# Corpus-based study on syntactic complexity of texts by L1 and L2 learners

*Ronggen* Zhang[*]

Department of Fundamental Teaching, Shanghai Publishing and Printing College, Shanghai, China

**Abstract.** Based on 90 argumentative texts from the corpora of LOCNESS and WECCL, the paper attempts to make a study of the syntactic complexity of those texts. All the data are processed online by the Web-based L2 Syntactic Complexity Analyser, and then processed with the corpus tool AntConc and IBM SPSS Statistics 20. The findings include: for one thing, L1 learners use more complex sentences with more verb phrases, and more embedded clauses, in their writings, compared with L2 learners. For another, in writing L2 learners may use dependent clauses with longer size than those by L1 learners despite the L1 learners tend to use sentences embedded with dependent clauses more often. Finally, some pedagogical suggestions are made on how to improve L2 writing course teaching.

## 1 Introduction

This study is based on corpus, and all data are processed by computer. The data include 90 argumentative papers of Chinese English majors and US undergraduates based on two corpora, namely the Louvain Corpus of Native English (LOCNESS) [1] and Written English Corpus of Chinese Learners (WECCL) [2]. LOCNESS' 45 articles cover some social problems, such as money as the source of all evil, crime, male or female social contribution, feminism, death penalty, legalization of marijuana, the role of teachers, euthanasia, rules and regulations, and WECCL's other 45 articles include papers on some social problems, such as the consequences of expensive education failure, plastic pollution and college students living outside campus. The research attempts to find the differences of texts by L1 and L2 learners in respective of syntactic complexity.

Syntactic complexity is an important index to measure the quality of L2 writing. According to Ortega (2003), syntactic complexity refers to the range of forms and the complexity of these forms in the process of language production [3]. Casaneve (1994) described syntactic complexity as the ratio of clauses per T unit (C / T) and subordinate clauses per T unit (DC / T). [4]. And Lu, & Ai (2015) illustrated syntactic complexity as the range of syntactic structures generated and the complexity of these structures [5]. Barrot, & Agdeppa (2021) found the output unit length index, phrase complexity index, weighted clause ratio and the number of words per article differentiate the writing proficiency levels between by L1 and L2 learners [6].

[*] Corresponding author: zrgen@163.com

Zeng (2011) compared several syntactic complexity indices of argumentative writing of second year English majors. [7]. He and Li, et al (2018) found sentence complexity, unit length and specific phrase structure are the three main indicators of syntactic complexity of highly scored students' writings [8]. Zheng (2012) found the number of clauses in t unit (C / T) and the number of dependent clauses in clause (DC / C) were not distinguishable between grades and semesters, which could not be considered as effective indices of syntactic complexity [9]. Jiang (2019) concluded that except for the mean sentence length and the ratio of T unit to sentence, other syntactic complexity indices could effectively reflect the learner's writing level; in terms of subordinate structures and coordinate structures, there are significant differences between L1 and L2 learners [10].

The above literature review reveals to us that most of the studies on syntactic complexity are made on L2 learners, while there are fewer comparative studies on both L1 and L2 learners. Therefore, this paper attempts to elaborate this field by comparing the texts from LOCNESS and WECCL.

Research Question

1. What are the significant differences in syntactic complexity of the texts by L1 and L2 learners?

2. What is the pedagogical significance of the research findings?

Purpose of this research

The purpose of this study is to find out the differences between L1 and L2 learners' writing through data mining and analysis of the corpus, and to provide some suggestions for L2 learners.

Significance of this research

The significance of this study is to provide instructors and students with a systematic and computer-aided method to analyse the text from the perspective of syntactic complexity, so as to strengthen the teaching of English writing.

## 2 Material and methodology

1) Sampling

The corpora concerned are based on the 90 pieces of argumentative writings from the Louvain Corpus of Native English (LOCNESS) [1] and Written English Corpus of Chinese Learners (WECCL) [2].

2) Data mining: all the data are processed by using the software such as AntConc, SPSS 19, the Web-based L2 Syntactic Complexity Analyser [11， etc. For the purpose of statistics, WECCL is valued as 1 while LOCNESS is valued as 2.

**Table 1.** Syntactic complexity indices concerned in the research.

| Index Abbreviation | Index Name | Index Abbreviation | Index Name |
|---|---|---|---|
| W | Word number | C/T | Clause per T-unit |
| S | Sentence | DC/C | Dependent clause per clause |
| C | Clause | DC/T | Dependent clause per T-unit |
| T | T-unit | T/S | T-unit per sentence |
| MLS | Mean length of sentence | CT/T | Complex T-unit ratio |
| MLT | Mean length of T-unit | CP/T | Coordinate phrase per T-unit |
| MLC | Mean length of clause | CP/C | Coordinate phrase per clause |
| C/S | Clause per sentence | CN/T | Complex nominal per T-unit |
| VP/T | Verb phrase per T-unit | CN/C | Complex nominal per clause |

Here are the definitions of some more syntactic complexity indices: W/T stands for total number of words / total number of T units), W/C for total number of words / total number of clauses), and W/S for total number of words / total number of sentences. A T-unit is main clause plus any dependent clauses attached to or embedded in it (Hunt, 1966) [12].

**Table 2.** Descriptive statistics of the syntactic complexity indices for WECCL / LOCNESS.

| | | **min** | **max** | **mean** | **ST D.** | | | **min** | **max** | **mean** | **ST D.** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| W/S | WEC | 10.79 | 29.67 | 17.01 | 3.69 | DC/C | WEC | 0.2 | 0.58 | 0.36 | 0.1 |
| | LOC | 13.32 | 32.68 | 19.47 | 4.51 | | LOC | 0.22 | 0.58 | 0.42 | 0.09 |
| W/C | WEC | 7.18 | 14.85 | 9.58 | 1.73 | DC/T | WEC | 0.21 | 1.1 | 0.61 | 0.25 |
| | LOC | 6.63 | 13.19 | 9.38 | 1.58 | | LOC | 0.29 | 1.36 | 0.82 | 0.28 |
| W/T | WEC | 10.41 | 27.38 | 15.39 | 3.13 | T/S | WEC | 0.96 | 1.4 | 1.11 | 0.11 |
| | LOC | 12.19 | 27.43 | 17.32 | 3.47 | | LOC | 0.93 | 1.32 | 1.12 | 0.1 |
| MLS | WEC | 10.79 | 29.67 | 17.01 | 3.69 | CT/T | WEC | 0.16 | 0.75 | 0.45 | 0.15 |
| | LOC | 13.32 | 32.68 | 19.47 | 4.51 | | LOC | 0.25 | 0.79 | 0.54 | 0.15 |
| MLT | WEC | 10.41 | 27.38 | 15.39 | 3.13 | CP/T | WEC | 0 | 1 | 0.33 | 0.19 |
| | LOC | 12.19 | 27.43 | 17.32 | 3.47 | | LOC | 0.07 | 0.88 | 0.44 | 0.18 |
| MLC | WEC | 7.18 | 14.85 | 9.58 | 1.73 | CP/C | WEC | 0 | 0.75 | 0.21 | 0.14 |
| | LOC | 6.63 | 13.19 | 9.38 | 1.58 | | LOC | 0.04 | 0.55 | 0.24 | 0.11 |
| C/S | WEC | 1.18 | 2.83 | 1.8 | 0.36 | CN/T | WEC | 0.75 | 3.1 | 1.68 | 0.55 |
| | LOC | 1.34 | 3.21 | 2.1 | 0.46 | | LOC | 1.08 | 4.14 | 1.99 | 0.62 |
| VP/T | WEC | 1.32 | 3.17 | 2.23 | 0.4 | CN/C | WEC | 0.52 | 2.38 | 1.06 | 0.37 |
| | LOC | 1.59 | 3.46 | 2.42 | 0.43 | | LOC | 0.58 | 1.85 | 1.09 | 0.34 |
| C/T | WEC | 1.05 | 2.17 | 1.62 | 0.27 | | | | | | |
| | LOC | 1.32 | 2.54 | 1.87 | 0.31 | | | | | | |

Table 2 shows that the mean of each syntactic complexity index for WECCL is smaller than that for LOCNESS, except that of W/S and MLC. That is, the US undergraduates (L1 learners) do better than Chinese English majors (L2 learners), in terms of the mean length of sentence, the mean number of clause per sentence, the mean number of verb phrase per main clause, the mean number of coordinate phrase per clause and the mean number of complex nominal per clause, and the like. This reveals to us that L1 learners use more complex sentences with more verb phrases, and more embedded clauses, in their writings, compared with L2 learners. However, L2 learners only do slightly better than L1 learners in terms of W/C and MLC, i.e. in writing L2 learners may use dependent clauses with longer size than those by L1 learners despite the L1 learners tend to use sentences embedded with dependent clauses more often.

**Table 3.** Correlation between WECCL and LOCNESS in syntactic complexity indices.

| | **W/S** | **W/C** | **W/T** | **MLS** | **MLT** | **MLC** | **C/S** | **VP/T** |
|---|---|---|---|---|---|---|---|---|
| CORP | .289** | -0.061 | .284** | .289** | .284** | -0.061 | .352** | .230* |
| | DC/C | DC/T | T/S | CT/T | CP/T | CP/C | CN/T | CN/C |
| CORP | .338** | .372** | 0.072 | .269* | .289** | 0.126 | .260* | 0.041 |

* P<0.05; ** P<0.01

In table 3, Corpus is positively correlated with most of the syntactic complexity indices, i.e. W/S, W/T, MLS, MLT, C/S, VP/T, DC/C, DC/T, CT/T, CP/T, and CN/T. Especially, Corpus is the most positively correlated with DC/T(.372**),C/S(.352**), and DC/C(.338**), that is, there are significant differences between WECCL and LOCNESS in terms of the number of dependent clause per main clause, the number of clause per sentence, and the number of dependent clause per clause. This suggests that the native undergraduates are liable to use more complex sentences with embedded clauses in their writings, compared with L2 learners. Nevertheless, WECCL is weakly correlated with W/C and MLC, which means that the L2 learners may use longer dependent clauses in their writings despite the number of those dependent clauses is fairly smaller than that of the L1 learners. All these findings correspond to those data in table 2.

**Table 4.** T-test, paired sample correlation coefficient.

| CORP& | DC/T | C/T | DC/C | C/S | W/S | MLS | CP/T | W/T | MLT | CT/T | CN/T | VP/T | CP/C | T/S | MLC | CN/C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CC | 0.37 | 0.39 | 0.34 | 0.35 | 0.29 | 0.29 | 0.29 | 0.28 | 0.28 | 0.27 | 0.26 | 0.23 | 0.13 | 0.07 | -0.06 | 0.04 |
| Sig. | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.03 | 0.24 | 0.50 | 0.57 | 0.70 |

Table 4 further confirms there are significant differences between the writings of L1 and L2 learners in terms of syntactic complexity indices, except number of coordinate phrases, number of T-units per sentence, mean length of clause, and number of complex nominal phrases per clause.

## 3 Conclusion

To summarize the results of the analyses from above, here comes the following points:

First, for one thing, L1 learners use more complex sentences with more verb phrases, and more embedded clauses, in their writings, compared with L2 learners. For another, in writing L2 learners may use dependent clauses with longer size than those by L1 learners despite the L1 learners tend to use sentences embedded with dependent clauses more often.

Corresponding to findings above, some pedagogical suggestions are made as follows:

First, L2 learners should be encouraged to practise writing more often, for the writing course is a practical course needing frequent exercise.

Second, L2 learners should read more original materials written by native speakers of English, which are filled with more complex sentences.

Finally, L2 learners should read more original English grammar books so as to get more familiar with English verb phrases, syntactic structures, and so for.

## References

1. S. Granger, The computer learner corpus: A versatile new source of data for SLA research. In Granger, S. (ed.) Learner English on Computer. Addison Wesley Longman: London & New York. (1998).
2. W. Wen, M. Liang, X. Yan, Foreign Language Teaching and Research Press. (in Chinese) (2008).
3. L. Ortega, Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics* **24** (4), 492-518. (2003).

4. C. P. Casanave, Language development in students' journals. *Journal of Second Language Writing*, **3** (3), 179-201. (1994).
5. X. Lu, H. Ai, Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, **29**, 16-27. (2015).
6. J. Barrot, J. Agdeppa, Complexity, Accuracy, and Fluency as Indices of College -Level L2 Writers ' Proficiency. *Assessing Writing*, **47** (1), 100510. (2021).
7. X. Zeng, A Comparison of Syntactic Complexity in Timed and Untimed Compositions of English Majors, *Journal of PLA University of Foreign Languages*, **5**:67-74.(in Chinese) (2011).
8. X. He, Z. Li, et al, A Study on the Syntactic Characteristics of Highly Scored English Written Texts- Data Mining of English Written Texts Based on Juku Correction Network, *Modern Education Technology*, **28**(12),74-79.(in Chinese). (2018).
9. Y. Zheng, Chinese Learners' Lexical and Syntactic Development in English L2 Writing, Doctoral dissertation, Shanghai Foreign Studies University. (in Chinese). (2012).
10. M. Jiang, Towards Syntactic Complexity and Semantic Cohesion in Argumentative Writings, Master's thesis, Guangxi University). (in Chinese). (2019).
11. X. Lu, Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, **15**(4), 474-496. (2010).
12. K. Hunt, Recent measures in syntactic development. *Elementary English*, **43**, 732–739. (1966)