Exploratory and Confirmatory Factor Analyses of L2 Linguistic Complexity Measures

Hang Li[1] & Shuting Zhang[1]

[1] School of International Studies, Zhejiang University, Hangzhou, China

Correspondence: Shuting Zhang, School of International Studies, Zhejiang University, Hangzhou, 310058, China. E-mail: zhangshuting@zju.edu.cn

Received: October 25, 2020 Accepted: November 30, 2020 Online Published: December 16, 2020

doi:10.5539/ijel.v11n1p192 URL: https://doi.org/10.5539/ijel.v11n1p192

Abstract

This study investigated the latent structure of L2 linguistic complexity as a multidimensional construct and analyzed the relationship between the sub-constructs of L2 linguistic complexity by employing exploratory and confirmatory analyses of a set of linguistic complexity measures indexing different sources of L2 linguistic complexity. Based on relevant theories and empirical studies, 11 automated measures indexing distinct sources of syntactic and lexical complexity were selected and used to assess the linguistics complexity of 930 EFL argumentative essays, which were then equally divided into two subsamples. Sample 1 was used for exploratory factor analysis while sample 2 was used for confirmatory factor analysis. The results show that L2 linguistic complexity is a multi-dimensional construct composed of clausal subordination, phrasal elaboration and lexical complexity. Furthermore, regarding the relationships among the three sub-constructs, it was found that lexical complexity and phrasal elaboration are moderately correlated; while clausal subordination employs rather different means of complexification than that employed by phrasal elaboration and lexical complexity. Findings of the study provide empirical evidence for the multidimensionality of L2 linguistic complexity in L2 argumentative writing and lend support to the hypothesis that lexical complexity and grammatical complexity constitute separate, independent dimensions of L2 performance and proficiency, and that there was a certain level of trade-off effect between them.

Keywords: L2 linguistic complexity measures, multidimensionality, exploratory factor analysis, confirmatory factor analysis

1. Introduction

Complexity, especially linguistic complexity has featured prominently in L2 writing research (Bulté & Housen, 2012). Though no consensus has been reached on its definition (Bulté & Housen, 2014), over the last decade, theoretical models of syntactic complexity (Norris & Ortega, 2009) and linguistic complexity (Bulté & Housen, 2012), as well as an increasing amount of empirical evidence have both pointed towards the multidimensionality of this construct (Biber, Gray, & Poonpon, 2011; Bulté & Housen, 2014; Crossley & McNamara, 2014; Mazgutova & Kormos, 2015). One way to examine linguistic complexity and its sub-dimensions is by performing a factor analysis of the complexity measures that are used to gauge the observable properties of the construct. Over the years, a wealth of complexity measures has been proposed in L2 research (Bulté & Housen, 2012). Although some of them measure exactly the same thing and are thus redundant, others are believed to be measuring distinct qualities or dimensions of L2 complexity (Norris & Ortega, 2009). Therefore, by doing exploratory and confirmatory analysis of a set of linguistic complexity measures indexing distinct sources of L2 linguistic complexity, the present study hopes to gain further understanding about the complexity measures selected, and gather empirical evidence for the multidimensionality of L2 linguistic complexity.

2. Literature Review

2.1 L2 Linguistic Complexity

Complexity is in itself a highly complex construct (Bulté & Housen, 2012, 2014; Norris & Ortega, 2009). In their taxonomy of L2 complexity, Bulté and Housen (2012) distinguished between a relative and absolute approach to the notion of complexity. Relative complexity, defined in relation to language users, implies cost and difficulty of processing or learning; whereas absolute complexity derives from objective inherent properties of linguistic and/or systems thereof, and defines language complexity in objective, quantitative terms. Bulté and Housen (2012) further defined L2 complexity as being composed of propositional complexity, discourse-interactional complexity and linguistic complexity. Of the three, linguistic complexity, which can be evaluated across such domains as phonology, lexis, morphology and syntax, has received the most attention (Bulté & Housen, 2012). The present study, taking an absolute approach to the construct of complexity, focuses on the two domains of lexis and syntax, as they are both crucial for effective communication and are also the targets of most of the current repertoire of measures in L2 research.

2.1.1 L2 Lexical Complexity and Its Measures

In L2 writing literature, lexical complexity, sometimes also termed "lexical richness" (Engber, 1995; Lu, 2012; Read, 2000), is generally viewed as the possession of a wide variety of basic and sophisticated words that can be accessed quickly (Lu, 2012; Wolfe-Quintero, Inagaki, & Kim, 1998). It is thus hypothesized as being composed of lexical variation or diversity, lexical sophistication and lexical density (Read, 2000; Lu, 2012).

Lexical variation refers to the range of a learner's vocabulary as displayed in his or her language use (Lu, 2012). A widely used measure gauging lexical variation is type-token ratio (TTR). However, this measure is extremely sensitive to sample size, and tends to decrease with the increase of the text length. A number of measures that are mathematical transformations of TTR have been proposed over the years to mitigate the sample size effect, such as MSTTR, RTTR, CTTR, logTTR, vocd-D, HD-D etc. Of the various lexical diversity measures, MTLD (the Measure of Textual Lexical Diversity, calculated as the mean length of word strings that maintain a given TTR value (0.720)) was found to be least affected by text length (McCarthy, 2005; McCarthy & Jarvis, 2010) since the value of MTLD is the average number of words required for the text to reach a point of stabilization (where neither the introduction of repeated types nor even a considerable string of new types can markedly affect the TTR trajectory), though it should be used with texts of at least 100 tokens (Koizumi, 2012).

Lexical sophistication, defined as a person's command of less-common words (Javier, 2013), or the proportion of relatively unusual or advanced words in the learner's text (Read, 2000, p. 203), is usually measured by referencing to large and well-balanced corpus, such as the British National Corpus (BNC), for word frequency. Typically, lexical sophistication is gauged either by a ratio measure, with the number of sophisticated types/tokens being the numerator and the total types/tokens being the denominator (Lu, 2012; Wolfe-Quintero, Inagaki, & Kim, 1998); or a frequency measure, where the overall commonness of the words used in the text is assessed (Jarvis, 2013).

Lexical density refers to the proportion of content words or lexical words to the total number of words in a text (Lu, 2012). It should be noted that notable variability exists in terms of how lexical words were defined. Meanwhile, the validity of this measure seems dubious, as no empirical studies of either L2 writing or speaking reported significant correlation between it and overall quality of L2 production.

With regard to the validity of the dimensionality of lexical complexity described above, little empirical evidence exists in L2 literature. Lu (2012), in a study on the relationship between lexical complexity and the quality of L2 learners' oral narrative, examined the correlations between measures of lexical density, variety and sophistication. Lu argued that the three dimensions were indeed different constructs as there were no strong correlations between them. He did propose, however, that the findings be confirmed by a factor analysis.

The other hypothesis worth examining is the distinctiveness of lexical complexity. Guided by Levelt's (1989) model of speaking, and based on empirical study on the differences between the quality of native and non-native speakers' oral narrative, Skehan (2009) and Foster and Tavakoli (2009) proposed that at least for non-native users, lexical complexity and grammatical complexity constitute separate, independent dimensions of L2 performance and L2 proficiency, rather than being different aspects of the same L2 performance-proficiency area. This proposal was supported by Bulté and Housen (2014), who found that, for a group of college-level ESL learners taking a four-month intensive EAP course, lexical and syntactic complexity did not develop in parallel over the said period of time. Clearly, findings from a factor analysis of L2 linguistic complexity measures could offer additional empirical evidence regarding the relationship between lexical and grammatical complexity.

2.1.2 L2 Syntactic Complexity and Its Measures

In L2 writing, syntactic complexity, sometimes termed "grammatical complexity" (Biber, Gray, & Poonpon, 2011; Wolfe-Quintero, Inagaki, & Kim, 1998), was viewed in terms of the range and sophistication of the syntactic structures produced (Lu, 2011; Wolfe-Quintero, Inagaki, & Kim, 1998). Drawing on both empirical findings and systemic and functional linguistics (Halliday & Mathiessen, 1999), Norris and Ortega (2009) proposed a multidimensional framework of L2 syntactic complexity, which includes the following measurable sub-constructs: (i) complexity via subordination; (ii) overall or general complexity; (iii) subclausal complexity via phrasal elaboration; (iv) clausal complexification via coordination; and (v) the variety, sophistication and acquisitional timing of forms produced.

Complexity via subordination, which Norris and Ortega (2009) proposed as a powerful index of complexification at intermediate level, can be measured by any metric with clause (or subordinate or dependent clause) in the numerator, regardless of the denominator of choice. In other words, many popular measures, such as clauses per sentence (cS) and dependent clauses per clause (dcC), are measures of the same construct (Norris & Ortega, 2009).

Overall or general complexity can be measured by any length-based index with a potentially multiple-clausal unit of production in the denominator (Norris & Ortega, 2009), such as mean length of sentence (MLS) and mean length of T-unit (MLT). Such measures, however, are not without problems, for they are not well motivated from a linguistic perspective (Biber, Gray, & Poonpon, 2011). Norris and Ortega (2009) pointed out that, all measures with a denominator that is potentially multiple-clausal in scope can become longer or more complex in several ways that cannot be determined by the numerical results that these measures yield. As a result, such measures "can only be interpreted as a global or generic metric of linguistic complexity" (Norris & Ortega, 2009).

Subclausal complexity via phrasal elaboration, though a fairly recent development in both L1 and L2 complexity research (Bulté, & Housen, 2014), has been established as an integral part of syntactic complexity based on findings from both corpus-based analysis of conversational and written discourses (Biber, 2006; Biber, Gray, & Poonpon, 2011) and L2 developmental studies on syntactic complexity (Bulté & Housen, 2014; Crossley & McNamara, 2014; Mazgutova & Kormos, 2015). Norris and Ortega (2009) proposed that subclausal complexity be measured by mean length of clause (MLC), since clause length taps complexification subclausally or at the phrasal level. To capture specific means by which clauses get lengthened, researchers have also used length measures that gauge average noun phrase length (McNamara et al. 2014); and ratio measures such as coordinate phrases per clause (cpC) (Ai & Lu, 2013; Lu, 2011), and complex nominals per clause (cnC) (Ai & Lu, 2013; Lu, 2011), which includes both headed nominals, i.e., nouns with pre- or post-modifiers, and non-headed nominals, i.e., nominal clauses, as well as gerunds and infinitives in subject position.

Clausal complexification via coordination was proposed to be chiefly relevant for data at beginning levels of L2 development (Norris & Ortega, 2009), since sentence coordination is usually introduced in the early stages of English instruction. This sub-construct has been measured by the ratio between the number of T-units and the number of sentences (tS).

The last sub-construct of syntactic complexity proposed by Norris and Ortega (2009), i.e., the variety, sophistication and acquisitional timing of forms produced, focuses on specific syntactic forms that are deemed difficult and/or carry certain developmental significance. Though some researchers have made hypotheses of L2 developmental stages for complexity features (e.g., Biber, Gray, & Poopon, 2011), these hypotheses have yet to be verified through fine-tuned L2 developmental studies. As a result, this sub-construct is yet to be operationalized.

Little empirical evidence exists in L2 literature regarding the distinctiveness of these sub-constructs. Lu (2011) looked into the relationship between pairs of 14 indices of L2 syntactic complexity by examining the correlations between them. Computed by a computational system for automatic measurement of L2 syntactic complexity (Lu, 2010, 2011), these indices were also claimed to be aligned with the four sub-constructs of syntactic complexity proposed by Norris and Ortega (2009), i.e., overall complexity, complexity via subordination, phrasal elaboration and coordination (Lu, 2017). The results indicated that first, measures generally correlate strongly with other measures of the same type or involving the same structure. Second, measures correlate moderately to strongly with other measures gauging the same sub-construct. Third, measures gauging phrasal elaboration (MLC, cpC and cnC) exhibit low to weak negative correlations with subordination measures (e.g., cS and dcC). Fourth, the two measures of overall complexity, i.e., MLS and MLT correlate moderately to highly with most other measures. Finally, the coordination sub-construct measure, tS, shows mostly low to weak correlations with other measures, except for cS, which is hardly surprising as the two measures share the same denominator.

2.1.3 Factor Analysis on Multidimensionality of Linguistic Complexity

So far, the multidimensionality of linguistic complexity remains much of a hypothesis and rarely have researchers analyzed the multi-dimensionality of syntactic complexity using a factor-analysis approach. The only extant study was carried out by Yoon (2017), who used exploratory factor analysis in an attempt to explore the underlying constructs of 14 linguistic complexity measures. Yoon found that unit-length measures (MLS and

MLT), together with C/T and T/S, both of which measure clause-level syntactic complexity load on one factor. MLC, and CP/C and CN/C, which are phrase-level complexity measures, load on another factor. Meanwhile, lexical diversity, measured by vocD, lexical sophistication measured by word length and word frequency, and morphological diversity load together on the third factor. It should be noted that the study did not tackle the issue of overloading of measures such as MLT and T/S, nor did it verify the factor structure obtained through confirmatory factor analysis. Thus, it could only be viewed as a preliminary attempt at empirically exploring the validity of various dimensions of linguistic complexity. As pointed out by Yoon (2017), evidence for more confirmatory argument is needed.

2.2 Aim of the Current Study

Given the above theoretical and practical limitations of extant studies about L2 linguistic complexity, the primary goal of the study is to investigate the factor structure of L2 linguistic complexity measures. To this end, a number of automated measures indexing distinct sources of syntactic and lexical complexity are selected and used to assess the linguistic complexity of argumentative essays by EFL leaners of intermediate level. Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) are performed to uncover the factor structure of L2 linguistic complexity measures. The general research questions that guide the study are as follows:

1) What is the underlying factor structure of the linguistic complexity measures selected?

2) What are the relationships between the factors, suppose a multidimensional structure is detected?

3. Method

3.1 Participants and Data

The participants were 930 first-year non-English major students in a university in southeast China. All students were matriculated undergraduates from diverse fields of study including social sciences, humanities, engineering and natural sciences, and all had learned English as a foreign language for at least six years. Their CET-4 (College-English Band 4) score ranged from 376 to 696, with an average of 553 (SD = 50.74) (Note 1). CET-4 is a test targeting intermediate-level learners of English in Chinese colleges. Higher-level learners who have passed CET-4, can attempt CET-6. The two tests together form the battery of College English Test (CET). Moreover, in the sample used, about 70% of the learners' CET score fall between one standard deviation of the mean. The students' English proficiency can thus be said to be at the intermediate level. Two timed impromptu writing tasks were designed, with half of the students (N = 465) finishing task one and the other half (N = 465) finishing task two. Both writing tasks were argumentative essays. In task one, the prompt asked students to discuss whether or not children should be allowed more free time to play. In task two, the prompt asked students to discuss whether colleges should ban campus tourism. Since both prompts require writers to "justify their beliefs, and support interpretations of why events follow each other by giving reasons" (Robinson, 2005), they are believed to elicit causal reasoning, which, as predicted by Robinson (2007, 2011), can lead to increased syntactic complexity in language production. The students had 30 minutes to finish either of the tasks in class. The essays on the play topic averaged 219.2 words (SD = 40.4), and the essays on the campus topic averaged 194.8 words (SD = 39.9). No significant differences were found in a pair-sample t-test of the scores awarded to these two groups of essays: t (930) = 0.256, p = 0.821. Therefore, these two prompts can be viewed as of equal difficulty. Prior to analysis, the data was cleaned of formatting and spelling errors.

3.2 Measures Selected

The construction of measurement models implies close relationships among the indicators of a latent variable. However, severe multicollinearity may cause problems in parameter estimation; therefore, theoretically and/or empirically redundant measures should not be included in a model (Ockey & Choi, 2015). In accordance with this principle, a total of 11 measures were chosen, each of which was believed to gauge a distinct source of L2 linguistic complexity. Since the participants of the study were intermediate-level L2 learners, T-units per sentence (tS), a measure proposed to be chiefly relevant for data at beginning levels of L2 development (Norris & Ortega, 2009), was not included in the present study. Specifically, syntactic complexity was measured by eight indices and lexical complexity by three. Table 1 lists all 11 measures.

Table 1. Linguistic complexity measures

Sub-construct	Measure
Clausal subordination	Dependent clauses per clause (dcC)
	Verb phrases per T-unit (vpT)
Phrasal elaboration	Mean length of clause (MLC)
	Complex nominals per clause (cnC)
	Modifiers per noun phrase (SYNNP)
	Coordinate phrases per clause (cpC)
Overall syntactic complexity	Mean length of T-unit (MLT)
	Syntactic similarity (SYNSTRUTt)
Lexical complexity	Lexical density (LD)
	Lexical diversity (MTLD)
	Lexical sophistication (LS2)

(i) *Clausal subordination*. Two measures were chosen to gauge the level of clausal subordination, i.e., dependent clauses per clause (dcC) and verb phrases per T-unit (vpT). Clauses are defined in the present study as structures with a subject and a finite verb, including independent, adjective, adverbial and nominal clauses. Compared with measures with clauses in the numerator, such as cS and cT, dcC offers a more direct measurement of the level of clausal subordination. vpT, which measures the number of finite and nonfinite verb phrases per T-unit, complements dcC, since nonfinite verb phrases are not included in the definition of clauses in the present study.

(ii) *Phrasal elaboration*. Four measures were chosen to index subclausal complexity. Mean length of *clause* (MLC), calculated by dividing the number of words by the number of clauses, can measure the overall complexity of clauses. Three other measures that can gauge specific ways by which phrasal elaboration is realized were also chosen. (1) Complex nominals per clause (cnC) measures the number of both headed nominals and non-headed nominals per clause. Specifically, complexity nominals include nouns plus adjective, possessive, prepositional phrase, adjective clause, participle, or appositive; nominal clauses; and gerunds and infinitives in subject position. (2) The average number of modifiers per noun phrase (SYNNP), measures the average length of headed nominals. (3) Coordinate phrases per clause (cpC), which measures the number of coordinated noun, verb, adjective and adverb phrases per clause, gauges phrasal elaboration via coordination.

(iii) *Overall syntactic complexity*. Two measures were chosen to measure overall syntactic complexity. Mean length of T-unit (MLT) was *chosen* over mean length of sentence (MLS), since the accuracy of the latter can be affected by the existence of run-on sentences. The other measure chosen was SYNSTRUTt, a Coh-Metrix index that measures syntactic similarity (Note 2). By gauging the average parse tree similarity between all combinations of sentence pairs across paragraphs of the text (McNamara et al., 2014), SYNSTRUTt can measure the uniformity and consistency of the syntactic constructions in the text at clausal, phrasal and POS (part of speech) levels. The present study held that while MLT measures the overall sophistication of multiple-clause units in a text, SYNSTRUTt indicates the variety of syntactic constructions produced.

(iv) *Lexical complexity*. Three measures were chosen to measure the three components of lexical complexity, i.e., lexical density, lexical variation and lexical sophistication (Read, 2000). Lexical density (LD) was calculated by dividing the number of lexical words by the total number of words in a text. Following Lu (2012), in the present study, lexical words were defined as nouns, adjectives, verbs (excluding modal verbs, auxiliary verbs, “be”, and “have”), and adverbs with adjective base, including those that can function as both an adjective and adverb and those formed by attaching the -*ly* suffix to an adjective root. The Measure of Textual Lexical Diversity (MTLD) was chosen to measure lexical variation, as it was found to not vary as a function of text length (Jarvis, 2013). As for lexical sophistication, a type/type measure (LS2), calculated by dividing the number of sophisticated word types by the total number of word types in a text (Laufer, 1994), was chosen. Following Lu (2012), words not on the list of the 2000 most frequent words generated by the British National Corpus (BNC) were considered to be sophisticated.

The present study adopted an automatic approach to assessing linguistic complexity, since it “affords speed, flexibility and reliability” (Crossley & McNamara, 2014). Human raters may be better at analyzing L2 productions, however, they are also prone to error and subjectivity. Employing human raters is also time-consuming and resource-intensive, as raters need to be trained and monitored so as to ensure rating quality

(Higgins, Xi, Zchner, & Williamson, 2011).

Three automated analyzers were used in the present study to compute the 11 indices chosen. The lexical diversity measure MTLD, the average noun phrase length measure SYNNP, and the syntactic similarity measure SYNSTRUTt were computed by Coh-Metrix 3.0 (Graesser et al., 2004; McNamara & Graesser, 2012; McNamara et al., 2014). Coh-Metrix was initially designed to assess the cohesion characteristics of a text that contribute to the coherence of the mental representation of the text (McNamara et al., 2014). Of the 106 indices computed by Coh-Metrix, MTLD and three other indices gauge lexical diversity; another seven are measures that McNamara et al. (2014) explicitly discussed as appropriate for examining syntactic complexity and have been used to investigate L2 writing syntactic complexity and its relationship to writing quality (e.g., Crossley & McNamara, 2014), among which were SYNNP and SYNSTRUTt. Coh-Metrix analyzes the structural representations of sentences in parse trees generated by the Charniak parser. For texts written by L1 speakers, the Charniak parser reports an average accuracy of 89% for expository and narrative texts (Hempelmann, Rus, Graesser, & McNamara, 2006). However, the accuracy of the Charniak parser for L2 writing remains unknown, although it can be assumed to be lower than when the parser is used for L1 writing (Crossley & McNamara, 2014).

The other two lexical complexity measures, i.e., Lexical density measure LD and lexical sophistication measure LS2 were computed by Lexical Complexity Analyzer (LCA) (Ai & Lu, 2010; Lu, 2012). The remaining six syntactic complexity measures were computed by L2 Syntactic Complexity Analyzer (L2SCA) (Lu, 2010, 2011), which was designed to automate syntactic complexity measures of L2 English texts (Lu, 2010). For the production units and structures identified by L2SCA, Lu reported F-scores (standardized measure of inter-annotator agreement) ranging from 0.830 to 1.000, and correlation coefficients ranging from 0.834 to 1.000 between the syntactic complexity indices generated by L2SCA and human annotators (Lu, 2010).

3.3 Statistical Analysis

The present study aimed to investigate the factor structure of L2 linguistic complexity. To achieve this aim, both EFA and CFA were applied following the procedure in Yang (2009) and Liao et al. (2015). The two sets of writing samples, i.e., one on task 1 and one on task 2, were each randomly divided into two subsets, resulting in four sets of data, i.e., Task 1-1 (n = 232) and Task 1-2 (n = 233); and Task 2-1 (n = 233) and Task 2-2 (n = 232). Task 1-1 and Task 2-1 were then combined to form sample 1 (n = 465); and Task 1-2 and Task 2-2 were combined to form sample 2 (n = 465).

A set of EFA was first performed on subsample 1 to identify the indices that best measure the same and separate underlying constructs. To pursue this goal, maximum likelihood extraction was used to extract the initial factors; and an oblique rotation was applied to examine the composition of the factors extracted. SPSS 19.0 was used for the EFA. The underlying factor structure uncovered by EFA was then verified by CFA, using subsample 2. One more function that can be performed by CFA is to identify method effects. A method effect exists when additional covariation among indicators is introduced by the measurement approach (Brown, 2006). In the present study, several measures including MLC, cnC, cpC and dcC, all have the number of clauses as the denominator, which could result in shared method variance. Unfortunately, EFA is incapable of estimating method effects. In CFA, however, method effects can be specified as part of the error theory of the measurement model, thus producing conceptually more viable models. Confirmatory factor analysis was performed by using AMOS 21.0.

4. Results

4.1 Descriptive Statistics

Table 2 shows the descriptive statistics of the two samples on all 11 measures. All values of skewness and kurtosis were within the accepted range (±3 for skewness and ±10 for kurtosis) for univariate normality (Kline, 2011).

Table 2. Means, standard deviation (SD) and distribution for sample 1 and 2.

	M		SD		Skewness		Kurtosis	
	Sample1	Sample2	Sample1	Sample2	Sample1	Sample2	Sample1	Sample2
dcC	0.36	0.35	0.10	0.09	-0.07	-0.09	-0.18	-0.31
vpT	2.25	2.22	0.45	0.40	0.74	0.57	0.70	0.23
MLC	9.01	9.00	1.38	1.27	0.65	0.45	0.50	0.01
cnC	0.94	0.95	0.24	0.22	0.31	0.34	0.08	0.13
SYNNP	0.69	0.71	0.15	0.13	0.19	0.17	-0.03	-0.18
cpC	0.21	0.23	0.11	0.12	0.57	0.60	-0.10	0.02
MLT	14.53	14.44	2.53	2.39	0.63	0.60	0.36	0.39
SYNSTRUTt	0.10	0.11	0.03	0.03	0.36	0.21	-0.32	-0.41
LD	0.54	0.54	0.03	0.03	0.17	0.17	-0.37	0.28
MTLD	86.06	88.19	19.95	19.28	0.56	0.50	0.41	0.45
LS2	0.16	0.16	0.05	0.04	0.37	0.28	0.07	-0.17

Note. Sample 1: N = 465; Sample 2: N = 465.

4.2 Item-Level EFAs for L2 Linguistic Complexity Measures

In the first phase of data analysis, the 11 L2 linguistic complexity measures were evaluated with exploratory factor analysis (EFA) using principal axis extraction followed by oblimin rotation. The Kaiser-Meyer-Olkin measure of sampling adequacy (KMO = 0.682) and Barlett's measure of sphericity (p = 0.000) indicated that the data was for suitable for factor analysis (Kaiser, 1974).

Maximum likelihood extraction was used to analyze common factor variability. Loadings with an absolute value less than 0.40 were suppressed. The results showed a three-factor structure, which explained 54.15% of the variance. Most of the measures gauging the same hypothetical sub-constructs loaded together. However, lexical density (LD) did not have sufficient loadings on any factors; and mean length of T-unit (MLT) double-loaded with the first two factors. These two measures were thus dropped from the original inventory.

The final EFA was conducted on the remaining nine measures. The Kaiser-Meyer-Olkin measure of sampling adequacy (KMO = 0.683) and Barlett's measure of sphericity (p = 0.000) indicated that the data was suitable for factor analysis. The maximum likelihood extraction yielded three factors with eigenvalues larger than 1, accounting for 51.92% of the variance (Table 3). The direct oblimin rotation was applied to interpret the three-factor structure.

Table 3. Exploratory factor analysis: three-factor solution.

	Factor		
	1	2	3
Proportion of variance explained	23.92	16.22	11.78
Direct oblimin rotation factor pattern			
Measure			
MLC	0.99		
cnC	0.69		
SYNNP	0.41		
cpC	0.51		
dcC		0.73	
vpT		0.92	
SYNSTRUTt		-0.52	
MTLD			0.51
LS2			0.58
Interrelation			
Factor			
1	1.000	-0.214	0.493
2	-0.214	1.000	-0.057
3	0.493	-0.057	1.000

Based on factor loading pattern and theories of L2 linguistic complexity, the three factors extracted were labelled as phrasal elaboration, clausal subordination and lexical complexity respectively. The correlation between factor 1 and 3 reached 0.493. When correlations exceed 0.32, there is a 10% (or more) overlap in variance among factors (Tabachnick & Fiddell, 2007). Therefore, the use of oblique rotation was justified.

According to the two-indicator rule of CFA (Abell, Springer, & Kamata, 2009), a multifactor model could be identified if the model had at least two or more indicators for each factor. Therefore, the current model should be identifiable in a confirmatory factor analysis.

4.3 CFA: Model Specification and Revision

Although EFAs provide useful information of construct dimensionality, CFAs can be used to test whether a hypothesized factor model fits the sample data. Based on the results of exploratory factor analyses, a three-factor model was hypothesized. A first-order factor analysis was performed to test the hypothesized multidimensionality of L2 linguistic complexity with sample 2.

Prior to model specification, sample 2 was screened for multivariate normality and multicollinearity. Mardia's coefficient was 2.27, which suggested multivariate normal distribution. With regard to the correlations between the nine measures, it is found that the majority of the measures did correlate significantly with one another. This is to be expected, given that these measures were hypothesized to gauge different aspects of the same construct. But it would be ideal if their correlations were only moderate, i.e., no higher than 0.80, so as to indicate that they are not varying as a function of each other. Among all pairs of indices, the highest correlation was between MLC and cnC (r = 0.75, p = 000). The correlation coefficients between MLC and cpC, and MLC and dcC also reached 0.55 and -0.35 (p = 0.000). Since these four measures all share the same denominator, the significant correlations between them could indicate the existence of shared method variance. Therefore, covariance parameters between the error associated with MLC and the error associated with cnC, cpC and dcC were also included in the model (Figure 1).

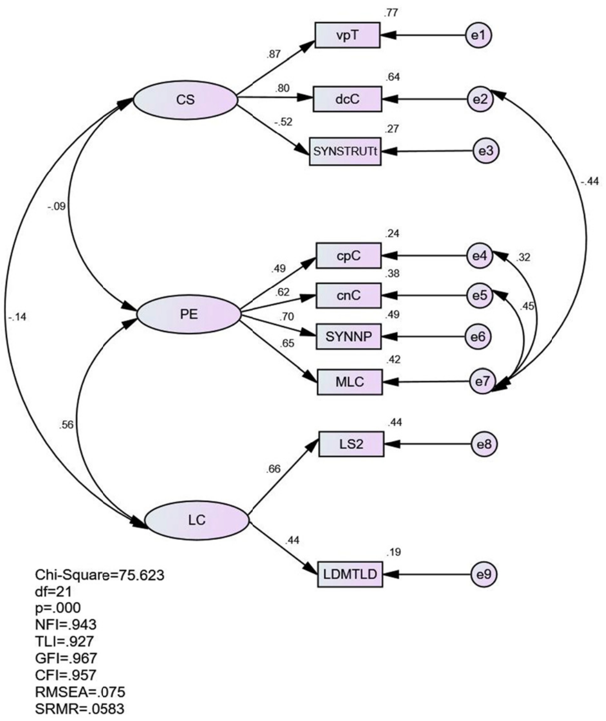

Figure 1. Initially hypothesized model of L2 linguistic complexity

Note. CS = clausal subordination; PE = phrasal elaboration; LC=lexical complexity.

As shown by Figure 1, in terms of model adequacy as a whole, the initially hypothesized three factor model of L2 linguistic complexity produced a chi-square value of 75.623 with 21 degrees of freedom ($p < 0.05$), indicating a global misfit of the model to the sample data. Other fit indices were also used to examine model fit, including the normed fit index (NFI), the Tucker-Lewis index (TLI), the goodness of fit index (GFI), and the comparative fit index (CFI). As shown by Figure 1, the model produced a NFI of 0.943, a TLI of 0.927, a GFI of

0.967, and a CFI of 0.957, suggesting an adequate fitting model. However, the root means square error of approximation (RMSEA), and the standardized root-mean square residual (SRMR) both exceeded the value of 0.05, indicating the model was beyond acceptance. Since the model was not a good fit to the data, model parameters were not interpreted.

To disentangle the mixed results from the fit indices, significance of individual parameters was examined. The loadings of the nine observed variables on the respective factors were all significant ($z > 2$) at 0.01 level. Except for the covariances between clausal subordination and phrasal elaboration, and between clausal subordination and lexical complexity, all variances and covariances among factors, as well as error variances were statistically significant. However, inspection of standardized residuals indicated some localized point of ill fit in this solution. Specifically, the standardized residual covariances between SYNSTRUTt and all four measures of phrasal elaboration exceeded the absolute value of 3; and the standardized residual covariance between vpT and cnC also exceeded the absolute value of 2. Furthermore, modification indices, which reflect an approximation of how much the overall model $\chi 2$ would decrease if the fixed or constrained parameter was freely estimated, were also examined. According to the modification indices, estimating a covariance parameter between the error associated with SYNSTRUTt and the factor PE (phrasal elaboration), and between the errors associated with vpT and cnC would result in a decrease in $\chi 2$ of at least 23.894 and 16.915 respectively.

Based on the aforementioned results of model estimation and literature in L2 writing, a series of post hoc modification procedures were performed to examine alternative models. Two changes were made from the initially hypothesized model. First, SYNSTRUTt, a measure of syntactic similarity was hypothesized to cross-load with both clausal subordination and phrasal elaboration. As a measure of the uniformity and consistency of syntactic constructions in a text at clausal, phrasal and POS (part of speech) levels, SYNSTRUTt should, by definition, be multifaceted in nature. It is therefore reasonable to assume that SYNSTRUTt was associated with both clausal and phrasal-level complexity. Second, a covariance parameter between the errors associated with vpT and with cnC was estimated. This change indicates that these two measured something in common beyond their respective factors. A close look at the numerator of both measures helped make sense of this exgoenous common cause. Verb phrase, in the present study, covers finite verbs, as well as non-finite verbs includinng gerundives, infinitives and participals. Complex nominals, on the other hand, involve such a variety of structures as adjective clauses and participles as noun-modifiers, as well as nominal clauses and gerunds and infinitives in subject position. The overlap in the numerators of these two measures could very likely be the source of their correlated error.

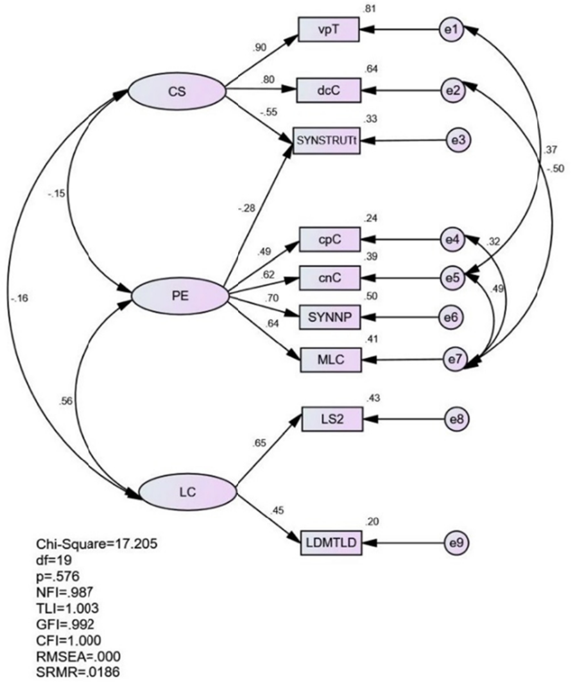

Figure 2. Final version of the hypothesized model of L2 linguistic complexity

Note. CS= clausal subordination; PE= phrasal elaboration; LC=lexical complexity.

Figure 2 presents the final version of the first-order hypothesized model of L2 linguistic complexity with the aforementioned two revisions based on substantive and statistical points of view. This model provided an excellent fit to the data: the chi-square statistic was not significant; both RMSEA and SRMR were below 0.05; and NFI, TLI, GFI and CFI were all considerably above the 0.95 threshold. All loadings of the nine indicator variables on the respective latent factors were significant ($z > 2$) at 0.05 level, as were all the variances and covariances among latent factors. No indicator of ill fit was found in the solution. The modification indices did not suggest any model revisions either.

5. Discussion

5.1 Underlying Factors of L2 Linguistic Complexity Measures

In this study, the factor structure of L2 linguistic complexity measures was examined using both EFA and CFA. Four sub-constructs of L2 linguistic complexity with 11 measures were initially hypothesized. However, only three factors were extracted in the EFAs, and two of the measures, i.e., mean length of T-unit (MLT) and lexical density (LD), were dropped as the former double-loaded on two factors, and the latter did not have sufficient loadings on any factor. Based on factor loadings and theories of L2 linguistic complexity, the three factors extracted were labelled as phrasal elaboration, clausal subordination and lexical complexity. This three-factor structure was later validated by CFAs. As shown by Figure 2, L2 linguistic complexity was represented by three underlying factors measured by nine observed variables. Each factor was well represented by its proposed observed variables—clausal subordination is represented by verb phrases per T-unit (vpT), dependent clauses per clause (dcC), and syntactic similarity measure (SYNSTRUTt); phrasal elaboration is measured by coordinate phrases per clause (cpC), complex nominals per clause (cnC), number of modifiers per noun phrase (SYNNP) and mean length of clause (MLC); and lexical complexity is measured by Measure of Textual Lexical Diversity (MTLD) and type-based lexical sophistication measure (LS2). SYNSTRUTt, which measures syntactic similarity at both clausal and phrasal levels, was also negatively loaded on phrasal elaboration. The negative loadings of SYNSTRUTt on both clausal subordination and phrasal elaboration suggests that the more complex a text is on clausal and phrasal levels, the less similar its syntactic constructions will be, though it seems that syntactic similarity as measured by SYNSTRUTt is more heavily influenced by clausal-level rather than phrasal-level complexity.

Figure 2 also shows quite clearly that some of the covariance of the observed variables is due to method effect, i.e., measurement approach. For instance, in the present study, one clausal subordination measure, i.e., dcC, as well as three of the four phrasal elaboration measures, use the number of clauses as their denominator. This resulted in a substantial amount of shared method variance among them. Norris and Ortega (2009) cautioned against the use of redundant measures that tap the same kind of complexity, as they provide redundant information which might be mistaken for robust evidence for the findings. In a similar vein, it could be argued that when using indices which share all or part of their numerator or denominator, it is important for researchers to point out that some of the consistency in the findings may be due to measurement effect, which is a kind of nonrandom measurement error and is thus construct-irrelevant.

5.2 Relationships Between the Sub-Constructs of L2 Linguistic Complexity

Figure 2 also presents the relationships among the three factors identified. None of the factor correlations exceeded 0.80, indicating that they were distinct variables (Brown, 2006). This finding lends support to the argument that complexity is not a single unified construct, and it is thus not reasonable to suppose that any single measure will adequately represent this construct (Biber, Gray, & Poonpon, 2011).

More importantly, the present study also offered empirical evidence regarding the relationships among the three factors. On the one hand, there was a moderate positive correlation ($r = 0.56$) between lexical complexity and phrasal elaboration. This relatively close relationship is not surprising, since phrasal elaboration is often realized through lexical means, for instance, the use of coordinate phrases, and the use of lexical constructions such as attributive adjectives, nouns and prepositional phrases as pre- or post-modifiers of nouns. It seems plausible that texts that are lexically diverse and sophisticated are also complex at phrasal level.

On the other hand, clausal subordination was much less associated with either phrasal elaboration or lexical complexity ($r = -0.16$; $r = -0.15$), and the weak correlations between them were also negative. Lu (2011) reported weak negative correlations between subordination measures and clause-based measures of phrasal elaboration, i.e., MLC, cnC and cpC. Skehan (2009) also found from L2 speakers' performance on a spoken narrative task that their lexical sophistication as measured by lambda, was negatively correlated with subordination-based measures of complexity. In the present study, clausal subordination was mainly assessed by the number of verb phrases, including both finite verbs and non-finite verbs, in a text. In other words, clausal subordination mainly represents achieving linguistic complexity through grammatical means; whereas phrasal elaboration and lexical complexity mainly represent linguistic complexification through lexical means. Thus, the findings of the present study seem to lend support to the hypothesis that lexical complexity and grammatical complexity constitute separate, independent dimensions of L2 performance and proficiency, and that there was a certain level of trade-off effect between them (Skehan, 2009). It should be noted that L1 research has also offered theoretical argument as well as empirical evidence concerning the differences between lexical and grammatical complexity. Theoretically, systemic functional linguistics posits that language development proceeds from expressing ideas by means of mostly parataxis, i.e., coordination; to an expansion by which hypotaxis, i.e., subordination, is added as a source to express the logical connection of ideas via grammatical intricate texts; and to finally the emergence of and reliance on grammatical metaphor (achieved through nominalization, among other processes), which leads to advanced language that exhibits lower levels of subordination but much higher levels of lexical density and more complex phrases (Halliday & Mathiessen, 1999). Empirical evidence for this hypothesized developmental trajectory, especially the change from grammatical complexity to lexical complexity, was provided by a series of corpus-based analyses of L1 conversational versus written discourse (Biber, 2006; Biber, Gray, & Poonpon, 2011). Specifically, it was found that written discourse such as academic writing was characterized by non-clausal features embedded in noun phrases, such as prepositional phrases as post-nominal modifiers and attributive adjectives and nouns as nominal pre-modifiers; whereas the most strongly favored type of structural complexity in conversation is finite dependent clauses functioning as constituents in other clauses (Biber, Gray, & Poonpon, 2011). Biber et al. argued that since grammatical features such as finite dependent clauses are frequently produced in conversation by all L1 speakers, they must be acquired at relatively early stage; whereas, many types of complex phrasal embedding which are produced mostly only in formal writing are likely to be acquired late, typically in adulthood. These findings and theoretical arguments suggest the distinctiveness of grammatical means and lexical means of complexification, to which the findings of the present study provide further empirical support.

6. Conclusion

Critical as linguistic complexity is to the investigation of L2 performance, L2 proficiency and L2 development, as a construct, it is still poorly defined in L2 research (Bulté & Housen, 2014). By applying exploratory and confirmatory factor analyses on a selection of popular measures of L2 linguistic complexity, the present study revealed the sub-constructs underlying these measures and the relationships among these sub-constructs. The distinctiveness of these sub-constructs offered solid empirical evidence for the multidimensionality of L2 linguistic complexity, and a strong argument against the idea of a one-size-fits-all measure of L2 linguistic complexity. Findings regarding the relationships among the three sub-constructs revealed that lexical complexity and phrasal elaboration are moderately correlated; while clausal subordination employs rather different means of complexification than that employed by phrasal elaboration and lexical complexity. Moreover, error variances among some of the measures indicate that method effect is hard to avoid when indices that are similar in the way of measurement are used together, and that caution is needed when interpreting the results of studies using such measures.

As the present study further clarified the constructs underlying some popular automated measures, their application in automated essay scoring systems seems to be more justifiable. Recent empirical studies on the development of L2 writing complexity found that there was a disassociation between L2 syntactic development and judgments of L2 writing quality. Specifically, while L2 learner growth was associated with greater nominal style and phrasal complexity, judgment of essay quality by human raters was based on structures aligned with spoken discourse, i.e., clausal complexity (Bulté & Housen, 2014; Crossley & McNamara, 2014). With regard to raters' judgment of lexical qualities of L2 essays, it is also found that raters were sensitive to accuracy, but not range or sophistication (Fritz & Ruegg, 2013). Therefore, it seems the use of automated means in the judgment of lexical complexity and phrasal elaboration could compensate for raters' lack of sensitivity to these sub-constructs of linguistic complexity.

Of course, automated indices are not without limitations. Despite the speed, flexibility and reliability that can be afforded by linguistic software tools such as Coh-Metrix and L2 Syntactic Complexity Analyzer, they may still be too rigid to accurately and fully identify, segment, and parse the L2 learner productions, thus creating measurement noise (Bulté & Housen, 2014). Therefore, empirical investigation of the reliability and validity of manual and automated measures of L2 linguistic complexity is of the essence if the field of L2 writing is to gain real understanding of the construct it is interested in. It is also desirable to use human rating to provide concurrent validity for the use of those indices provided by automatic text analysis software. However, no rating scale of syntactic complexity is currently in existence. Meanwhile, the present study is also limited in some other methodological aspects and thus caution is needed when interpreting the findings of the study. Specifically, the writer sample of the study was limited to intermediate-level L2 learners of a single L1 background. Moreover, the writing tasks were argumentative tasks only so that the findings may not be generalized to other rhetorical tasks. Future studies should consider assessing linguistic complexity using a variety of genres and involving learners of diverse L2 proficiency levels and L1 backgrounds so as to further clarify the construct definition of L2 linguistic complexity and improve the way it is operationalized.

References

Abell, N., Springer, D. W., & Kamata, A. (2009). *Developing and validating rapid assessment instruments*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195333367.001.0001

Ai, H., & Lu, X. (2010). *A web-based system for automatic measurement of lexical complexity*. Paper presented at the 27th Annual Symposium of the Computer-Assisted Language Consortium (CALICO-10). Amherst, MA. June 8−12.

Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In A. Diáz-Negrillo, N. Ballier & P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data* (pp. 249–264). Amsterdam: John Benjamins. https://doi.org/10.1075/scl.59.15ai

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.23

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*(1), 5–35. https://doi.org/10.5054/tq.2011.244483

Brown, T. A. (2006). *Confirmatory Analysis for Applied Research*. New York: The Gilford Press.

Bulté, B., & Housen, A. (2012). Defining and operationalizing L2 complexity. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency—Investigating complexity, accuracy and fluency in SLA* (pp. 21–46). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.32.02bul

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

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, *26*, 66–79. https://doi.org/10.1016/j.jslw.2014.09.006

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*, 139–155. https://doi.org/10.1016/1060-3743(95)90004-7

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

Fritz, E., & Ruegg, R. (2013). Rater sensitivity to lexical accuracy, sophistication and range when assessing writing. *Assessing Writing*, *18*, 173−181. https://doi.org/10.1016/j.asw.2013.02.001

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, *36*, 193–202. https://doi.org/10.3758/BF03195564

Halliday, M. A. K., & Matthiessen, C. (1999). *Construing experience through meaning: A language-based approach to cognition*. London, England: Cassell.

Hemplemann, C. F., Rus, V., Graesser, A. C., & McNamara, D. S. (2006). Evaluating state-of-the-art treebank-style parsers for Coh-Metrix and other learning technology environments. *Natural Language Engineering*, *12*, 131−144. https://doi.org/10.1017/S1351324906004207

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language*, *25*(2), 282–306. https://doi.org/10.1016/j.csl.2010.06.001

Jarvis, S. (2013). Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 13–45). Amsterdam: John Benjamins. https://doi.org/10.1075/sibil.47.03ch1

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*, 31–36. https://doi.org/10.1007/BF02291575

Kline, R. B. (2011). *Principles and practices of structural equation modeling* (2nd ed.). New York, NY: Guilford.

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction*, *1*(1), 60–69. https://doi.org/10.7820/vli.v01.1.koizumi

Levelt, W. J. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Liao, C., Kuo, B., Deenang, E., & Mok, M. (2015). Exploratory and confirmatory factor analyses in reading-related cognitive component among grade four students in Thailand. *Educational Psychology*, *36*(1), 1102–1114. https://doi.org/10.1080/01443410.2015.1058342

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures of college-level ESL Writers' Language Development. *TESOL Quarterly*, *45*(1), 36–62. https://doi.org/10.5054/tq.2011.240859

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing*, *34*(4), 493–511. https://doi.org/10.1177/0265532217710675

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for academic purposes programme. *Journal of Second Language Writing*, *29*, 3–15. https://doi.org/10.1016/j.jslw.2015.06.004

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Unpublished Doctoral dissertation. Memphis: University of Memphis.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, voc-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 188–205). Hershey, PA: IGI Global. https://doi.org/10.4018/978-1-60960-741-8.ch011

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

National College English Testing Committee. (2006). *Syllabus for College English Test*. Shanghai, China: Shanghai Language Education Press.

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

Ockey, G. J., & Choi, I. (2015). Structural equation modeling reporting practices for language assessment. *Language Assessment Quarterly*, *12*, 305–319. https://doi.org/10.1080/15434303.2015.1050101

Read, J. (2000). *Assessing vocabulary*. Oxford: Oxford University Press. https://doi.org/10.1017/CBO9780511732942

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching*, *43*, 1–32. https://doi.org/10.1515/iral.2005.43.1.1

Robinson, P. (2007). Criteria for classifying and sequencing pedagogic tasks. In M. P. García Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 7–26). Clevedon, UK: Multilingual Matters. https://doi.org/10.21832/9781853599286-004

Robinson, P. (2011). Second language task complexity, the cognition hypothesis, language learning, and performance. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 3–38). Amsterdam: John Benjamins.

https://doi.org/10.1075/tblt.2

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics*, *30*(4), 510–532. https://doi.org/10.1093/applin/amp047

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Upper Saddle River, NJ: Pearson Allyn & Bacon.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity* (Report No. 17). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Yang, H. (2009). *Exploring the complexity of second language writers' strategy use and performance on an integrated writing test through structural equation modeling and qualitative approaches*. Unpublished doctoral dissertation. Austin: The University of Texas at Austin.

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

Yoon, H. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System (Linköping)*, *66*, 130–141. https://doi.org/10.1016/j.system.2017.03.007

Notes

Note 1. The CET-4 scoring system has a total of 710 (M = 500, SD = 70) (National College English Testing Committee, 2006).

Note 2. The parse tree similarity (Sim) is computed by the following formula: Sim = nodes in the common tree/ (the sum of the nodes in the two sentence trees minus nodes in the common tree) (McNamara et al., 2014).

Copyrights