit ratio (CT_T), coordinate phrases per T-unit (CP_T), coordinate phrases per clause (CP_C), complex nominals per T-unit (CN_T) and complex nominals per clause (CN_C). The descriptive statistics for the indices are shown below in Table 5, from which we can see the mean values for most of the indices (with the exceptions of VP_T, C_T, T_S, and CT_T) are growing from the first year to the third. It indicates that learners are having more complex clauses in their writing. Meanwhile, the values for standard deviation are increasing over the years (with the exceptions of VP_T, C_T, T_S, and CT_T), which indicates that individual differences are widening from the first year to the third year in terms of clausal complexity. The results of Kruskal-Wallis tests (shown in Table 4) indicate that there are significant differences for all the indices but one (complex T-unit ratio, Sig.=0.211). And the results of further pairwise comparisons indicate variably that, for coordinate phrases per T-unit, there are significant differences for the three year-group pairs (Sig. <0.05); for half of the remaining indices: verb phrases per T-unit, clauses per T-unit, coordinate phrases per clause, complex nominals per T-unit and complex nominals per clause, there are significant differences noticed for the 1st-3rd year pair and 2nd-3rd year pair (Sig.<0.05); for three other indices (clauses per sentence, dependent clauses per clause and dependent clauses per T-unit), significances are only found for the 1st-3rd year pair (Sig.<0.05); for T-units per sentence, significant differences are noticed for the 1st-2nd year pair and the 1st-3rd year pair (Sig.<0.05).

Therefore, seven indices: coordinate phrases per T-unit, verb phrases per T-unit, clauses per T-unit, coordinate phrases per clause, complex nominals per T-unit, complex nominals per clause, and T-units per sentence, are applicable in predicting the development of beginner writing. Coordination, subordination, and modification highlight themselves as the skills that matter for beginners, which should be the focus of language teaching and learning for EFL beginners.

Table 4. Descriptive statistics, results of Kruskal-Wallis tests, and pairwise comparisons for clausal complexity (by year)

Variables	Year	N	Mean	Std. Deviation	Kruskal-Wallis test (Sig.)	Pairwise Comparisons		
						1st Year – 2nd Year (Sig.)	2nd Year – 3rd Year (Sig.)	1st Year – 3rd Year (Sig.)
C_S	1st Year	136	1.38534	.471894	.003*	.089	.741	.002*
	2nd Year	165	1.50455	.535233				
	3rd Year	199	1.57484	.644091				
	Total	500	1.50010	.569940				
VP_T	1st Year	136	1.77969	.540176	.000*	1.000	.000*	.000*
	2nd Year	165	1.75798	.507549				
	3rd Year	199	2.09521	.884150				
	Total	500	1.89810	.706820				
C_T	1st Year	136	1.44745	.435089	.010*	1.000	.026*	.036*
	2nd Year	165	1.43228	.373447				
	3rd Year	199	1.58106	.570670				
	Total	500	1.49562	.480772				
DC_C	1st Year	136	.25066	.110745	.036*	.398	.848	.030*
	2nd Year	165	.27007	.111422				
	3rd Year	199	.28749	.146592				
	Total	500	.27172	.127067				
DC_T	1st Year	136	.39156	.284635	.028*	.852	.332	.026*
	2nd Year	165	.41555	.264258				
	3rd Year	199	.51057	.453924				
	Total	500	.44685	.359727				
T_S	1st Year	136	.96288	.206830	.000*	.000*	.075	.046*
	2nd Year	165	1.04395	.208280				
	3rd Year	199	.98922	.180641				
	Total	500	1.00012	.199597				
CT_T	1st Year	136	.27802	.155612	.211	---	---	---
	2nd Year	165	.27613	.141586				
	3rd Year	199	.30706	.178859				
	Total	500	.28895	.161403				
CP_T	1st Year	136	.17239	.176785	.000*	.044*	.000*	.000*
	2nd Year	165	.22518	.200496				
	3rd Year	199	.32302	.252077				
	Total	500	.24976	.225503				
CP_C	1st Year	136	.12014	.091754	.000*	.077	.002*	.000*
	2nd Year	165	.15641	.128268				
	3rd Year	199	.20845	.156890				
	Total	500	.16726	.137014				
CN_T	1st Year	136	1.04821	.524601	.000*	.479	.000*	.000*
	2nd Year	165	1.14204	.586235				
	3rd Year	199	1.81462	.963594				
	Total	500	1.38421	.824544				
CN_C	1st Year	136	.71355	.242758	.000*	.159	.000*	.000*
	2nd Year	165	.78478	.296342				
	3rd Year	199	1.14536	.414580				
	Total	500	.90892	.388328				

4 Conclusion

Among the various indicators used in the academic community, some indices, especially the conventional unit length indicators and indices for clausal complexity are more applicable for predicting the development of beginner writing. In terms of unit length, the three indicators of mean length of sentence, mean length of T-unit, and mean length of clause can be served as the main descriptive variables for the development of the language of beginners; for clausal complexity, seven indices: coordinate phrases per T-unit, verb phrases per T-unit, clauses per T-unit, coordinate phrases per clause, complex nominals per T-unit, complex nominals per clause and T-units per sentence, are the reliable indicators for beginner writing development. But indices for noun phrase complexity and verb-VAC complexity show no significant difference in the Kruskal-Wallis tests.

The results indicate that the major task for beginners is to enhance the complexity of their language use via coordination, subordination, and modification. Modification requires the knowledge and mastery of specific words, collocations, and idiomatic usage. Coordination calls more for knowledge of semantic relations, such as synonyms and antonyms. The use of similar forms for parallel construction also plays an important role. Subordination depends on the learner's knowledge of clauses for logical connection and effective communication. Some of the techniques and means for achieving these purposes may seem to pose some difficulty for beginners, which includes the correct use of prepositional phrases and infinitive forms of verbs. Therefore, classroom teaching and syllabus design should pay more attention to the teaching of vocabulary, phrases, and the expansion of clause structure.

Acknowledgments

The present research is funded by the Innovative Practice Base for Developmental Integration of Information Technology with Foreign Languages Teaching and Research and the National Social Science Fund of China (17BYY042, A Typological Study of the Function-order Interactions of the Modifiers of English and Chinese).

References

1. K. Kyle, S. Crossley, Assessing syntactic sophistication in L2 writing: A usage-based approach. Language Testing, 34(4), 513-535 (2017)
2. K. Kyle, S. A. Crossley, S. Jarvis, Assessing the validity of lexical diversity indices using direct judgements. Language Assessment Quarterly, 18(2), 154-170 (2021)
3. K. Kyle, S. Crossley, M. Verspoor, Measuring longitudinal writing development using indices of syntactic complexity and sophistication. Studies in Second Language Acquisition, 43(4), 781-812 (2021)
4. T. Mostafa, S. A. Crossley, Verb argument construction complexity indices and L2 writing quality: Effects of writing tasks and prompts, Journal of Second Language Writing (2020)
5. J. E. Casal, J. J. Lee, Syntactic complexity and writing quality in assessed first-year L2 writing, Journal of Second Language Writing (2019)
6. J. Jiang, P. Bi, H. Liu, Syntactic complexity development in the writings of EFL learners: Insights from a dependency syntactically-annotated corpus, Journal of Second Language Writing (2019)

7. D. Mazgutova, J. Kormos, Syntactic and lexical development in an intensive English for Academic Purposes programme, Journal of second language writing (2015)
8. X. Wu, A. Mauranen, L. Lei, Syntactic complexity in English as a lingua franca academic writing, Journal of English for Academic Purposes (2020)
9. K. Uzun, Performance prediction strengths of noun and verb phrases in L2 writing: Comparison of density and complexity variables, Assessing Writing (2021)
10. R. Khany, N. B. Kafshgar, Analysing texts through their linguistic properties: A cross-disciplinary study, Journal of Quantitative Linguistics (2016)
11. D. Biber, B. Gray, S. Staples, Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement, Journal of English for Academic Purposes (2020)
12. D. Man, M. H. Chau, Learning to evaluate through that-clauses: Evidence from a longitudinal learner corpus, Journal of English for Academic Purposes, 37, pp. 22-33 (2019)
13. X. Lu, H. Ai, Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds, Journal of second language writing, 29, pp. 16-27 (2015)
14. T. Brunner, Structural nativization, typology and complexity: noun phrase structures in British, Kenyan and Singaporean English, English Language & Linguistics, 18(1), pp. 23-48 (2014)
15. M. Nasseri, Is postgraduate English academic writing more clausal or phrasal? Syntactic complexification at the crossroads of genre, proficiency, and statistical modelling. Journal of English for Academic Purposes, 49 (2021)
16. S. Staples, R. Reppen, Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings, Journal of Second Language Writing (2016)
17. K. Kyle, Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication (2016)
18. X. Lu, Automatic analysis of syntactic complexity in second language writing, International journal of corpus linguistics, 15(4), pp. 474-496 (2010)