A Corpus-based Exploration into the Sentence Development in EFL Writing in China

Chang-zhong SHAO
School of Foreign Languages, Linyi University, China

Keywords: Corpus, Sentence development, EFL in China.

Abstract. Based on a self-built corpus, this research aims to explore whether the English writing proficiency in sentences of college English majors in China has been improved according to the mean length of sentences, the usage of coordinating conjunctions and subordinating conjunctions. WordSmith 4.0 is employed as the retrieval program. The survey found that the subjects do significant improvement in their compositions in the aspects of the mean length of sentences and the usages of subordinating conjunctions. No obvious changes of the usages of coordinating conjunctions have been observed. The variety of conjunctions is not enough.

Introduction

Among the basic listening, speaking, reading, and writing skills, writing is usually reckoned to be a significant part, which has been considered as one of the hardest parts in English. And it is also seen to be an effective measure to identify one’s English proficiencies. Writing plays a key role not only in language learning, but also in all kinds of English examinations. Take the TEM4 (Text for English Major) for example, writing section occupies 25% of the total scores. Unfortunately, writing has puzzled most language learners in China, even the English majors.

Following the international boom based on learner corpus, thus, Chinese scholars have laid more emphasis on L2 writing research. Literature review shows that two kinds of learner corpus-based studies: computer-assisted error analysis (CEA) and cross-sectional comparison between native and non-native speakers. [1] Other domestic research focuses on the lexical development and the function of lexical cohesive devices. Sentence, as the most fundamental unit in an article, affects the overall quality in a composition. The process of writing is composing sentence into discourse based on certain grammatical rules to achieve the goal of conveying the writers’ idea to the readers.[2] Studies on sentence, nevertheless, are great shortages.

It is well-known that Chinese students are weak at writing in TOEFL and IELTS. Numerous English majors deem that they have learnt nothing during their four- years’ undergraduate education.

This research based on a corpus made by ourselves aims to examine whether the English writing proficiency of college English majors in China moves forward in the perspective of sentences development. It studies on sentences in term of syntactic fluency and syntactic complexity. The standard of syntactic fluency is the mean length of sentence (MLS). As for syntactic complexity, we study on the coordination and subordination of sentences by the usage of conjunctions.

Syntactic complexity focuses on the range and sophistication degree of language output, which is illustrated by writing proficiency of writers to quickly gain access to and make full use of various types of sentences and complicated sentences.[3]

Sentence fluency is defined as a measure of sheer number of words or structural units a writer can include in their writing within a period. After a review of indices of fluency in the literature of writing, three best indices are concluded, named, T-unit length (total number of words divided by total number of T-units), error-free T-unit length (total number of words in error-free T-units divided by total number of error-free T-units).[4]

T-units is referred to as a minimal terminable unit which comprises a main and independent clause, be whatever subordinate clauses embedded in or attached to it.[3] Some scholars believe that T-units are the measure of syntactic complexity as well.
This research primarily explores the mean length of sentence, coordinating conjunction and subordinating conjunctions. Due to technical limitations, we cannot check into T-units. By combining the mean length of sentences with the conjunctions, however, we can do a study on T-units from the side.

This longitudinal study collects 182 English majors from Linyi University and their compositions as part of the final exam in 6 academic stages. Ultimately, 861 writing samples with an approximate number of 172,200 words are created to be a writing corpus. Among them, 30 articles are selected randomly as a sample. The main purpose of present research is to investigate whether English majors’ writing competence or language proficiency is improved.

**Literature Review**

In the past decades, numerous researches on L2 writing based on corpus have been done a lot. Linguists home and abroad have made a great achievement while imperfections in their studies also exist. The followings are the major publications made by foreign scholars.

A study statistically analyzes syntactic complexity embodied in writings in terms of essay and narrative by using words per clause (W/C) and clauses per T-units (C/T). The following is the finding: W/C is positively related to quality for argumentative essays but not for narratives. [5]

Some studies are conducted to explore relationship between the use of connectives and the writing quality, finding that in good essays there are more connectives.[6]

Empirical studies show that the use of connectives is especially problematic for EFL and ESL learners.[7]

A cross-linguistic study (Hong Kong University students comparing with American students) reveals that native English speakers tend to use some connectives like “after”, “before”, “however”, “so”, “still”, etc. more frequently although they use a wider variety of connectives.

All above shows that the usage of connectives is problematic for EFL learners and huge differences between native speakers and foreign language learners are existed. More attention should be paid on connectives in teaching and writing.

Chinese researchers have also come to realize the importance of sentences and conjunctions in English writing. A research on lexical and syntactic development finds that the curves for LD (lexical density), LV (lexical variation), AWL (Academic Word families) and Off-list families, together with MLT (mean length per T-unit) and MLC (mean length per clause), show learner increases. It also comes across that the syntactic multi-dimensional development is further empirically proven by the high and relations between MLT and VP/T (verb phrases per T-unit), MLT and CN/T (complex nominal per T-unit), MLT and C/T (clauses per T-unit), as well as MLT and DC/C (dependent clause per clause). [1]

When the sentences are organized coherently and the appropriate words are used in it, the process of communication between writer and reader will be quite smooth.[2]

A work based on grading 19 compositions by senior English majors, sums up 4 problems reflected in those compositions and two of them are: difficulty in producing long sentences and lack of complex sentences.[8]

The results mean that the mean length of sentence increases with the improvement of learners’ writing ability. Meanwhile, this change is complex and dynamic. Students, however, are especially weak at producing long sentences and complex sentences.

A study of cohesion features in the argumentative writing discovers that the usage of conjunction is the main cause to the quality of argumentative writings and much attention should be focused on how to teach the use of lexical cohesion and conjunction devices to improve students’ writing ability.[9]

A search on developmental features in L2 writing proficiency finds that the empirical test shows that the subjects do make statistically significant improvement in compositions in the aspects of fluency, lexical complexity and formality. [10]
There is a significant decrease in the density of cohesive devices at the higher proficiency level. [11] Apparently, just as the studies above show that the usages of effective cohesive devices and conjunctions can enhance the quality of the article.

While many studies conclude that Chinese learners of English are particularly weak at writing, these are, among other things, by no means empirical studies supported by data and statistical tests.[10]

The existing developmental studies done in China are, in fact, cross-sectional, due to their corpus-collection methods, which feature the comparison from students of different grades or with different language proficiency.[1]

To sum up, lacking of data and statistical tests support, too much cross-sectional studies and lacking researches on sentences are the main limitations in L2 writing researches.

Research Methodology

Research Questions
Based on the preceding theoretical knowledge and previous studies related to sentences in writing, the main objective of the study is to investigate if English majors in China move up in their writing proficiency in terms of sentence features such as syntactic complexity and syntactic fluency. To better understand the developmental feature of the writing across all levels, two sub-questions are listed as followed and serve as the guideline of the progress of the whole research:

1. Whether the EFL learners’ writing ability enhances according to the mean length of sentence?
2. How do the conjunctions change in the writing samples?

Data Collection
This longitudinal study chooses 182 English majors from Linyi University. Among them, 89 students (Business English: 43, Translation English: 46) come from Grade 2009 and 93 students (Business English: 45, Translation English: 48) come from Grade 2010. And their compositions as part of the final exam in 6 academic stages are collected. The topics of those 6 articles in sequence are: Students’ Pursuit for Famous Brand (120), Water Shortage (150), Holiday bargains: a trick to make you lose a pound for a penny (unknown), Package is Everything (250), Craze for Civil Service Examinations (200) and One Way to Solve Energy Crisis (unknown). It takes a week to collect the whole data. Data collectors input the compositions from the test paper to computer without any modification. Ultimately, 861 writing samples with an approximate number of 172,200 words are created to be a writing corpus.

Among the corpus, 30 English majors from Grade 2010 (Translation English) are selected randomly. Students whose compositions are missed in some semesters are eliminated. There is not significant different in writing strategy between male and female. So, gender can be ignored since English majors on a national scale are on the same condition. Finally, 180 compositions from their final test of 6 semesters are used as samples for examination. After the collection and selection of compositions, all the 180 texts are saved as text files.

In order to retrieve all the instances of English connectives and sentences, the present research employs WordSmith 4.0 as the retrieval program. In a word, Word-Smith (which means “word” and “creator”) is a paid software package primarily for linguists, for work in the field of corpus linguistics. It is a collection of modules for searching patterns in a language. The software is also available in several languages.

WordSmith 4.0 can operate under windows on any text files and get the specified research items from a corpus, show the overall occurrence number, list and highlight all the occurrence of the searched items in lines context (Wu, 2012, p.24). The pack-age provides 3 tools: Concord, Keywords and Wordlist. Each of the modules is offering several features available to certain other features of the analyzed text corpus. And Concord tool is mainly used in this thesis.
Concord is used to create concordances, so all the hits from a search within a previously defined body text. In present study, by using Concord, we can concord the particular words (connectives) we have chosen to study in batch.

As we have noted in Chapter 1, this thesis mainly investigates into the mean length of sentence, coordinating conjunction and subordinating conjunctions. The major functions of WordSimth are to search words, but it can also search punctuations.

We assume that it is a sentence, if a capital letter appears after a period, a question mark, a colon or an exclamation point. Thus, we use Concord tool to search period, question mark, colon and exclamation point and the number of sentences can be realized. Then, with the total number of words divided by the number of sentences, we can get the mean length of sentences.

Coordinating conjunction: 7 coordinating conjunctions are searched, they are: and, but, so, or, for, nor and yet.

As for subordination, the classification of the subordinate clause is: attributive clause, adverbial clause and noun clauses. And their main connections are as following:

Attributive clause: that, which, who, whom, whose, as, where, when, why…

Adverbial clause: when, as, while, as soon as, before, after, since, till, until be-cause, in that, now that, where, though, although, even if, even though, if, unless, so that, in order that…

Noun clauses: that, who, whose, when, what, whom, which, where, how…

Obviously, a lot of repetition exists and the most commonly used are picked to retrieve: that, who, whom, whose, what, which, whatever, whichever, when/while, how, why, where, whenever, until/till, as, as though, since, because, so that, in other that, if/provided/in case/whether, although/though/even if, before and after.

Data Presentation and Discussion

In this part, results of this research are presented and discussed. All the 180 sampled articles are analyzed concerning the mean length of sentence, coordinating conjunctions and subordinating conjunctions. Excel will be used to help describing statistics.

Syntactic Fluency: The Mean Length of Sentence

Without further introduction, the results of the mean length of sentence will be showed as Table 1 and Figure 1.

<table>
<thead>
<tr>
<th>Number of Sentences</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Words</td>
<td>5008</td>
<td>6185</td>
<td>6925</td>
<td>7675</td>
<td>7748</td>
<td>6779</td>
</tr>
<tr>
<td>Mean Length of Sentence</td>
<td>12.3</td>
<td>14.6</td>
<td>14.2</td>
<td>14.9</td>
<td>15.1</td>
<td>15.9</td>
</tr>
</tbody>
</table>

Figure 1. Results of the Mean Length of Sentence.
As demonstrated in Table 1, the mean length of sentences from term 1 to term 6 increases in sequence except term 3. The mean of term 1 is 12.33 and term 6 is 15.95. The growth rate is reaching to 29.36%. Transparently, a huge improvement is displayed. The growth rate in order is: 1-2: 18.9%, 2-3: -2.8%, 3-4: 5.0%, 4-5: 1.4% and 5-6: 5.2%. Obviously, there is a largest improvement between term 1 and term 2, reaching a peak 18.9%. And a small decline appears between term 2 and term 3.

We come back to check the compositions of term 3 and find that a dense mass of short sentences exists, such as: “However, we had no money.” “I will tell you some examples.” Furthermore, there are many dialogues, for instance: “I said to my friend ‘yes, I have attracted by it.” “If you buy two, we can cheaper.” “How do you think it?” Generally, insufficient as oral English is, the length of sentences may descend. And short sentence needs more development.

The total number of sentences and total number of words are rising orderly from term 1 to term 5 while term 6 sharply drops. And the words standard testing requested are: term1: 120 words, term 2: 150 words, term 3: unknown, term 4: 250 words, term 5: 200 words and term 6: unknown.

Since there is a great decline in term 1, we check up the statistical data of term 6, only finding that: 7 articles out of 30 are less than 200 words. The statistic is: 159,199,173,158,166,179 and 180. This situation has never occurred before and at the beginning we think it is sampling error. Then we count the total number of the left 18 compositions in this class. Among the 18 articles, 4 are less than 200 words: 171,190,176 and the least one 146. Only one among the whole 48 compositions is more than 300 words. And 11 out of 48 are less 200 words, 22.9% in the total 48. We insist that this phenomenon is caused by too much questions or difficulties in the test.

Table 2. The Minimum and Maximum Sentences.

<table>
<thead>
<tr>
<th>Term</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term1</td>
<td>8.29</td>
<td>18.10</td>
</tr>
<tr>
<td>Term2</td>
<td>9.35</td>
<td>20.30</td>
</tr>
<tr>
<td>Term3</td>
<td>10.53</td>
<td>18.83</td>
</tr>
<tr>
<td>Term4</td>
<td>10.56</td>
<td>20.62</td>
</tr>
<tr>
<td>Term5</td>
<td>11.71</td>
<td>20.93</td>
</tr>
<tr>
<td>Term6</td>
<td>12.30</td>
<td>23.67</td>
</tr>
</tbody>
</table>

Figure 2. The Minimum and Maximum Sentences.

Table 2 and Figure 2 are the minimum and maximum length of sentences in each term. The minimum length of sentence is rising all the time from term 1 to term 6. While the maximum length is also increasing except term 3. Reasons have been discussed above. An obvious trend is that the length of sentence is going up.

**Syntactic Complexity**

Syntactic complexity plays a significant role in the research of syntactic, and it is generally supposed that syntactic complexity well represented syntactic features and has a lot to do with writing proficiency and language level. As we have noted before, T-units clause is an effective of syntactic complexity. Since we cannot do a survey on it directly, we study on syntactic features of writing in terms of coordinating conjunctions and subordinating conjunctions.

Coordinating Conjunctions
As we have concerned in part 3, seven coordinating conjunctions (and, but, or, so, for, yet and nor) are searched by WordSmith. Among them, “yet” and “nor” are not list in Table 3, for nobody use them in the whole 6 terms. And the results are as followed:

Table 3. Coordinating Conjunctions.

<table>
<thead>
<tr>
<th></th>
<th>and</th>
<th>but</th>
<th>so</th>
<th>or</th>
<th>for</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>38</td>
<td>8</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>57</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>11</td>
<td>7</td>
<td>6</td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>15</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>42</td>
</tr>
<tr>
<td>5</td>
<td>24</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>32</td>
</tr>
<tr>
<td>6</td>
<td>23</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>35</td>
</tr>
<tr>
<td>Total</td>
<td>140</td>
<td>47</td>
<td>29</td>
<td>12</td>
<td>12</td>
<td>240</td>
</tr>
</tbody>
</table>

Illustrated in Table 3, 240 coordinating conjunctions are used in the whole 6 terms. “and” accounts for a major part in all the 5 coordinating conjunctions for all terms. It represents more than half of all coordinating conjunctions, reaching to 58.33%. Next are “but” (19.58%) and “so” (12.08%) while “or” (5%) and “for” (5%) are on the equal footing. In addition, “yet” and “nor” are never used in the all 3 years. From Figure 3, no obvious regular pattern is discovered from term 1 to term 6. Thus, we calculate the usage of freshman, sophomore and junior: 81 (33.75%), 92 (38.33%) and 67 (27.92%). From freshman to sophomore, there is an ascent and next to junior, a decline.

As freshman, the English writing proficiency of students is limited, and simple coordinating conjunctions are commonly used. For instance, in term 1, 38 out of 57 conjunctions is “and”. Then, with the development of writing proficiency, students try to use other conjunctions such as “for” and “or”. As for the decline, they may learn other cohesive devices in junior. We will combine the usage of subordinating conjunctions with coordinating conjunctions to explain this fall.

Subordinating Conjunctions

As it is noted before, 31 common-used subordinating conjunctions have been searched. The quantitative approach attempts to figure out the frequencies of different kinds of subordinating conjunctions in the articles. Several of the 31 subordinating conjunctions are never used in all the 180 compositions. Conjunctions are omitted in the table 4 such as “while” and “whether”, since they have the same functions with “if” and “when”. And the results are showed in Table 4, Table 5 and Graph 4:

Table 4. Common-used Subordinating Conjunctions.

<table>
<thead>
<tr>
<th></th>
<th>that</th>
<th>which</th>
<th>if</th>
<th>when</th>
<th>as</th>
<th>what</th>
<th>who</th>
<th>because</th>
<th>though</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>10</td>
<td>17</td>
<td>11</td>
<td>2</td>
<td>0</td>
<td>8</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>21</td>
<td>11</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>70</td>
<td>30</td>
<td>38</td>
<td>31</td>
<td>1</td>
<td>8</td>
<td>11</td>
<td>18</td>
<td>6</td>
</tr>
</tbody>
</table>
From the frequency of each term, we can clearly observe the changes in conjunctions. Apart from term 2 and term 6, each is rising according to the usages of subordinating conjunctions. And the growth rate successively is -6.99%, 75.19%, 47.21%, 18.52% and -45.83%. We can see that there is a decline not only in the usage of subordinating but coordinating conjunction from term 1 to term 2 and term 5 to term 6. It is slowdown in term 2 while there is a plunge in term 6. The former one is normal when the later one is aberrant. In addition, the mean length of sentence increases although students use fewer conjunctions.

We choose compositions in which there are few conjunctions in term 6 and discover that students have the tendency to employ a various cohesive device in their writings such as independent genitive structure, past participle and present participle. For example: “Nowadays, with the development of industry and society, energy crisis became a danger and it is a sign that we should realize we need a new way to solve energy crisis.” “Using energy in circle, I think, is a good way to solve energy crisis.” The decline in conjunctions does not mean a decline in writing proficiency.

Among the subordinating conjunctions, “that” holds highest percentage 31.94%, followed by “which” 14.30%, “if” 11.96% and “when” 10.28%. Conjunctions such as “provided” and “in case that” are never used. Generally, “that”, “which” and “if” is the top three most frequently used subordinating conjunctions, reaching to 58.20%. That means the subordinating conjunction variety is not enough.
Conclusion

This study aims to explore whether the English writing proficiency of English majors in China have been improved regarding the development of sentences in their English writing. Syntactic complexity and syntactic fluency are concerned in this thesis. According to syntactic complexity, the mean length of sentence is searched. As for syntactic fluency, we learn coordination and subordination by study on different kinds of conjunctions. The main results from this research can be encapsulated as follows:

(1) According to the mean length of sentences, it increases all the time apart from a small down in term. Great improvement has been obtained by the English majors.

(2) No obvious changes of the usages of coordinating conjunctions have been observed while the usages of subordinating conjunctions rise sharply except term 2 and term 6. However, the variety of conjunctions is not enough.

Imperfections in this study still exist. This is a quantitative research, rather than a qualitative one. It is, among other things, by no means empirical studies supported by data and statistical tests. It is doubtful that whether the sample is representative or not. Meanwhile, on account of technical restriction, we cannot study the T-units. Therefore, many questions in this study need further exploration.

References


