The Modern Language Journal

Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices

KRISTOPHER KYLE[1] and SCOTT A. CROSSLEY[2]
[1] *University of Hawai'i at Manoa, Second Language Studies, 1890 East-West Road, Honolulu, HI, 96822 Email: kkyle@hawaii.edu*
[2] *Georgia State University, Applied Linguistics, PO Box 4099, Atlanta, GA 30302–4099 Email: scrossley@gsu.edu*

Syntactic complexity is an important measure of second language (L2) writing proficiency (Larsen–Freeman, 1978; Lu, 2011). Large-grained indices such as the mean length of T-unit (MLTU) have been used with the most consistency in L2 writing studies (Ortega, 2003). Recently, indices such as MLTU have been criticized, both for the difficulty in interpretation (e.g., Norris & Ortega, 2009) and for a potentially misplaced focus on clausal subordination (e.g., Biber, Gray, & Poonpon, 2011). In this article, we attempt to address both of these criticisms by using traditional indices of syntactic complexity (e.g., MLTU), fine-grained indices of clausal complexity, and fine-grained indices of phrasal complexity to predict holistic scores of writing quality. In 4 studies, we used indices of each index type to predict holistic writing quality scores in independent essays on the Test of English as a Foreign Language (TOEFL). We then used all index types in a combined analysis to predict a holistic writing score. The results indicated that fine-grained indices of phrasal complexity were better predictors of writing quality than either traditional or fine-grained clausal indices, though a single fine-grained index of clausal complexity contributed to the combined model. These results provide some support for Biber et al.'s (2011) claims regarding complexity and academic L2 writing proficiency.

Keywords: assessment; syntactic complexity; natural language processing; L2 writing

COMPLEXITY HAS BEEN AN IMPORTANT measure of second language (L2) writing proficiency and development for the past 45 years. Drawing on previous work in first language (L1) writing proficiency (as measured by grade level; Hunt, 1965), early L2 researchers adopted complexity indices to investigate L2 proficiency (Larsen–Freeman, 1978; Larsen–Freeman & Strom, 1977). At the syntactic level, complexity has historically been operationalized through large-grained indices that measure complexity at the clause or sentence level (e.g., the length of clauses, T-units, and/or sentences). L2 research using such indices has indicated that more proficient language users produce longer and more varied syntactic structures (Lu, 2011; Ortega, 2003; Wolfe–Quintero, Inagaki, & Kim, 1998). The widespread use of large-grained syntactic complexity indices has recently been the topic of much interest and criticism in the wider field of applied linguistics (Biber et al., 2011; Bulté & Housen, 2012; Norris & Ortega, 2009). Among the criticisms leveled, two in particular have featured prominently in the literature. These include the difficulty in interpreting large-grained indices and a disproportionate focus on clausal complexity (e.g., clausal subordination).

A number of researchers have criticized the granularity (i.e., specificity) of large-grained syntactic complexity indices (Larsen–Freeman, 2009; Norris & Ortega, 2009; Wolfe–Quintero et al., 1998). While relatively consistent and positive relationships have been reported between measures such as mean length of T-unit (MLTU) and L2 development, very little is known about the specific structures that emerge as learners develop because large-grained indices are not sensitive

The Modern Language Journal, 102, 2, (2018)
DOI: 10.1111/modl.12468
0026-7902/18/333–349 $1.50/0

enough to provide such information. A number of linguistic structures, ranging from phrasal dependents such as adjectives and adverbs to dependent clauses, can increase the length of a T-unit, but the types of structures found within a T-unit are not captured by large-grained indices. This suggests that using fine-grained indices of syntactic complexity may provide a clearer understanding of the relationship between syntactic complexity and L2 writing proficiency. Furthermore, recent research in corpus linguistics has suggested that the disproportionate focus on large-grained clause units may be misplaced, especially in academic writing. Biber et al. (2011), for example, demonstrated that clausal subordination is a feature of informal speech, while complex noun phrases are a hallmark of academic writing. In this article, we provide an overview of the use of T-unit and clause-based indices in L2 writing studies, followed by the more recent research focusing on phrasal complexity indices. We then propose a number of fine-grained indices of clausal and phrasal complexity and discuss their operationalization. Finally, we use traditional indices of syntactic complexity (e.g., MLTU and mean length of clause, MLC), fine-grained indices of clausal complexity, and fine-grained indices of phrasal complexity to predict holistic scores of writing quality in a high-stakes writing assessment.

TRADITIONAL INDICES OF SYNTACTIC COMPLEXITY

Over the past 45 years, L2 writing researchers have primarily operationalized syntactic complexity using large-grained, length-based indices (Larsen–Freeman, 1978; Ortega, 2003; Wolfe–Quintero et al., 1998). In Ortega's (2003) research synthesis of L2 writing studies, for example, 25 of the 27 included studies operationalized syntactic complexity as MLTU either as the sole index or in combination with other indices. MLTU has also prominently featured as an index of proficiency (as measured by program levels and/or writing quality scores) and development in more recent studies (Cumming et al., 2005; Knoch, Rouhshad, & Storch, 2014; Lu, 2011). Although differences across studies have been observed, the general trend in the literature suggests that more proficient L2 learners tend to write longer T-units. This trend, coupled with the relative ease of calculating the measure by hand, provides evidence of the utility of MLTU as an index of writing proficiency and general linguistic development (as originally intended by Hunt, 1965). However useful MLTU may be as an index of proficiency, the interpretative usefulness of an MLTU score is opaque (Norris & Ortega, 2009).

The following two example T-units, which elaborate on the five-word T-unit *the man kicked the ball*, illustrate the difficulty of interpreting indices such as MLTU. Each T-unit includes 12 words, but the linguistic structures are very different: (a) *The athletic man in the jersey kicked the ball over the fence;* (b) *Because he wanted to score a goal, the man kicked the ball.* In the first example, the subject phrase is elaborated with an adjective (*athletic*) and a prepositional phrase (*in the jersey*). In addition, the verb phrase is elaborated with a locative prepositional phrase (*over the fence*). In the second example, the original T-unit is elaborated with a *because* subordinate clause (*because he wanted to score a goal*). These examples demonstrate the reductive nature of an MLTU score and indicate the difficulty in interpreting such a score. An MLTU score only gives a general indication of how elaborated a particular main clause is but says nothing about the type(s) of elaboration included. Another similar (and often used) measure is the mean length of clause (MLC) which may provide a general indication of phrasal complexity. However, as noted by Norris and Ortega (2009), this index is similarly flawed from an interpretation standpoint because clauses can be lengthened both by the addition of complements and through phrasal elaboration. Thus, a particular MLC score provides very little information with regard to the specific linguistic features of the clauses in a writing sample.

More clearly interpretable indices of syntactic complexity have been both proposed and employed in a number of studies (though with substantially less frequency than MLTU). Related measures of clausal subordination, such as the number of clauses per T-unit (C/TU) and dependent clauses per clause (DC/C), clearly indicate the amount of clausal subordination in a text (Lu, 2011; Ortega, 2003). Indices of clausal coordination, such as Bardovi–Harlig's (1992) coordination index or the related number of T-units per sentence (T/S; Lu, 2011; Wolfe–Quintero et al., 1998), also provide interpretable information regarding syntactic complexity, but are rarely used (Norris & Ortega, 2009). In addition to criticisms of opaque indices such as MLTU and MLC, recent investigations have questioned the extant focus on clausal complexity in the analysis of L2 writing proficiency.

PHRASAL COMPLEXITY AND ACADEMIC WRITING

Biber et al. (2011) challenged the use of indices related to clausal complexity based on results of a corpus analysis comparing the linguistic features of informal speech and academic writing. Using the Biber Tagger (Biber, 1988; Biber et al., 2004) to tag each corpus for fine-grained lexicogrammatical features (see Table 1) related to both clauses and phrases, Biber et al. found that clausal complexity was a distinctive feature of informal conversations, while phrasal complexity was a distinctive feature of academic writing. Based on this evidence, the researchers suggested that indices of clausal complexity were likely inappropriate for measuring L2 writing proficiency. They hypothesized that, as L2 academic writers become more proficient, their writing would move from being characterized by features of informal speech to features of academic writing. From the perspective of grammatical forms, the hypothesized developmental sequence would move from writing characterized by finite dependent clauses to nonfinite dependent clauses, and eventually to dependent phrases. From the perspective of syntactic functions, development would first be characterized by the addition of clausal constituents, and would then move to the addition of noun phrase modifiers.

A number of follow-up studies have explored whether the L1 corpus findings are applicable to the measurement of L2 writing proficiency (Biber, Gray, & Staples, 2014; Parkinson & Musgrave, 2014; Taguchi, Crawford, & Wetzel, 2013). Biber et al. (2014) conducted an analysis similar to the one conducted in Biber et al. (2011), but instead of analyzing L1 reference corpora, they analyzed responses to the speaking and writing performance tasks that are part of the Test of English as a Foreign Language (TOEFL). They divided the texts into four categories: independent and integrated speaking and independent and integrated writing. Generally, they found that similar differences existed between L2 texts as were found in the L1 texts used in Biber et al. (2011). For instance, writing samples were reported to have more complexity at the phrasal level (particularly with regard to noun phrases), while spoken texts include more finite clauses and *verb* + *to* constructions. In terms of writing proficiency, the results indicated that only two indices significantly interacted with holistic scores: High-scoring written integrated texts included more attributive adjectives and *verb* + *that* clause constructions than lower-scoring written integrated texts. Additionally, a component score consisting of a number of features (including both lexical features and features of phrasal complexity) explained 15.2% of the variance in TOEFL independent essay scores.

TABLE 1
Clausal and Phrasal Features Analyzed in Biber et al. (2011)

Category	Index
Finite adverbial clauses	Total finite adverbial clauses
	Because clause
	If clause
	Although clause
Finite complement clauses	verb + *that* clause
	verb + *WH* clause
	adjective + *that* clause
	noun + *that* clause
Finite noun modifier clauses	*that* relative clauses
	WH relative clauses
Nonfinite adverbial clauses	*to* adverbial clause
Nonfinite complement clauses	verb + *-ing* clause
	verb + *to* clause
	adjective + *-ing* clause
	adjective + *to* clause
	noun + *of* + *-ing* clause
	noun + *to* clause
Nonfinite noun modifier clauses	nonfinite relative clause
Adverbials	adverbs as adverbials
	prepositional phrases as adverbials
Noun modifiers	attributive adjectives
	nouns as nominal premodifiers
	total prepositional phrases as nominal modifiers
	of as postmodifier
	in as postmodifier
	on as postmodifier
	with as postmodifier
	for as postmodifier

Taguchi et al. (2013) investigated differences in clausal complexity and phrasal complexity between high- and low-scoring L2 placement essays. They report that high-scoring essays tended to include more features of phrasal complexity (e.g., more post noun modifying prepositional phrases) and fewer features of clausal complexity (e.g., *that* relative clauses) than low-scoring essays. Taguchi et al. only report descriptive statistics, however, making the scale of differences between high- and low-scoring essays difficult to interpret. In a related study, Parkinson and Musgrave (2014) investigated differences in clausal and phrasal features

between essays written by L2 English for Academic Purposes (EAP) students and Master's (MA) students writing in English as an L2. They found that EAP students showed characteristics of lower levels of development (i.e., reliance on attributive adjectives), while the MA students demonstrated the characteristics of higher levels of development (e.g., phrasal modifiers). However, because writing prompt, genre, and writing proficiency level were not controlled for, the implications of the study are difficult to determine.

Other studies have used Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, Graesser, McCarthy, & Cai, 2014) indices related to phrasal complexity to measure L2 writing proficiency and development. Crossley and McNamara (2014), for example, reported significant growth in the number of words before the main verb (NW→MV) and the number of modifiers per noun phrase (M/NP) between essays written at the beginning and the end of a semester ($p = .007$; $\eta^2_p = .122$). They also found a positive relationship ($p = .023$, $r = .213$) between M/NP and writing quality scores, suggesting that higher-quality essays tend to include more noun phrase modifiers. Guo, Crossley, & McNamara (2013) also found a positive relationship between M/NP and holistic scores of writing quality in TOEFL integrated ($r = .264$) and independent ($r = .377$) essays. Lu (2010, 2011) also used phrasal complexity indices (i.e., complex nominals per clause [CN/C] and complex nominals per T-unit [CN/T]), to examine differences between college levels in a large learner corpus study. Lu found that students at higher levels tended to include more complex nominals, as indicated by significant differences between complex nominal use across most levels. Unlike the fine-grained indices provided by the Biber Tagger, indices such as M/NP and CN/T still face limitations with regard to interpretation. Nonetheless, these findings provide further evidence of the importance of phrasal complexity features in the development of L2 writing.

CURRENT STUDY

In this study, we extend previous work by examining the relationship between fine-grained indices of clausal and phrasal complexity and holistic scores of writing quality using a newly developed and freely available text analysis tool. To provide a baseline analysis, we also compare their performance with traditional indices of syntactic complexity (e.g., MLTU and MLC). This study is guided by the following research questions:

RQ1. What is the relationship between traditional indices of syntactic complexity and holistic scores of writing quality?

RQ2. What is the relationship between fine-grained indices of clausal complexity and holistic scores of writing quality?

RQ3. What is the relationship between fine-grained indices of phrasal complexity and holistic scores of writing quality?

RQ4. What is the complementary relationship between the three types of syntactic complexity indices and holistic scores of writing quality?

METHOD

Learner Corpus

The written proficiency corpus used in this study is made up of argumentative essays written as part of the Test of English as a Foreign Language (TOEFL). The essays comprise responses to two independent prompts (240 texts each) that asked test takers to compose an essay that asserts and defends an opinion on a particular topic based on life experience (see Table 2). Test takers were given 30 minutes to complete the writing task and were expected to produce at least 300 words. See Table 3 for an overview of this corpus.

Each essay was given a score on a 5-point scale by at least two raters trained by ETS. If the scores given by the raters differed by 1 point or less, scores were averaged. If any two scores given by raters differed by more than 1 point, a third rater was consulted in order to adjudicate the score. Scores range from 1.0 to 5.0 in .5 point intervals.

TABLE 2
Writing Prompts for Independent Essays in TOEFL Public Use Dataset

Test Form	Prompt Instructions
1	Do you agree or disagree with the following statement? It is more important to choose to study subjects you are interested in than to choose subjects to prepare for a job or career. Use specific reasons and examples to support your answer.
2	Do you agree or disagree with the following statement? In today's world, the ability to cooperate well with others is far more important than it was in the past. Use specific reasons and examples to support your answer.

TABLE 3
Overview of Writing Proficiency Corpus

Prompt	*N*	Number of Words	Mean Score	Standard Deviation
1	240	77,238	3.83	0.86
2	240	74,252	3.47	0.91

TABLE 4
Abbreviated TOEFL Rubric for Independent Writing Tasks

Score	Descriptors
5	An essay at this level largely accomplishes all of the following: a) effectively addresses the topic and task b) is well organized and well developed, using clearly appropriate explanations, exemplifications, and/or details, c) displays unity, progression, and coherence, d) displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors
2	An essay at this level may reveal one or more of the following weaknesses: a) limited development in response to the topic and task, b) inadequate organization of connection of ideas, c) inappropriate or insufficient exemplifications, explanations, or details to support or illustrate generalizations in response to the task, d) a noticeably inappropriate choice of words or word forms, e) an accumulation of errors in sentence structure and/or usage

The holistic rating score used included descriptors related to the completion of the task, organization, development of ideas, coherence, word use, and syntax. See Table 4 for the score descriptors for low- and high-proficiency essays.

Syntactic Complexity Indices

One practical barrier in the calculation of indices of syntactic complexity is the time, effort, and expertise required to complete them. Lu (2010), for example, reports that annotators spent 2 hours on average per 315-word essay counting eight structure types (e.g., clauses, T-units, complex nominals). Given that the manual analysis of fine-grained linguistic analyses requires even more time and expertise, it is no surprise that researchers often opt for indices that are more efficient to calculate and are common in the literature. In recent years,[1] however, technological advancements in the field of natural language processing (e.g., Charniak & Johnson, 2005; Chen & Manning, 2014; Klein & Manning, 2003) have paved the way for the accurate automatic annotation of linguistic features. Publicly available tools such as Coh-Metrix (Graesser et al., 2004; McNamara et al., 2014), L2 syntactic complexity analyzer (SCA; Lu, 2010, 2011), and the recently developed Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC; Kyle, 2016) harness such advancements. These tools allow researchers to quickly analyze both traditional measures of syntactic complexity (i.e., SCA) and more fine-grained indices of phrasal and clausal complexity (i.e., TAASSC).

Traditional Syntactic Complexity Indices. To compute traditional syntactic complexity measures, we used Lu's (2010) L2 complexity analyzer (SCA) indices, which are reported by TAASSC. SCA includes 14 indices of syntactic complexity drawn from Wolfe–Quintero et al. (1998) and Ortega (2003), among others. Table 5 includes a description of each of the structures counted by SCA, and Table 6 comprises a list of the 14 SCA indices including a short description of each. For further information, refer to Lu (2010, 2011).

Fine-Grained Clausal Indices. We consider 30 fine-grained indices of clausal complexity. Twenty-eight indices calculate the average number of particular structures per clause. The fine-grained clausal indices included in TAASSC differ from traditional indices of syntactic complexity (i.e., MLC, DC/C, CP/C, and CN/C) included in SCA in three main ways. First, TAASSC counts the length of clauses as the number of direct dependents per clause instead of the number of words. This prevents structures that inherently include more words (e.g., prepositional phrases) to be given more weight than those that do not (e.g., adjectives). Second, instead of grouping structures (such as dependent clauses or complex nominals) together, TAASSC counts each type separately. Finally, TAASSC identifies clauses as a main verb and its associated structures (e.g., arguments

TABLE 5
A Description of Syntactic Structures Counted by SCA

Structure	Description	Examples
Word	a sequence of letters that are bounded by white space	*I* *ate*
Verb phrase	a finite or nonfinite verb phrase that is dominated by a clause marker	*ate pizza* *was hungry*
Complex nominal	a) nouns with modifiers b) nominal clauses c) gerunds and infinitives that function as subjects	*a) red car* *b)* I know *that she is hungry* *c) Running* is invigorating
Coordinate phrase	adjective, adverb, noun and verb phrases connected by a coordinating conjunction	She *eats pizza and smiles*
Clause	a syntactic structure with a subject and a finite verb	*I ate pizza* *because I was hungry*
Dependent clause	a finite clause that is a nominal, adverbial, or adjective clause	I ate pizza *because I was hungry*
T-unit	an independent clause and any clauses dependent on it	*I ate pizza* *I ate pizza because I was hungry*
Complex T-unit	a T-unit that includes a dependent clause	*I ate pizza because I was hungry*
Sentence	a group of words bounded by sentence-ending punctuation (., ?, !, ", …)	*I went running today.*

TABLE 6
A Description of SCA Variables

Index Abbreviation	Index Name	Index Description
MLS	mean length of sentence	number of words per sentence
MLT	mean length of T-unit	number of words per T-unit
MLC	mean length of clause	number of words per clause
C/S	clauses per sentence	number of clauses per sentence
VP/T	verb phrases per T-unit	number of verb phrases per sentence
C/T	clauses per T-unit	number of clauses per T-unit
DC/C	dependent clauses per clause	number of dependent clauses per clause
DC/T	dependent clauses per T-unit	number of dependent clauses per T-unit
T/S	T-units per sentence	number of T-units per sentence
CT/T	complex T-unit ratio	number of complex T-units divided by T-units
CP/T	coordinate phrases per T-unit	number of coordinate phrases per T-unit
CP/C	coordinate phrases per clause	number of coordinate phrases per clause
CN/T	complex nominals per T-unit	number of complex nominals per T-unit
CN/C	complex nominals per clause	number of complex nominals per clause

and adverbials), which may or may not include a finite verb. This operationalization diverges from Lu (2011) but is in line with previous research (Bardovi–Harlig & Bofman, 1989; Wolfe–Quintero et al., 1998).

TAASSC also includes two more general indices of clausal complexity. These indices take into account the total number of dependents per clause. The first index represents the average number of dependents per clause, while the second represents the standard deviation of the number of dependents per clause, which provides a measure of syntactic variation. Table 7 comprises a description of each of the fine-grained indices of clausal complexity in TAASSC.

Fine-Grained Phrasal Indices. TAASSC includes phrasal indices for seven noun phrase types and ten phrasal dependent types (see Table 8 for an overview of these structures). Three types of phrasal indices are included in TAASSC. The first type calculates the average number of dependents per each phrase type (e.g., nominal subjects) and for all phrase types. The second type calculates the

TABLE 7
Clausal Dependent Types Analyzed by TAASSC

Structure	Abbreviation	Description	Example of Structure
Adjective complement	acomp	An adjective that functions as a complement in a copular clause	*She [looks]$_{gov}$ [beautiful]$_{acomp}$*
Adverb modifier	advmod	A nonclausal adverb or adverb phrase that modifies a verb phrase	*[Accordingly]$_{advmod}$, I [ate]$_{gov}$ pizza.*
Adverbial clause	advcl	A clause that modifies a verb phrase	*The accident [happened]$_{gov}$ [as night fell]$_{advcl}$*
Agent	agent	The conceptual subject in a passive clause, which is introduced by the word "by"	*The man has been [killed]$_{gov}$ by the [police]$_{agent}$*
Auxiliary verb	aux	A verb that is not the main verb in a clause (e.g., aspect marker, modal verb, etc.)	*He[is]$_{aux}$ [running]$_{gov}$*
Bare noun phrase temporal modifier	tmod	A nonprepositional phrase noun modifier that specifies a time	*Last [night]$_{tmod}$, I [swam]$_{gov}$ in the pool*
Clausal complement	ccomp	A dependent clause that serves as a complement	*I am [certain]$_{gov}$ [that he did it]$_{ccomp}$*
Clausal coordination	cc	Clauses joined by a coordinating conjunction	*[Jill runs]$_{gov}$ and [Jack jumps]$_{cc}$*
Clausal prepositional complement	pcomp	Clausal complement that consists of a prepositional phrase that includes a clausal prepositional object	*They [heard]$_{gov}$ [about you missing classes]$_{pcomp}$*
Clausal subject	csubj	A clause that functions as the subject of another clause	*[What she said]$_{csubj}$ [is]$_{gov}$ not true*
Conjunction	conj	A verb phrase that includes a coordinating conjunction	*He [runs]$_{gov}$ and [jumps]$_{conj}$*
Direct object	dobj	A predicative noun phrase that is the recipient of the action of a transitive verb	*She [gave]$_{gov}$ me a [raise]$_{dobj}$*
Discourse marker	discourse	Fillers, interjections, and discourse markers that are not directly linked to the structure of the sentence	*[Well]$_{discourse}$, I [like]$_{gov}$ pizza*
Existential "there"	expl	"There" that functions as the subject of a clause	*[There]$_{expl}$ may [be]$_{gov}$ a solution to the problem*
Indirect object	iobj	A noun phrase that functions as the dative object of the verb	*She [gave]$_{gov}$ [me]$_{iobj}$ a raise*
Modal auxiliary	modal	Modal verbs such as "may," "might," "could," and "should"	*He [may]$_{modal}$ [be]$_{gov}$ awesome.*
Negation	neg	A verb phrase that is negated	*He did [not]$_{neg}$ [kill]$_{gov}$ them.*
Nominal complement	ncomp	A noun or noun phrase that functions as a complement in a copular clause	*He [is]$_{gov}$ a [teacher]$_{ncomp}$*
Nominal subject	nsubj	A subject of a (nonpassive) clause that is a noun phrase	*The [baby]$_{nsubj}$ [is]$_{gov}$ cute* *The [athlete]$_{nsubj}$ [ran$_{gov}$] quickly*
Open clausal complement	xcomp	A nonfinite clausal complement	*I am [ready]$_{gov}$ [to leave]$_{xcomp}$* *[Going]$_{gov}$ [running]$_{xcomp}$ is fun*
Parataxis	parataxis	Clauses or phrases inserted into or placed next to a clause with no explicit markers of coordination or subordination	*That man, Jack [continued]$_{parataxis}$, [is]$_{gov}$ dangerous*
Passive auxiliary verb	auxpass	An auxiliary verb in a passive clause	*Kennedy has [been]$_{auxpass}$ [killed]$_{gov}$*

TABLE 7 (*Continued*)

Structure	Abbreviation	Description	Example of Structure
Passive clausal subject	csubjpass	A clause that serves as the syntactic subject of a passive clause	*[That she lied]$_{csubjpass}$ was [suspected]$_{gov}$ by everyone*
Passive nominal subject	nsubjpass	A noun phrase that serves as the syntactic subject of a passive clause	*[Dole]$_{nsubjpass}$ was defeated$_{gov}$ by Clinton*
Phrasal verb particle	prt	The particle in a phrasal verb	*They [gave]$_{gov}$ [up]$_{prt}$ the fight*
Prepositional modifier	prep_*	A prepositional phrase that modifies the verb	*They [went]$_{gov}$ [into the store]$_{prep\_into}$*
Subordinating conjunction	mark	A subordinating conjunction that marks a subordinate clause	*Forces engaged in fighting [after]$_{mark}$ insurgents [attacked]$_{gov}$*
Undefined dependent	dep	A clausal dependent that could not be identified by the parser	*N/A*

Note. "gov" represents the governor of the dependent; *prepositional modifier representations include the actual preposition.

occurrence of particular dependent types (e.g., adjective modifiers) regardless of the type of noun phrase they occur in. The final phrasal index type calculates the average occurrence of particular dependent types in particular types of noun phrases (e.g., adjective modifiers occurring in nominal subjects).

Noun phrases in English can consist of pronouns; except in very rare cases, pronouns do not take direct dependents (relative clauses being an exception). Due to the potential for pronouns as phrases to skew counts of dependents, TAASSC includes two versions of each index, one that includes pronoun noun phrases in its counts and one that does not. In this study, only the TAASSC indices that ignore pronouns were used.

Basic TAASSC phrasal indices represent the average number of phrasal dependents per phrase type (e.g., the average number of dependents per nominal subject). The sentence *The man in the red hat gave that tall man some money*, for example, includes four nominal phrases, which are a nominal subject, a prepositional object, an indirect object, and a direct object. Together, these four nominal phrases include four determiners, two adjective modifiers, and one prepositional phrase for a total of seven phrasal dependents. The average number of adjective modifier dependents per nominal, for example, is .5 (2/4).

For the largest-grained indices, standard deviations are also calculated. In a normal distribution of data, standard deviations indicate how far values from the mean must be to include 68.2% of the data. While a mean value indicates central tendencies in a dataset, standard deviations indicate how well the mean represents the data. In syntactic analysis, standard deviations can be used to measure variation.

Overall, we investigated 66 indices of phrasal complexity (and variation) calculated by TAASSC. See Table 9 for an overview of the phrasal indices.

Analysis

In order to investigate the relationship between indices of syntactic complexity and holistic scores of writing quality in TOEFL independent essays, a stepwise multiple linear regression analysis was conducted for each index type (i.e., traditional, fine-grained clausal, and fine-grained phrasal indices), and then with all three index types. First, normality was checked using the visualization component of the WEKA statistical package (Hall et al., 2009). Any variables that violated a normal distribution were discarded. Pearson correlations were then conducted on the remaining variables to determine whether they were meaningfully correlated with holistic essay score. Any variables that did not reach an absolute correlation value of $r \geq .100$ with the holistic essay score (which represents the threshold for a "small" effect according to Cohen, 1988) and a significance of $p < .001$ were removed from further consideration. Next, the remaining variables were checked for multi-collinearity to ensure that the final model consisted only of unique indices and that multicollinear indices did not exaggerate the results of the multiple regression analysis (Tabachnick & Fidell, 2014). For each pair of variables with

TABLE 8
Phrase Types and Dependent Types Analyzed by TAASSC

Structure	Abbreviation	Description	Example of Structure
Phrase types			
Nominal subject	nsubj	A subject of a (nonpassive) clause that is a noun phrase	*[The man in the red hat]$_{nsubj}$ gave that tall man some money.*
Passive nominal subject	nsubj_pass	A noun phrase that serves as the syntactic subject of a passive clause	*[The tall man]$_{nsubj\_pass}$ was given money by the man in the red hat*
Agent	agent	The conceptual subject in a passive clause, which is introduced by the word "by"	*The tall man was given money by [the man in the red hat]$_{agent}$*
Nominal complement	ncomp	A noun or noun phrase that functions as a complement in a copular clause	*He is [a tall man]$_{ncomp}$*
Direct object	dobj	A predicative noun phrase that is the recipient of the action of a transitive verb	*The man in the red hat gave that tall man [some money]$_{dobj}$.*
Indirect object	iobj	A noun phrase that functions as the dative object of the verb	*The man in the red hat gave [that tall man]$_{iobj}$ some money*
Prepositional object	pobj	A noun or noun phrase that functions as the object of a prepositional phrase	*The man in [the red hat]$_{pobj}$ gave that tall man some money*
Dependent types			
Determiners	det	Articles, demonstratives, and quantifiers	*[The]$_{det}$ man in [the]$_{det}$ red hat gave [that]$_{det}$ tall man [some]$_{det}$ money*
Adjective modifiers	amod	An adjective that modifies a noun or noun phrase	*The man in the [red]$_{amod}$ hat gave that [tall]$_{amod}$ man some money*
Prepositional phrases	prep	A prepositional phrase that modifies a noun or noun phrase	*The man [in the red hat]$_{prep}$ gave that tall man some money*
Possessives	poss	A possessive pronoun or noun with a possessive "s" that modifies a noun or noun phrase	*That is [her]$_{poss}$ red car [Charlie's]$_{poss}$ car is red.*
Verbal modifiers	vmod	A nonfinite verb or verb phrase that modifies a noun or noun phrase	*I don't have anything [to say]$_{vmod}$ to you*
Nouns as modifiers	nn	A noun that modifies a noun or noun phrase	*[Oil]$_{nn}$ prices are rising*
Relative clause modifiers	rcmod	A relative clause is a clause that modifies a noun or noun phrase, and is often (but not always) marked by a "wh" word	*I saw the person [you love]$_{rcmod}$ The person [who brought pizza]$_{rcmod}$ is a hero.*
Adverbial modifiers	advmod	An adverb that modifies a noun or noun phrase	*Today was a [really]$_{\text{advmod}}$ hot day*
Conjunction "and"	conj_and	The conjunction "and" when used to join two nouns or noun phrases	*Jack [and]$_{conj\_and}$ Jill*
Conjunction "or"	conj_or	The conjunction "or" when used to join two nouns or noun phrases	*Jack [or]$_{conj\_or}$ Jill*

absolute correlation values of $r >= .700$, only the variable with the highest correlation with holistic score was kept (Crossley, Salsbury, & McNamara, 2012).

The remaining variables were entered into a stepwise multiple regression that used the Akaike information criterion (AIC) method (Akaike, 1974). Finally, a follow-up 10-fold forced entry

TABLE 9

An Overview of the Phrasal Indices Included in TAASSC Index Type	Average	Standard Deviation	Total
Number of dependents per nominal	8	8	16
Occurrence of particular dependents	10		10
Occurrence of particular dependents per particular nominal	40		40
Total	58	8	66

TABLE 10
Correlations Between Holistic Essay Score and SCA Variables Entered Into Regression Model

Variable	Prompt 1 Mean *(SD)*	Prompt 2 Mean *(SD)*	Combined Mean *(SD)*	Correlation With Holistic Score
Mean length of clause	9.154 (1.744)	10.307 (1.961)	9.730 (1.942)	0.240
Coordinate phrases per clause	0.190 (0.112)	0.201 (0.119)	0.196 (0.115)	0.190

TABLE 11
Summary of SCA Multiple Regression Model

Entry	Predictors Included	*R*	R^2	R^2 Change	β	*SE*	*B*
1	Mean length of clause	.240	.058	.058	.110	.201	.240

Note. Estimated constant term = 2.360; β = unstandardized beta; *SE* = standard error; *B* = standardized beta.

linear regression was conducted using the indices included in the stepwise model to ensure that the model was consistent across the dataset. The next step in the statistical analysis was to determine how generalizable the model was across topics by comparing the multiple regression models between prompts using a Fisher *r* to *z* transformation. This analysis tests whether the differences between two correlation values are due to chance (Dunn & Clark, 1969). Finally, comparisons between the SCA and TAASSC models were conducted using a Fisher *r* to *z* transformation to examine which models explained a greater amount of variance.

RESULTS

Traditional Syntactic Complexity Indices

All 14 of Lu's (2010, 2011) SCA indices demonstrated normal distributions. Eleven of these indices did not reach the minimum thresholds of $r \geq 0.100$ and $p < .001$ with TOEFL essay quality scores and were removed from the analysis. Of the remaining three variables, one (complex nominals per clause) was removed due to multicollinearity with mean length of clause. The remaining two variables (mean length of clause and coordinate phrases per clause) were entered into a stepwise regression (see Table 10 for descriptive statistics and correlations between these variables and the human ratings). The resulting model, which included one variable (mean length of clause), explained 5.8% ($r = .240$, $R^2 = .058$) of the variance in holistic essay scores (see Table 11 for the model). The 10-fold cross-validated model explained 8.2% of the variance in holistic essay scores. The model explained 2.7% ($r = .163$, $R^2 = .027$) of the variance in prompt 1 scores and 8.9% ($r = .298$, $R^2 = .089$) of the variance in prompt 2 scores. A Fisher's *r* to *z* transformation indicated that the amount of variance explained by the model across the two prompts did not differ significantly ($z = -1.56$, $p = .119$).

TABLE 12
Correlation Between Holistic Essay Score and Clausal Complexity Variables Entered Into Regression

Variable	Prompt 1 Mean *(SD)*	Prompt 2 Mean *(SD)*	Combined Mean *(SD)*	Correlation With Holistic Score
Nominal subjects per clause	0.656 (0.090)	0.655 (0.097)	0.656 (0.093)	−0.172

TABLE 13
Summary of Clausal Complexity Multiple Regression Model

Entry	Predictors Included	*r*	R^2	R^2 change	β	*SE*	*B*
1	nominal subjects per clause	.172	.030	.030	−1.639	.429	−.172

Note. Estimated constant term = 4.501; β = unstandardized beta; SE = standard error; B = standardized beta.

Fine-Grained Clausal Complexity

Sixteen clausal complexity indices violated the assumption of normality and were removed from further consideration. Of the remaining 15 variables, 14 did not reach the minimum correlation thresholds of $r >= 0.100$ and $p < .001$ and were removed from further consideration. The remaining variable (nominal subjects per clause; see Table 12 for descriptive statistics and the correlation between this variable and the human ratings) was entered into a stepwise regression. The resulting model explained 3.0% ($r = .172$, $R^2 = .030$) of the variance in holistic essay scores (see Table 13 for the model). The 10-fold cross-validated model explained 4.9% ($r = .222$, $R^2 = .049$) of the variance, suggesting that the model was consistent across the dataset. The model explained 2.2% ($r = .147$, $R^2 = .022$) of the variance in prompt 1 scores and 3.8% ($r = .194$, $R^2 = .038$) of the variance in prompt 2 scores. A Fisher's *r* to *z* transformation indicated that the amount of variance explained by the model across the two prompts did not differ significantly ($z = -.530$, $p = .596$).

Fine-Grained Phrasal Complexity

Thirty-nine phrasal complexity indices violated the assumption of normality and were removed from further consideration.[2] Of the remaining 22 variables, 9 did not reach the minimum correlation thresholds of $r >= 0.100$ and $p < .001$ and were removed from further consideration. Of the remaining 13 variables, 1 was removed due to multi-collinearity. The remaining 12 variables (see Table 14) were entered into a stepwise regression. The resulting model, which included six variables, explained 18.9% ($r = .435$, $R^2 = .189$) of the variance in holistic essay scores (see Table 15 for the model). The model indicated that indices related to phrasal elaboration were predictive of essay quality. The 10-fold cross-validated model explained 18.7% ($r = .432$, $R^2 = .187$) of the variance, suggesting that the model was consistent across the dataset. The model explained 14.5% ($r = .381$, $R^2 = .145$) of the variance in prompt 1 scores and 22.9% ($r = .479$, $R^2 = .229$) of the variance in prompt 2 scores. A Fisher's *r* to *z* transformation indicated that the amount of variance explained by the model across the two prompts did not differ significantly ($z = -1.31$, $p = .190$).

Combined Analysis

The 15 indices entered into the previous models were then considered together. All indices were normally distributed, none of them were collinear, and, therefore, all were entered into a stepwise regression. The resulting model, based on seven indices, explained 20.3% of the variance ($r = .450$; $R^2 = .203$) in holistic scores of writing quality (see Table 16 for the model). The model indicated that indices related to phrasal elaboration and the use of nonfinite clauses were predictive of essay quality. The 10-fold cross-validated model explained 19.6% ($r = .443$, $R^2 = .196$) of the variance, suggesting that the model was consistent across the dataset. The model explained 15.7% ($r = .396$, $R^2 = .157$) of the variance in prompt 1 scores and 24.4% ($r = .494$, $R^2 = .244$) of the variance in prompt 2 scores. A Fisher's

TABLE 14
Correlations Between Holistic Essay Score and Phrasal Complexity Variables Entered Into Regression

Variable	Prompt 1 Mean *(SD)*	Prompt 2 Mean *(SD)*	Combined Mean *(SD)*	Correlation With Holistic Score
Dependents per nominal	1.040 (0.173)	1.055 (0.175)	1.048 (0.174)	0.311
Dependents per object of the preposition	0.980 (0.215)	0.934 (0.215)	0.957 (0.216)	0.301
Prepositions per nominal	0.119 (0.050)	0.144 (0.059)	0.131 (0.056)	0.289
Prepositions per object of the preposition	0.118 (0.074)	0.121 (0.072)	0.120 (0.073)	0.278
Adjectival modifiers per object of the preposition	0.149 (0.091)	0.183 (0.096)	0.166 (0.095)	0.263
Dependents per object of the preposition (*SD)*	0.938 (0.162)	0.910 (0.162)	0.924 (0.163)	0.256
Dependents per nominal (*SD*)	1.036 (0.141)	1.041 (0.148)	1.038 (0.145)	0.248
Dependents per direct object (*SD*)	0.960 (0.199)	0.991 (0.258)	0.975 (0.231)	0.245
Dependents per direct object	1.197 (0.307)	1.276 (0.340)	1.237 (0.326)	0.233
Dependents per nominal subject (*SD*)	0.869 (0.256)	0.952 (0.231)	0.911 (0.247)	0.206
Determiners per nominal	0.309 (0.093)	0.343 (0.091)	0.326 (0.093)	0.152
Adjectival modifiers per direct object	0.219 (0.127)	0.271 (0.156)	0.245 (0.144)	0.151

TABLE 15
Summary of Phrasal Complexity Multiple Regression Model

Entry	Predictors Included	*r*	R^2	R^2 Change	β	*SE*	*B*
1	Dependents per object of the preposition	0.301	0.091	0.091	0.562	0.215	0.137
2	Prepositions per object of the preposition	0.349	0.122	0.031	1.901	0.551	0.156
3	Adjectival modifiers per object of the preposition	0.368	0.135	0.014	0.916	0.469	0.098
4	Dependents per direct object (*SD*)	0.403	0.163	0.028	0.492	0.179	0.128
5	Dependents per direct object (no pronouns)	0.412	0.169	0.007	0.207	0.129	0.076
6	Dependents per nominal subject (*SD*)	0.435	0.189	0.020	0.512	0.152	0.143

Note. Estimated constant term = 1.307; β = unstandardized beta; *SE* = standard error; *B* = standardized beta.

TABLE 16
Summary of Combined Multiple Regression Model

Entry	Predictors Included	*r*	R^2	R^2 Change	β	*SE*	*B*
1	Dependents per object of the preposition	0.301	0.091	0.091	0.565	0.214	0.138
2	Prepositions per object of the preposition	0.349	0.122	0.031	1.756	0.549	0.145
3	Adjectival modifiers per object of the preposition	0.368	0.135	0.014	0.725	0.470	0.078
4	Dependents per direct object (*SD*)	0.403	0.163	0.028	0.508	0.178	0.132
5	Dependents per direct object	0.412	0.169	0.007	0.192	0.129	0.070
6	Dependents per nominal subject (*SD*)	0.435	0.189	0.020	0.544	0.151	0.152
7	Nominal subjects per clause	0.450	0.203	0.014	−1.141	0.401	−.120

Note. Estimated constant term = 2.076; β = unstandardized beta; *SE* = standard error; *B* = standardized beta.

TABLE 17
Examples from TOEFL Essays: Mean Length of Clause

Score	Example	Length of Clause
1	I selected agree to this question.	6
	Because I regret it.	4
		Mean = 5
5	With this in mind, it is still possible to argue that	11
	colleges do not exist for the sole purpose of producing effective social agents	13
		Mean = 12

r to *z* transformation indicated that the amount of variance explained by the model across the two prompts did not differ significantly ($z = -1.33$, $p = .184$).

DISCUSSION

The use of large-grained syntactic complexity indices has recently been the topic of much interest and criticism (Biber et al., 2011; Bulté & Housen, 2012; Norris & Ortega, 2009). Two criticisms, in particular, have gained traction: (a) Large-grained indices are difficult to interpret, and (b) there has been a disproportionate focus on clausal complexity (e.g., clausal subordination) at the expense of phrasal complexity indices. One reason that researchers have relied on large-grained indices such as MLTU is the amount of time and effort involved in the calculation of more fine-grained indices. The recent development of freely available text analysis tools (e.g., Kyle, 2016; Lu, 2011), however, has made it possible to efficiently calculate fine-grained indices of syntactic complexity.

This article investigates the two criticisms leveled at large-grained indices by comparing traditional indices of syntactic complexity, fine grained-indices of clausal complexity, and fine-grained indices of phrasal complexity ability to model holistic scores of writing quality in TOEFL independent essays. The results suggest that indices of phrasal complexity are the best syntactic predictors of writing quality scores, though a fine-grained clausal complexity index provided complementary explanatory power. These findings generally support Biber et al.'s (2011) claims regarding the importance of phrasal complexity over clausal complexity in academic writing. Furthermore, these results support the use of fine-grained indices that indicate the use of particular linguistic structures (e.g., Norris & Ortega, 2009). The following discussion elaborates on each model.

Traditional Syntactic Complexity Indices

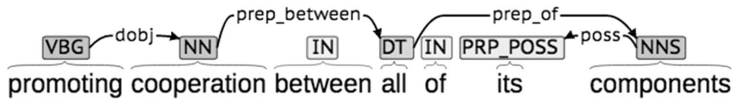
The relationship between indices of syntactic complexity calculated by the SCA and TOEFL independent essay scores was significant, but small. None of the indices that measure clausal subordination directly met the selection criteria. This provides support for Biber et al.'s (2011) corpus-based suggestion that clausal subordination may be inappropriate for measuring L2 academic writing proficiency. Two indices met the index selection criteria (MLC and CP/C), and a single index (MLC) was included in the predictor model. The results indicate that higher rated essays tend to include longer clauses. These findings support previous studies, such as Lu (2010, 2011), who found similar results across university levels (i.e., as university level increased, writers used longer clauses and more coordinate phrases per clause). These results also align with the findings from Ortega's (2003) synthesis of L2 writing studies, which found either neutral or positive relationships between MLC and writing proficiency. Overall, however, these differences demonstrated small effects and explained only a small portion of the variance in holistic scores of writing quality. To better illuminate these findings, we present examples of shorter and longer clauses found in in the TOEFL writing data across low and high-scoring essays in Table 17.

Furthermore, as previously discussed, results from large-grained indices such as MLC are difficult to interpret with much precision. Such results suggest that learners write longer clauses, but leave gaps in our understanding regarding the specific structures that contribute to increased clause length (e.g., prepositional phrases as adverbials, prepositional phrases as noun modifiers, infinitive phrases, etc.). Furthermore, it is unclear whether the inclusion of particular clause-lengthening structures is uniform across participants and/or score levels (Larsen–Freeman, 2009; Norris & Ortega, 2009).

TABLE 18
Examples of Infinitive Clauses in High-Scoring Essays

Grammatical Construction	Example
infinitive – clausal complement	*It is our responsibility* ***[to make]***$_{infinitive\ verb}$ ***[our children understand …]***$_{clausal\ complement}$
infinitive – direct object	*The issue of deciding to choose and* ***start***$_{infinitive\ verb}$ ***[a career]***$_{direct\ object}$
adverb modifier – infinitive – direct object	*Parents can contribute by signaling teachers about* ***how***$_{adverb\ modifier}$ ***to teach***$_{infinitive\ verb}$ ***their children***$_{direct\ object}$

FIGURE 1
Phrasal Complexity: Prepositions per Object of the Preposition



Fine-Grained Clausal Complexity

The relationship between fine-grained indices of clausal complexity and TOEFL independent essay scores was also significant but small. A single index, nominal subjects per clause, met the inclusion criteria. A linear regression using this variable explained 3.0% of the variance in essay scores. The negative correlation between the number of nominal subjects per clause and holistic essay scores suggests that higher scoring essays tend to include more nonfinite clauses, such as infinitive clauses. This result, while weak, provides some support for Biber et al.'s (2011) hypothesis that writers will move from writing characterized by finite dependent clauses to writing characterized by nonfinite dependent clauses (such as infinitive clauses). To better demonstrate this, we provide examples of infinitive clause in high-scoring TOEFL essays in Table 18.

Fine-Grained Phrasal Complexity

The relationship between fine-grained indices of phrasal complexity and holistic essay quality scores was significant and demonstrated a medium effect size. Twelve indices related to phrasal complexity met the selection criteria and were entered into a stepwise regression. The resulting model included six indices related to prepositional object modifiers, direct object modifiers, and nominal subject modifiers, and explained approximately 19% of the variance in essay scores. These results provide support for the importance of phrasal complexity in indexing L2 writing proficiency (as measured by writing quality scores), and bolster claims made by Biber et al. (2011). The results also provide further support for the use of fine-grained indices that indicate the use of particular linguistic structures over more linguistically opaque indices such as MLC (e.g., Norris & Ortega, 2009). However, the analysis also indicated a potential weakness in fine-grained indices: their rareness. For instance, a number of the structures examined in TAASSC were rare in TOEFL independent essays, leading to nonnormal distributions and the exclusion of their related indices. Indirect objects, for example, were extremely rare in the data, as were passive constructions. Those indices that were frequent in the data and were included in the predictor model are discussed in the following sections.

Prepositional Object Dependents. Indices related to prepositional object modifiers accounted for 13.6% of the variance in holistic essay scores. High-scoring essays tended to include objects of the preposition with more dependents. Specifically, high-scoring indices tended to include prepositional objects with more adjective modifiers (see Figure 1) and prepositional phrase modifiers (see Figure 2). These results generally align with Biber et al.'s hypothesized developmental stages, wherein the writing of more proficient L2 writers is characterized by the use of noun phrase modifiers.

Direct Object Dependents. Indices related to direct object dependents accounted for 3.5% of the variance in holistic essay scores. High-scoring

FIGURE 2
Phrasal Complexity: Adjectives per Object of the Preposition

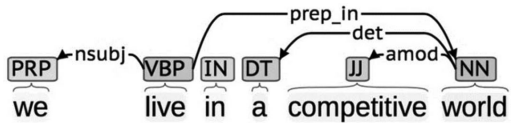
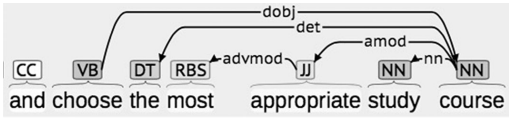


FIGURE 3
Phrasal Complexity: Dependents per Direct Object



essays tended to include direct objects with more dependents and a wider range of dependents. See Figure 3 for an example of direct object dependents in TOEFL essays. These results provide further support for the importance of noun phrase modifiers as an indicator of L2 academic writing proficiency (i.e., Biber et al., 2011).

Nominal Subject Dependents. An index related to nominal subject dependents accounted for 2.0% of the variance in holistic essay scores. High-scoring essays tended to include nominal subjects with a wider range of dependents. See Figure 4 for an example of nominal subject dependents in TOEFL essays.

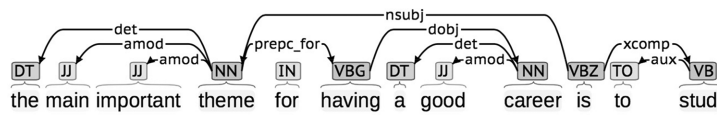
Combined Model

The combined complexity model explained 20.3% of the variance in holistic essay scores and was nearly identical to the phrasal complexity model. The only divergence was the addition of the fine-grained clausal complexity index, *nominal subjects per clause*, which contributed 1.4% of the explained variance. Neither MLC nor CP/C were included in the model. From both an interpretation and explanatory power perspective, these results highlight the advantage of using fine-grained indices and, in particular, the use of indices related to phrasal complexity. Our initial results using large-grained indices only suggested that more proficient L2 writers (as measured by a TOEFL independent essay holistic score) tended to write more elaborated clauses, and that the type of elaboration is not important. The fine-grained results supplement the finding regarding MLC by providing more detailed information about the types of elaboration that more proficient L2 writers tend to use. For example, more proficient L2 writers tended to include more elaborated direct objects and objects of the preposition, and a wider range of elaboration across direct objects and subjects. In particular, higher quality essays tend to include more preposition objects that are modified by adjectives and prepositional phrases. The fine-grained clausal complexity index also indicates that the relationship between clause length and writing quality scores can also be attributed to the increased use of nonfinite clauses. Furthermore, the model consisting of fine-grained indices of clausal and phrasal complexity was significantly better ($z = -3.71$, $p < .001$) than the MLC model in explaining the variance in writing quality scores. These results, in combination with the release of freely available text analysis tools, provides further rationale for the use of fine-grained indices of phrasal complexity in particular, and also fine-grained indices of clausal complexity. Additionally, in this article, the use of fine-grained indices provides clearer insights into the construct of writing quality than large-grained indices.

LIMITATIONS AND FUTURE DIRECTIONS

This study provides evidence for measuring L2 writing development using indices related to phrasal complexity. However, the study does have limitations. First, our results are based on a very narrow type of L2 writing (independent TOEFL essays written in response to two prompts), and caution should be used when generalizing these results to other writing contexts. Future research should investigate whether the trends observed in the context of TOEFL independent essays are also observed in other writing contexts. Second, this study explored whether linear relationships exist between linguistic indices and writing

FIGURE 4
Phrasal Complexity: Dependents per Nominal Subject



quality scores. While the investigation of linear trends is a logical first step in such an analysis, nonlinear trends should also be explored. Future research should investigate nonlinear trends both cross-sectionally (e.g., Lu, 2011) and longitudinally (e.g., Verspoor, Schmid, & Xu, 2012). Third, this study investigated one aspect of writing development (syntactic complexity). The relationship between syntactic complexity and other important linguistic features of writing quality such as lexical and phrasal sophistication (e.g., Engber, 1995; Guo et al., 2013; Kyle & Crossley, 2016; Laufer & Nation, 1995) should also be explored to determine the relative influence of different linguistic features on writing quality scores. Finally, Stanford neural-network dependency parser, which TAASSC relies on to identify syntactic structures, achieves a state of the art tagging accuracy of around 90% (Chen & Manning, 2014). While parsers such as the Stanford constituency parser (Klein & Manning, 2003) have been found to accurately identify linguistic structures in L2 writing (Lu, 2010), the accuracy of the Stanford neural-network parser has not been formally evaluated with L2 texts. This is an area for future research.

CONCLUSION

This study investigated the predictive validity of three types of syntactic complexity indices related to clausal and phrasal complexity. The results indicated that fine-grained indices of phrasal complexity (e.g., number of dependents per prepositional object) were stronger predictors of writing quality than either traditional syntactic complexity indices (e.g., MLC) or fine-grained clausal complexity indices (e.g., number of subjects per clause). The combined analysis also indicated that the most accurate models will likely include fine-grained indices of both phrasal and clausal complexity. These results extend previous corpus-based findings regarding the importance of phrasal elaboration in academic writing to the measurement of L2 writing quality (i.e., Biber et al., 2011). Furthermore, this study presents an automated method for the investigation of syntactic complexity using fine-grained, interpretable indices (Norris & Ortega, 2009). It is hoped that this study will serve as a starting point for the principled replication of previous L2 writing studies and for the development of new studies that consider the role of syntactic constructions in explaining L2 writing quality.

NOTES

[1] The Biber Tagger (Biber, 1988; Biber et al., 2004), which was developed by Doug Biber, pre-dates the advancements noted in this paragraph.

[2] As is often the case in fine-grained linguistic analyses, most of the nonnormally distributed indices represented features that occurred rarely in the learner data (e.g., indirect objects) resulting in zero counts that strongly skewed the data.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Bardovi–Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly, 26*, 390–395.

Bardovi–Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition, 11*, 17–34.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., …, Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus*. Princeton, NJ: Educational Testing Service.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*, 5–35.

Biber, D., Gray, B., & Staples, S. (2014). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics, 37*, 639–668.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & F. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). Philadelphia/Amsterdam: John Benjamins.

Charniak, E., & Johnson, M. (2005). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In J. Eisner & P. Koehn (Eds.), *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 173–180). New Brunswick, NJ: Association for Computational Linguistics.

Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In Y. Marton (Ed.), *The 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 740–750). Stroudsburg, PA: Association for Computational Linguistics.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

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.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*, 243–263.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, *10*, 5–43.

Dunn, O. J., & Clark, V. (1969). Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association*, *64*, 366–377.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*, 139–155.

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, & Computers*, *36*, 193–202.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*, 218–238.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, *11*, 10–18.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels*. NCTE Research Report No. 3. Champaign, IL: National Council of Teachers of English.

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In D. Yarowsky & S. Kurohashi (Eds.), *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* (pp. 423–430). New Brunswick, NJ: Association for Computational Linguistics.

Knoch, U., Rouhshad, A., & Storch, N. (2014). Does the writing of undergraduate ESL students develop after one year of study in an English-medium university? *Assessing Writing*, *21*, 1–17.

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* (Unpublished doctoral dissertation). Georgia State University, Atlanta, GA.

Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12–24.

Larsen–Freeman, D. (1978). An ESL index of development. *TESOL Quarterly*, *12*, 439–448.

Larsen–Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*, 579–589.

Larsen–Freeman, D., & Strom, V. (1977). The construction of a second language acquisition index of development. *Language Learning*, *27*, 123–134.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307–322.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*, 474–496.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*, 36–62.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*, 555–578.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*, 492–518.

Parkinson, J., & Musgrave, J. (2014). Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purposes*, *14*, 48–59.

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics*. Harlow, UK: Pearson Education.

Taguchi, N., Crawford, W., & Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly*, *47*, 420–430.

Verspoor, M. H., Schmid, M. S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, *21*, 239–263.

Wolfe–Quintero, K., Inagaki, S., & Kim, H.–Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu, HI: University of Hawai'i Press.