Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 619-631
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

Design and Data Analysis of Diachronic Corpus

Changzhong Shao
School of Foreign Language, Linyi University, Linyi, Shandong, China

Abstract:

A diachronic corpus with 93 English as foreign language (EFL) learners in China was designed and built, and the data process and analysis of the corpus was achieved with the software Writer's Workbench, which incorporates the various tool types and optimum algorithms. The explored techniques are of great importance in mining data and finding pattern in EFL learners' writing in China.

Keywords*: Data analysis, diachronic corpus, EFL learners in China.*

I. INTRODUCTION

English writing has always been one of the major professional skills for English learners. English Teaching Syllabus for English learners in Colleges and Universities was published in March 2000 in China, which requires more attention to be paid to the comprehensive development of listening, speaking, reading and writing. At the same time, it emphasizes that we are supposed mainly to improve the capabilities of reading, writing and translation. As for Chinese college students who learned English as a foreign language (EFL), the writing skill has been a crucial index of their language proficiency development. English writing is also one of the difficult language skills for them. However, it is still difficult to write excellent articles in spite that they have affluent reading volume and vocabulary, therefore students' writing competence is still far from satisfactory.

In the writing of the second language, sentence complexity has been considered as the indicator in the study of English learners' language proficiency development. In terms of sentence complexity, it includes a wide range of indexes which can be used for assessing the writing development. Therefore, we will study sentence complexity in three aspects which involve in the average length of sentence, the percentage of sentence types and the beginning of sentences. The reason we choose the three aspects is that they can reflect most directly the writing characteristics of English learners.

We can draw the conclusion from the writing corpus of English majors that the complex sentence is limited in writing, the alternate use of long and short sentences is not fluent enough, and the sentence beginning is still dull and simple sentences are overused.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 619-631
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

The corpus provides a real language virtual space with foreign language learners. It can provide us the useful way to improve the writing skill by truly reflect students' problems in writing. This study will analyze data from the corpus, and then analyze the features of sentence structure from the English as a foreign language learner. We can find from the previous study that there are wide gaps existing in sentence complexity, therefore we can make further progress in the investigation of sentence structure. The innovation in this research is that we can not only narrow the gap existing in the previous study, but the study itself can be a challenge for traditional teaching methods. The objective of this thesis is to study the sentence development of English writing, which can provide a new method for English as foreign language learners to improve their writing skills.

II. LITERATURE REVIEW

The learner's writing competence is affected by many factors, and the sentence structure plays a crucial role in the writing quality. To improve the writing skill of L2 learners, many researches have been made by the scholars in China. Although great progress has been achieved, there are still many undiscovered fields we can explore.

As a crucial indicate of the complex sentence structure, the average length of sentence always plays an important role for the development of L2 learners. Based on the corpus of Chinese and American students' writing, the results show that the average length of writing of Chinese students' is lower than that of American students', and there are great differences in the distribution of different sentences length [1]. It is because Chinese students are not good at using complex structures as well as they are not skilled enough in using foreign languages. In addition, their thinking is restricted by traditional teaching models.

Besides the average length of sentence, the beginning of the sentence also determines the quality of the writing. For English learners, they are mostly beginning with the subject in English writing. This type of beginning tends to be monotonous and dull, making the article unattractive [2]. Therefore, to increase the diversity of sentence beginning can rich the expression of writings.

Language features are also one of the manifestations of sentence structure. The comparison of language features of Chinese and U.S. college compositions show that significant differences are existed in the use of the nine language features in their writings [3]. First of all, it is normal that the average length of sentence of American's is longer than Chinese students' [4]. However, the previous study indicates that the sentence length is highly related to the writing quality [5].

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 619-631
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

Unit length, unit density and the frequency of occurrence of various sentence patterns are also the basic features of sentence structure. Unit length refers to the average number of output units, the most common is W/T and W/C. T-unit means which contains a main sentence, all the additional clauses and non- clause structures and their indivisible minimum unit [6]. Clause refers to any subject-verb structure. Length of clause is the mainly characteristic of academic English [7], and it is one of the most effective variables to distinguish language levels [8]. Unit density is related to complexity ratio and subordinate sentence ratio, when the T unit includes more clause, the sentence structure is more complex.

Besides, passive sentences also belong to complex sentences, the use of passive sentences is a sign of higher writing capability. Complex sentences also include phrase contraction clause, such as adverb phrase, adjective phrase and nominal verb phrase. The top-level learners often use contraction clause [9]. The contraction clause can consolidate the meaning of several clauses into one sentence, thus increases the information and length of sentences.

Since 1960s, many scholars in other countries have also studied sentence structure already. Compared to L1 writing, their findings of sentence complexity are successfully applied to L2 writing.

The learner's writing competence is affected by many factors, and the sentence structure plays a crucial role in the writing quality. Grabe & Kaplan divide the capability of writing into three aspects: language knowledge, textual knowledge and sociolinguistic knowledge [10].

As for language knowledge, sentence structure is one of the basic elements of English writing, whether can English learners write a variety of sentence structure in the writing or not determines the quality of the article. However, to write more complex sentences maybe is difficult for L2 learners. The author found that better writers can write longer complex sentences, pause for shorter durations and at clause boundaries more often than poor writers [11].

Some scholars define complexity as lexical complexity, grammatical complexity and sentence complexity, while other scholars put the view that vocabulary and grammatical complexity are two different aspects of language output, and some scholars even consider the sentence complexity itself contains different subclasses [12]. Besides, sentence complexity has different manifestations. For example, in terms of structure, sentence complexity is showing compound sentences or complex sentences, the sign of complexity can also be variety.

From many studies of other experts, the quality of article depends on sentence complexity, which refers to sentence diversity and complexity [13]. Sentence diversity refers to the range of changes in sentence patterns, and it is the flexible use of various sentence patterns; complexity refers to the sentence complexity [14], subordinate clause, unqualified verb phrase, complex verb phrase, complex noun

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 619-631
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

phrases and nominal sentences, all of these are complex sentence patterns.

III. RESEARCH METHODOLOGY

This paper studies sentence complexity development about the writing of English as foreign language learners in China. This part mainly focuses on the research methodology of the study including the research questions, the sources of corpus, the related tool and software and procedures.

3.1 Research Questions

Drawing on previous research and achievements, we are trying to discuss the following three issues:

Question 1: Does the average length of sentence for English majors' writing develop with a linear tendency?

Question 2: Does the sentence types become more complex in writing of English majors?

Question 3: Does the beginning of sentence change more obvious in English major's writing?

3.2 The Corpus Employed in This Study

The corpus in this study is a self-built Diachronic Writing Corpus for Chinese English Majors. The data for this study consists a total of 411 compositions of the English major's final exam, which including 6 terms writing task of 3 years chosen by Business English and English Translation. In this corpus, the number of learners majoring in Business English is 45, and the number of learners majoring in English Translation is 48, totally 93 learners.

As the second language learners, the English learning experience of these students is generally the same with the influence of China's education system. Every one of them has almost 10 years' experience of English learning. They have engaged in English class from the third grade in their primary schools, later everyone is beginning to learning simple grammar when they are in middle school. After entering high school, the difficulty of learning English has also increased. In the end they passed the college entrance examination and joined English majors with almost equivalent English scores.

Unlike the previous synchronic studies, these are diachronic corpus writing materials of English majors. The focus of the study is to test the students' English proficiency and errors they have made over the past three years, thus to improve their writing skills.

In this study, the corpus is divided into three grades/six terms sub-corpora according to learner's exam writing: Grade 1, Grade 2, Grade 3; Term 1, Term 2, Term 3, Term 4, Term 5, Term 6. Just as showed in the following Table 1.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 619-631
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

TABLE I. Sources of Corpus

411 compositions	Grade 1 142 compositions	Term 1(29+36)
		Term 2(40+37)
	Grade 2 134 compositions	Term 1(37+28)
		Term 2(36+33)
	Grade 3 145 compositions	Term 1(33+31)
		Term 2(33+38)

3.3 Research Tool and Software

The retrieving tool employed in this study is Writer's Workbench (http//www.emo com/order/orderfront.htm). By using this tool, we can research the sentence complexity, the average length of the sentences, the type of different sentences and the proportion of sentence beginning.

Writer's Workbench (Hereinafter referred to as WWB) is a set of text analysis programs developed by AT&T Bell Labs. The software is installed as a plugin for Microsoft Office Word, which can be used with word software to analyze the corpus directly. In this way we can provide timely feedback on writing texts. The software is powerful with many useful applications, such as teaching function and research function. As for teaching function, WWB can not only assist teachers in writing teaching but can help students improve their writing skills. On the other hand, there are also researchers or company staff who use the software to modify academic papers or documents, thus to improve the quality of papers and documents. The advantage of WWB is that it can give a personalized analysis and guidance to the user's writing text. The analysis capabilities of WWB covered in articles, language features and many other aspects. When the users finish the draft, the text can be modified and edited. The structures and ideas of the article can be examined as well as the sentence structures, words, formats, segmentation and other language features.

As for WWB, it consists of 25 analysis programs in 6 categories: content, characteristics, verbs, clarity, words, punctuation. This study mainly focuses on the features of sentence, so we mainly use the Characteristic Style Statistics with Support function.

3.4 Procedure

According to the questions mentioned above, the analysis of sentence complexity in Chinese English major's argumentative writings are presented by the following procedures:

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 619-631
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

First of all, we install WWB software on our laptop desktop and fix this software into the word document, then we open the writings of the corpus in WWB software. In terms of sentence structure, the function which we have used is Characteristic Style Statistics with Support. We put all the writings into Characteristic Style Statistics with Support one by one, and we can get the relevant variables. Then the data will be collected into computer and analyzed by the WWB software to show the differences in writings among six terms. Next, the data is used in the comparison to evaluate whether the sentence complexity of English majors' writing develops with linear tendency or not. Furthermore, the study of sentence complexity will be explored. Therefore, the sentence complexity including the average length of sentences, the total sentence number of the writing, the beginning of the sentence and the sentence types can be researched.

IV. DATA COLLECTION AND DISCUSSION

In this part, the development trend of sentence complexity will be researched. The results of the six terms during three grades will be presented consecutively. Followed with the discussion on the changes of sentence complexity, the possible causes are illustrated accordingly.

4.1 The Average Length of the Sentences

As mentioned in Part Three, the sentence complexity can be judged from three aspects: the average length of the sentence, the type of sentence and the beginning of sentence. In this part, we will discuss the changes in the average length of sentences because the average length of sentences is an important indicator of the sentence complexity.

We will analyze sentences from four aspects: the average number of sentences, the average sentence length, percentage of sentences with 5 or more words shorter than average sentence length, and percentage of sentences with 10 or more words longer than average sentence length.

TABLE II. The Results of Average Sentence Length in the First Term

	ANS	**ASL**	**5 less than ASL %**	**10 more than ASL %**
Class 2	17	9.3	13%	5.3%
Class 9	17	9.7	16%	6%

(ANS: the average number of sentences; ASL: the average sentence length)

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 619-631
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

TABLE III. The Results of Average Sentence Length in the Second Term

	ANS	ASL	5 less than ASL %	10 more than ASL %
Class 2	17	11.1	23%	8.5%
Class 9	19	10.2	13%	5%

TABLE IV. The Results of Average Sentence Length in the Third Term

	ANS	ASL	5 less than ASL %	10 more than ASL %
Class 2	24	9.6	22%	7%
Class 9	26	9.2	12%	3%

TABLE V. The Results of Average Sentence Length in the Forth Term

	ANS	ASL	5 less than ASL %	10 more than ASL %
Class 2	23	11.2	22%	10%
Class 9	29	9.1	11%	4%

TABLE VI. The Results of Average Sentence Length in the Fifth Term

	ANS	ASL	5 less than ASL %	10 more than ASL %
Class 2	26	9	10%	5%
Class 9	28	7.4	2%	3%

TABLE VII. The Results of Average Sentence Length in the Sixth Term

	ANS	ASL	5 less than ASL %	10 more than ASL %
Class 2	21	10.3	15%	4%
Class 9	25	9.3	10%	3%

The 6 tables of data show the data of final composition of college English majors in three years or six semesters. As we can find that the average number of sentences shows a rising trend, but it is declined in the sixth semester. As for the average sentence length of the two classes, the overall trend is fluctuant rising. And the percentage of sentences with 5 or more words shorter than the average sentence length is unstable. In the first term, the data of class 2 is 13%, and class 9 is 16%, the data of the second term are 23% and 13%, the data of the third term are 22% and 11%, the data of the fourth data are 22% and 11%, the data of the fifth term are 10% and 2%, the data of the sixth term are 15% and 10%. While the and

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 619-631
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

percentage of sentences with 10 or more words longer than the average sentence length is increased in the first two term, and in the last four terms, it is declined in a fluctuant trend.

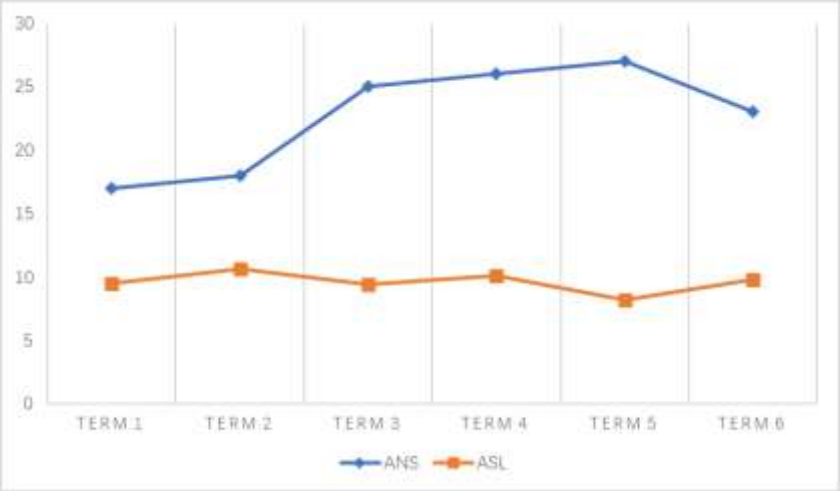

Figure 1: the average length of the sentence

The line with rhombus in Figure 1 is the average number of sentences of Class 2 and Class 9, while the line with square is the average sentence length. We can find from Figure 1 that the average sentence number of English majors' writing showed in a linear growth trend in the six terms besides the declined in the last term, and the sentence length showed a generally stable trend. This reflects that the student's writing competence is improved with the increase of grade. And the descend in the last term may indicate that students improve in their sentence type, thus to reduce the sentence numbers. The steady increase in sentence length indicates that students are relatively single in using sentence patterns and the complex sentence and compound sentence are also rare. Sentences in writings usually vary considerably in length. A lack of variety in sentence length may result in monotony. Hence the students should work to balance medium length sentences with shorter and longer sentences.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 619-631
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

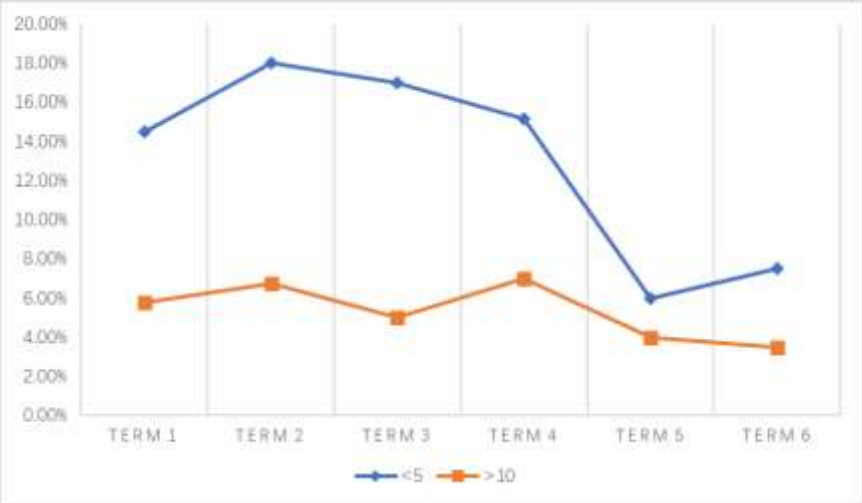

Figure 2: the average length of the sentence

In Figure 2, we can see that the sentence with 5 words less than the average sentence is raising in the first two terms, but the data of the first four semesters remained basically stable. This line indicates that the use of connective words is increasing, and the fluency between sentence and sentence is improving. But in the fifth term, it falls as the number of sentences increases. As for the sentences with 10 words more than the average sentence is generally increases in the first four terms. When the compound and complex sentences are raising, their writing competence is increasing as well. But it declines in the last two years, which results into the necessity that the writing competence by combining more types of sentences should be strengthened.

4.2 Percentage of Sentence Types

We will study the sentence types in this part. We mainly do research on the different percentage of simple sentence and complex sentence.

Table VIII shows the specific percentage of simple sentences in the two classes in six terms, while Tabe IX show the counterparts of complex sentences.

TABLE VIII. Simple Sentence

	Term 1 %	**Term 2 %**	**Term 3 %**	**Term 4 %**	**Term 5 %**	**Term 6 %**
Class 2	89	88	91	91	93	90
Class 9	81	88	92	89	97	87

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 619-631
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

TABLE IX. Complex Sentence

	Term 1 %	Term 2 %	Term 3 %	Term 4 %	Term 5 %	Term 6 %
Class 2	11	12	9	9	**7**	10
Class 9	19	12	8	11	3	13

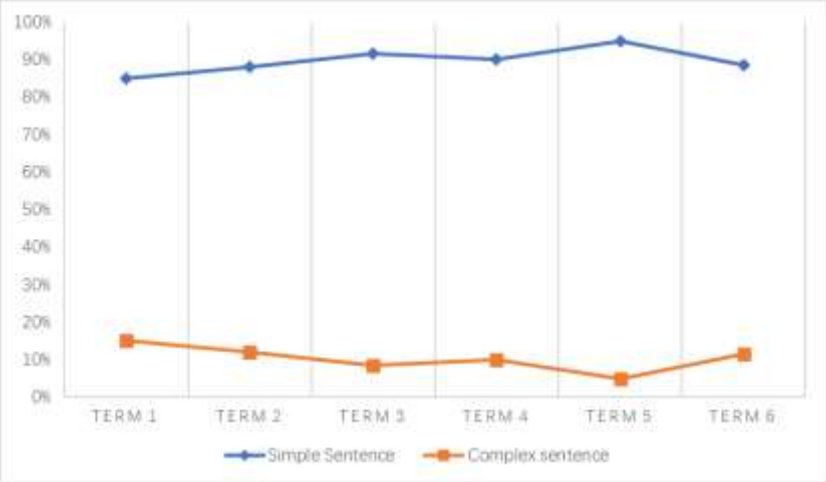

Figure 3: the sentence types

Figure 3 shows that the number of simple sentences develops linearly in the first five terms, while in the last term, the number of simple sentences is declining. In terms of complex sentence, its growth trend is unstable. It declines in the first three terms, while in the fourth term, it is higher than the third term. However, it is significantly decreased in the fifth term. But in the last term, it also showed a significant upward trend.

From Figure 3, it is also found that the large number of simple sentences indicates that students use too many simple sentences in their writing, which results in the accumulation of simple sentences. This will lead to confusion in the internal logic of the sentence, and the meaning of the sentence is not clear. To avoid this phenomenon, we can use conjunctions to merge the simple sentences and form a primary and secondary relations. And the sentence level will be more distinct as well as the structure will be more rigorous in this way, thus the sentence form will be more diversified.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 619-631
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

4.3 The Beginning of the Sentence

Many students tend to use nous and pronouns as the beginning in their writing, but this kind of opening will appear particularly single and dull. In this part, we will study the beginning of writing in terms of beginning with subjects and beginning with non-subjects.

Table X shows the numbers of sentences beginning with subjects, while Tabel VI, the numbers of sentences with non-subjects.

TABLE X. Sentence Beginning with Subjects

	Term 1 %	**Term 2 %**	**Term 3 %**	**Term 4 %**	**Term 5 %**	**Term 6 %**
Class 2	59	61	61	67	57	57
Class 9	57	72	61	63	51	58

TABLE VI. Sentence Beginning with Non-Subjects

	Term 1 %	**Term 2 %**	**Term 3 %**	**Term 4 %**	**Term 5 %**	**Term 6 %**
Class 2	41	39	39	33	43	43
Class 9	43	28	39	37	49	42

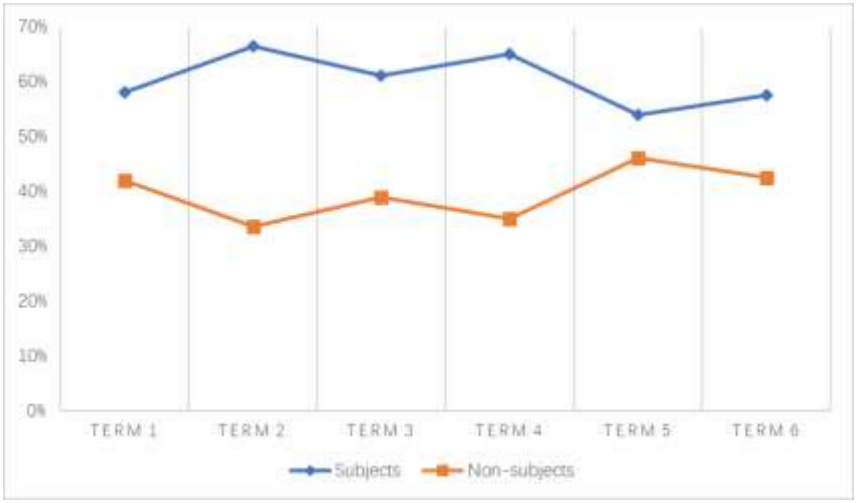

Figure 4: The Sentence Beginning

We can see from Figure 4 that the beginning of the sentence is fluctuating and unstable, the percentage of subjects beginning in the six terms are 58%, 66.5%, 61%, 65%, 54%, and 57.5%, and the percentage of

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 619-631
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

non-subjects beginning are 42%, 33.5%, 39%, 35%, 46%,42.5%. The figure shows that Chinese students still prefer sentences that begin with the subject. In fact, we can change the beginning in many ways in order to make the sentence more lively and vivid. For example, we can use adverbs, apposition, adverbial, predicative, object, and phrase as the beginnings of sentences.

V. CONCLUSION

This study is mainly data analysis of the diachronic corpus of college English as foreign language learner's writing. By using the Characteristic Style Statistics with Support function of WWB software, we can study the sentence development of the EFL learners' writing as well as finding different variables among different terms to improve the learners' writing competence.

With the research of the EFL learners' writing in China writing in six terms with the WWB software, we can find that the sentence complexity of the EFL learners in China is improved, and their writing competence also made corresponding progress. But the progress is imbalanced, and the learners' writing competence is also unstable. The sentence structure is becoming more complex in the first four terms; however, the sentence is tending simple in the last two terms.

Although the average sentence number is increasing in the three years, the number of complex sentences is also rare, and the beginning of sentence is still dull and single. The sentence number of each writing is increased with the growth of grade and its development has basically grown linearly. In term of sentence types, the using of complex sentence is not stable. In the first four terms, the number of simple sentences is increased while the number of complex sentences is decreased, the proportion of simple sentences is still high. Which indicates that students should improve their competence to combine different sentences, thus increase the variety of sentence patterns. In the respect of the sentence beginning, the subjects beginning of sentence is far more than the non-subjects beginning sentence. And students' progress in changing the beginning of a sentence is not obvious.

In the three aspects of sentence complexity, the average length of sentence is obviously increased, and the sentence type is still simple, students' writing beginning is little improved. The sentence type and a better beginning is crucial about the writing quality. This means that the EFL learners in China still have a large room in improving sentence complexity in their English writing.

Considering the number of the EFL learners in the corpus, we are not sure whether the selected sample is representative or not. The relationship between the elements of the corpus has not yet been measured. The data of this study can merely provide a comparative exploration for the follow-up researches, and the exploratory conclusions are expected more scholars to study sentence complexity in the future.

Forest Chemicals Review
www.forestchemicalsreview.com
ISSN: 1520-0191
September-October 2021 Page No. 619-631
Article History: Received: 22 July 2021 Revised: 16 August 2021 Accepted: 05 September 2021 Publication: 31 October 2021

The future research requires the addition of different topics, different conditions of writing, and larger samples of corpus, and it is necessary to study the groups of learners and also to study the individual case so as to obtain more comprehensive and applicable data.

ACKNOWLEDGEMENTS

This research was supported by 2019-2020 Excellent Course Foundation of Linyi University (Grant No. JPKT1911).

REFERENCES

[1] Hilton, H. (2008) The link between vocabulary knowledge and spoken L2 fluency. Language Learning Journal 36 (2): 153-66
[2] Ortega, L. (2003) Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. Applied linguistics 24: 492-518
[3] Wang, Y. X. (2017) The correlation between lexical richness and writing achievements of Chinese L2 learners. Applied Linguistics 66: 93-101
[4] Ortege, L. & H. (2008) *The Longitudinal Study of Advanced L2 Capacities* (London and New York: Routledge)
[5] Freed, B. (2000) If Fluency, Like Beauty, the Eyes, of the Beholder? H. Riggenback (ed.) Perspective on Fluency. Ann Arbor: The University of Michigan Press. ISBN9780472086047
[6] Liu, B. & Wang, Y. K. (2015) A Study on use of adjectives in English writing by non-English major postgraduates. Chinese Foreign Language 66(4): 45-53
[7] Wang, C. Y. (2019) Using Pedagogic Intervention to Cultivate Contextual Lexical Competence in L2: An Investigation of Chinese English Majors. NY: Palgrave Macmillan. ISBN 978-3-319-92715-2, ISBN978-3-319-92716-9
[8] Ellis, N. C. (1998) Emergentism, Connectionism and Language Learning. Language Learning 48: 31-664.
[9] Lennon, P. (1990) Investigating Fluency in EFL: A Quantitative Approach. Language Leaning 40: 387-417.
[10] Grabe, W & R. B. Kaplan. (1996) Theory and Practice of Writing: An Applied Linguistic Perspective. London: Longman. ISBN-13: 978-0582553835, ISBN-10: 0582553830
[11] Plakans, L. & A. Gebril. (2017) An assessment perspective on argumentation in writing. Journal of Second Language Writing 36: 85-86
[12] Wolfe-Quintero, K., S. Inagaki & H. Kim. (1998) Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity. Honolulu: University of Hawaii Press. ISBN 082482069X 9780824820695
[13] Kyle, K. & S. Crossley. (2016) The relationship between lexical sophistication and independent and source-based writing. Journal of Second Language Writing 34: 12-24
[14] James, C. (2013) Errors in Language Learning and Use: Exploring Error Analysis. London and New York: Routledge. ISBN9780582257634