A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers' Language Development

XIAOFEI LU
The Pennsylvania State University
University Park, Pennsylvania, United States

This article reports results of a corpus-based evaluation of 14 syntactic complexity measures as objective indices of college-level English as a second language (ESL) writers' language development. I analyzed large-scale ESL writing data from the Written English Corpus of Chinese Learners (Wen, Wang, & Liang, 2005) using a computational system designed to automate syntactic complexity measurement with 14 measures that have been proposed in second language writing development studies (Lu, 2010). This analysis allows us to investigate the impact of sampling condition on the relationship between syntactic complexity and language development, to identify measures that significantly differentiate between developmental levels, to determine the magnitude at which between-level differences in each measure reach statistical significance, to assess the pattern of development associated with each measure, and to examine the strength of the relationship between different pairs of measures. This research provides ESL teachers and researchers with useful insights into how these measures can be used effectively as indices of college-level ESL writers' language development.

doi: 10.5054/tq.2011.240859

Syntactic complexity is evident in second language (L2) writing in terms of syntactic variation and sophistication, or, more specifically, the range of syntactic structures that are produced and the degree of sophistication of such structures. Syntactic complexity has been recognized as an important construct in L2 writing teaching and research, as the growth of a learner's syntactic repertoire is an integral part of his or her development in the target language (Ortega, 2003). A large variety of measures have been proposed for characterizing syntactic complexity in L2 writing. These measures typically seek to quantify one or more of the following: length of production unit, amount of subordination or coordination, range of syntactic structures, and degree of sophistication of certain syntactic structures. Together with measures of fluency and accuracy, these complexity measures have been explored in numerous L2 writing development studies with the aim to find valid and reliable developmental indices by which L2 teachers and researchers can expediently determine and describe a learner's developmental level or global proficiency in the target language (Larsen-Freeman, 1978; Ortega, 2003; Wolfe-Quintero, Inagaki, & Kim, 1998).

A major challenge that has been facing researchers who apply syntactic complexity measures to large language samples is the lack of computational tools to automate syntactic complexity analysis and the labor-intensiveness of manual analysis. Consequently, most previous studies either focused on few measures[1] or analyzed relatively small amounts of data. For example, Ortega (2003) reviewed 25 L2 writing studies in a research synthesis, all of which examined some of the following six syntactic complexity measures: mean length of sentence, mean length of T-unit, mean length of clause, T-units per sentence, clauses per T-unit, and dependent clauses per clause. Among these, only four studies examined four or more measures. The number of samples analyzed among the 21 cross-sectional studies she reviewed ranged from 16 to 300 (mean = 84, standard deviation = 74), and the length of those samples ranged from 70 to 500 words (mean = 234, standard deviation = 110). The scenario has not changed much in more recent research. For example, Stockwell and Harrington (2003) applied one syntactic complexity measure, mean length of T-unit, to approximately 300 email messages; Ellis and Yuan (2004) applied one measure, clauses per T-unit, to 52 narratives; and Beers and Nagy (2009) applied two measures, mean length of clause and clauses per T-unit, to 41 essays.

In the search for the most reliable syntactic complexity measures as indices of language development in L2 writing, however, it is desirable to directly compare the full range of measures of interest within a single study using large-scale learner data. As Wolfe-Quintero et al. (1998) indicated, the choice and definition of measures among previous developmental index studies are often inconsistent, and the results reported on the same measures are often mixed. For example, 18 studies they reviewed examined the relationship between proficiency and clauses per T-unit, among which 7 reported a significant relationship, but 11 did not. This poses a problem in the cumulative state of knowledge offered therein. Ortega (2003) also cautioned that, in pooling previous results to compare the performance of different measures as developmental indices, the research synthesis approach

[1] This refers specifically to the number of syntactic complexity measures applied. It is common for studies to examine measures of syntactic complexity along with measures of other constructs, such as accuracy and fluency.

suffers from unidentified sources of error introduced by the variability among previous studies in the writing task used, sample size, corpus length, etc. Consequently, it is not always straightforward for L2 teachers and researchers to decide on the best measures to use based on the findings reported thus far.

The present study sets out to remedy this situation. I directly compared a comprehensive set of 14 syntactic complexity measures commonly used in L2 writing development research by analyzing large-scale college-level ESL writing data from the Written English Corpus of Chinese Learners (WECCL; Wen, Wang, & Liang, 2005) using a computational system designed to automate syntactic complexity measurement (Lu, 2010). Specifically, I aimed to investigate the impact of sampling condition on the relationship between syntactic complexity and language development, to identify measures that significantly differentiate between developmental levels, to determine the magnitude at which between-level differences in each measure reach statistical significance, to assess the pattern of development associated with each measure, and to examine the strength of the relationship between different pairs of measures. The research design allowed elimination of the inconsistency and variability discussed earlier and offered ESL writing teachers and researchers reliable, new insights into how these measures compare with and relate to each other as indices of ESL writers' language development.

One should note that valid developmental measures cannot be assumed to be valid measures of writing proficiency or quality (e.g., Perkins, 1983). The latter are generally based on rating scales such as the American Council on the Teaching of Foreign Languages writing proficiency guidelines (Breiner-Sanders, Swender, & Terry, 2001), which consider the writer's demonstrated ability to control a multiplicity of aspects (e.g., vocabulary, grammar, syntax, and organization) to effectively direct writing to specific audiences in ways that are appropriate for the requirements of the discourse and the target language culture. Pursuant to the goals of developmental index studies, I evaluated syntactic complexity measures as objective indices of "language development as it is manifested in a written modality" (Wolfe-Quintero et al., 1998, p. 2), not as indicators of writing proficiency or quality. Needless to say, a full picture of language development in L2 writing can only be obtained by engaging fluency, accuracy, and complexity measures at various linguistic levels, including vocabulary, morphology, syntax, semantics, pragmatics, and discourse. I hope that in-depth knowledge gained about syntactic complexity measures will enhance future investigations into the interactions between different types of developmental measures at various linguistic levels.

SYNTACTIC COMPLEXITY IN L2 WRITING DEVELOPMENT

Developmental index studies began in the late 1970s with the goal to identify a developmental yardstick for gauging global L2 proficiency (Larsen-Freeman, 1983). As Wolfe-Quintero et al. (1998) noted, such developmental indices would facilitate a more precise characterization of a learner's developmental level as well as a more objective assessment of the effect of a specific pedagogical treatment on language use. A fundamental issue that needs to be addressed to achieve this goal is the extent to which developmental measures of fluency, accuracy, and complexity that have been proposed are valid and reliable indices of a learner's developmental level or global proficiency in the target language.

With respect to syntactic complexity measures, this issue has been tackled primarily in cross-sectional studies that investigated how well such measures discriminate independently determined proficiency levels (Bardovi-Harlig & Bofman, 1989; Ferris, 1994; Henry, 1996; Homburg, 1984; Larsen-Freeman, 1978; Ortega, 2003). Some longitudinal studies also tracked learners' language development as indexed by changes in syntactic complexity of L2 writing over time (Casanave, 1994; Hunt, 1970; Ishikawa, 1995; Ortega, 2000, 2003; Stockwell & Harrington, 2003). Although these studies share a common goal, they differ from one another in several dimensions. First, the measures examined varied from study to study. Although the number of measures examined in any single study is typically small, the total number of measures that have been proposed is fairly large. Wolfe-Quintero et al. (1998) identified over 30 syntactic complexity measures proposed in previous L2 writing development studies. In general, most measures consider clauses, sentences, or T-units as production units and analyze them in terms of length (e.g., mean length of T-unit) or in relation to either one another (e.g., clauses per T-unit) or particular syntactic structures (e.g., complex nominals per T-unit). Second, the production units and syntactic structures involved in calculating the measures were sometimes inconsistently defined. For example, most researchers considered clauses as structures with a subject and a finite verb (e.g., Hunt, 1965; Polio, 1997), but some also counted nonfinite verb phrases as clauses (e.g., Bardovi-Harlig & Bofman, 1989). Furthermore, many studies failed to provide definitions for the relevant structures or interrater reliability correlations for structure identification, making it difficult to replicate them or to assess the reliability of the reported results. Among the 39 studies Wolfe-Quintero et al. (1998) reviewed, only 7 reported interrater reliability. Third, proficiency level was variably conceptualized in ways that were not always directly comparable, including program level (e.g., Larsen-Freeman, 1978), school level (e.g., Yau, 1991), rating scales (e.g., Henry, 1996), and short-term changes in classes (e.g., Ishikawa, 1995). Given that developmental index studies compare syntactic complexity measures to language proficiency measures to determine how well the former index developmental levels, a relevant question that arises is how well these conceptualizations of proficiency level reflect developmental level. Wolfe-Quintero et al. (1998) argued that "program level may be the most valid developmentally" (p. 9) and that some developmental measures may not discriminate among holistic ratings of writing samples from intact classes, because such samples might be developmentally similar. Finally, the size and type of the writing samples analyzed varied across studies as well. Because of the labor intensiveness of manual analysis, the size of the samples analyzed tended to be small. The tasks used for sample elicitation and the genres of the samples varied considerably. In terms of the learner's first language (L1) background, both homogeneous (or mostly homogeneous; e.g., Ishikawa, 1995; Stockwell & Harrington, 2003) and heterogeneous (e.g., Bardovi-Harlig & Bofman, 1989; Ferris, 1994) groups have been used. However, studies that used heterogeneous L1 groups tended to treat all learners as one single group, without considering the potential effect of their L1 background on syntactic complexity. Given the variability in research design, it is unsurprising that studies often reported inconsistent results on specific measures. As mentioned earlier, 18 studies Wolfe-Quintero et al. (1998) reviewed examined clauses per T-unit, among which 7 reported a significant relationship to proficiency, but 11 did not. This makes it challenging to interpret and utilize the cumulative knowledge presented about particular measures and more so to pool knowledge about different measures to evaluate how they compare with and relate to each other as developmental indices (Ishikawa, 1995; Ortega, 2003; Wolfe-Quintero et al., 1998).

The effects of different learner-, task-, and context-related factors on the relationship of syntactic complexity to language proficiency have also been extensively studied. Sotillo (2000) examined how different modes of computer-mediated communication affect syntactic complexity in advanced ESL writers' output and reported that the delayed nature of asynchronous discussions offers more opportunities to produce syntactically complex language. Way, Joiner, and Seaman (2000) investigated the effects of different writing tasks and prompts on samples written by beginning learners of French, measuring syntactic complexity using mean length of T-unit. They suggested that syntactic complexity was highest for the descriptive task and lowest for the expository task. Ortega (2003) examined the impact of instructional setting and proficiency sampling criterion on the relationship between proficiency and syntactic complexity. She found that ESL learners produced writing of higher syntactic complexity than EFL learners and that studies using holistic rating as the proficiency sampling criterion yielded narrower ranges of complexity values than those using program level. Ellis and Yuan (2004) studied how planning conditions affect Chinese learners' written narratives and reported that the lack of planning negatively affects syntactic complexity. Beers and Nagy (2009) examined how genre affects the relationship of syntactic complexity measures to rated quality of writing samples produced by middle school students. They showed that words per clause correlated positively with quality for expository essays, and clauses per T-unit correlated positively with quality for narratives. These studies pinpoint the importance of controlling for relevant factors in establishing the relationship of syntactic complexity measures to language proficiency.

Several studies examined the role syntactic complexity plays in L2 writing instruction or assessment. Buckingham (1979) argued that one goal for advanced composition instruction is to adjust the focus of vocabulary and syntax teaching so as to "produce in advanced student writers an increase in the clarity, complexity, and specificity of the linguistic units selected for communication" (p. 249). Perkins (1983) discussed the assumptions, procedures, and consequences for use of several syntactic complexity metrics as objective measures of students' ability to write. Silva (1993) reviewed 72 comparative studies of L1 and L2 writing and summarized salient differences between the two pertaining to composing processes and features of written texts, including fluency, accuracy, and morphosyntactic structure, and discussed implications of these findings for the practical concerns of assessment and instruction. Hinkel (2003) analyzed 1,083 L1 and L2 academic texts, concluded that advanced nonnative-English speaking students in U.S. universities overuse simple syntactic constructions, and proposed instructional methods for addressing this shortfall. Although not all these studies engaged the same syntactic complexity measures used in L2 writing development research, they suggest that research on developmental measures of syntactic complexity has useful applications in L2 writing instruction and assessment.

SYNTACTIC COMPLEXITY MEASURES INVESTIGATED

Measure Selection and Definition

Over 100 developmental measures of accuracy, fluency, and lexical and syntactic complexity employed in 39 L2 writing development studies were reviewed by Wolfe-Quintero et al. (1998) in a large-scale research synthesis. They identified measures that performed the best based on the cumulative evidence presented, and recommended some new measures for further research. Six syntactic complexity measures they reviewed were investigated in greater depth in a more focused research synthesis by Ortega (2003), who compared the results reported for each measure among 25 college-level L2 writing studies. The set of measures discussed in these two research syntheses represent a fairly complete picture of the range of measures adopted in L2 writing research. Fourteen measures are selected from this set and evaluated here. These include the six measures Ortega (2003) examined, a further five reviewed in Wolfe-Quintero et al. (1998) that were shown by at least one previous study to exhibit at least weak correlation with or effect for proficiency, and three new measures that Wolfe-Quintero et al. (1998) recommended for further research. These measures are categorized into five types, as summarized in Table 1 and described later. For each measure, Table 1 also lists the number of previous studies that reported varying degrees of correlation with or effect for proficiency, as tallied in Wolfe-Quintero et al. (1998).

TABLE 1
Syntactic Complexity Measures Evaluated

Measure	Code	***	**	*	X
Type 1: Length of production					
Mean length of clause	MLC		5	1	3
Mean length of sentence	MLS		5		5
Mean length of T-unit	MLT	4	19	5	12
Type 2: Sentence complexity					
Clauses per sentence	C/S		1		1
Type 3: Subordination					
Clauses per T-unit	C/T	1	6	4	7
Complex T-units per T-unit	CT/T			1	
Dependent clauses per clause	DC/C	1		1	1
Dependent clauses per T-unit	DC/T		1		2
Type 4: Coordination					
Coordinate phrases per clause	CP/C				
Coordinate phrases per T-unit	CP/T			1	
T-units per sentence	T/S		1		4
Type 5: Particular structures					
Complex nominals per clause	CN/C				
Complex nominals per T-unit	CN/T		1		
Verb phrases per T-unit	VP/T				

Note. MLC = mean length of clause; MLS = mean length of sentence; MLT = mean length of T-unit; C = clause; S = sentence; T = T-unit; CT = complex T-unit; DC = dependent clause; CP = coordinate phrase; CN = complex nominals; VP = verb phrases; X = Measures that show no correlation with or effect for proficiency.
*** Measures that highly correlate with proficiency ($r \geqslant 0.65$) or show an overall effect for proficiency with a significant difference between three or more adjacent proficiency levels ($p < 0.05$).
** Measures that moderately correlate with proficiency ($0.45 \leq r < 0.65$), or show an overall effect for proficiency for two or more proficiency levels ($p < 0.005$).
* Measures that weakly correlate with proficiency ($0.25 \leq r < 0.045$) or show a trend toward an effect for proficiency ($p < 0.10$).

Length of Production

The first type includes three measures that gauge length of production at the clausal, sentential, or T-unit level, including

1. mean length of clause (MLC): number of words divided by number of clauses;
2. mean length of sentence (MLS): number of words divided by number of sentences; and
3. mean length of T-unit (MLT): number of words divided by number of T-units.

Sentence Complexity

The second type comprises a sentence complexity ratio, i.e.,

4. clauses per sentence (C/S): number of clauses divided by number of sentences.

Subordination

The third type contains four ratios that reflect the amount of subordination, including

5. clauses per T-unit (C/T): number of clauses divided by number of T-units;
6. complex T-units per T-unit (CT/T): number of complex T-units divided by number of T-units;
7. dependent clauses per clause (DC/C): number of dependent clauses divided by number of clauses; and
8. dependent clauses per T-unit (DC/T): number of dependent clauses divided by number of T-units.

Coordination

The fourth type includes three ratios that measure the amount of coordination, namely,

9. coordinate phrases per clause (CP/C): number of coordinate phrases divided by number of clauses;
10. coordinate phrases per T-unit (CP/T): number of coordinate phrases divided by number of T-units; and
11. T-units per sentence (T/S): number of T-units divided by number of sentences.

Particular Structures

The final type comprises three ratios that consider particular structures in relation to larger production units, including

12. complex nominals per clause (CN/C): number of complex nominals divided by number of clauses;
13. complex nominals per T-unit (CN/T): number of complex nominals divided by number of T-units; and
14. verb phrases per T-unit (VP/T): number of verb phrases divided by number of T-units.

Definitions of Relevant Production Units and Structures

The definitions of the measures described earlier entail explicit definitions of the production units and structures involved. In cases of competing definitions, the most widely used is selected.

Sentence, Clause, and Dependent Clause

The definition of sentence is the least problematic. A sentence is a group of words punctuated with a sentence-final punctuation mark, usually a period, exclamation mark or question mark, and in some cases elliptical marks or closing quotation marks. Sentence fragments punctuated as complete sentences are counted as sentences, too (Hunt, 1965). Two approaches to counting clauses exist. Most studies considered clauses as structures with a subject and a finite verb, including independent, adjective, adverbial, and nominal clauses, but not nonfinite (including gerund, infinitive, and participle) verb phrases (Hunt, 1965; Polio, 1997). Some studies also counted nonfinite verb phrases as clauses (Bardovi-Harlig & Bofman, 1989). I do not consider nonfinite verb phrases as clauses but count them as verb phrases. A dependent clause is then a finite adverbial, adjective, or nominal clause, as is the case in most previous studies that analyzed dependent clauses (Cooper, 1976; Hunt, 1965; Kameen, 1979).

T-Unit and Complex T-Unit

Hunt (1970) defined T-unit as “one main clause plus any subordinate clause or nonclausal structure that is attached to or embedded in it” (p. 4). This definition has been consistently followed in L2 writing studies. A complex T-unit is then one containing at least one dependent clause (or subordinate clause in Hunt’s term; Casanave, 1994).

Coordinate Phrase, Complex Nominal, and Verb Phrase

Coordinate phrases include coordinate adjective, adverb, noun, and verb phrases. Complex nominals include (1) nouns plus adjective, possessive, prepositional phrase, adjective clause, participle, or appositive; (2) nominal clauses; and (3) gerunds and infinitives in subject, but not object position (Cooper, 1976). Verb phrases include both finite and nonfinite verb phrases.

METHOD

Data

I evaluated the 14 syntactic complexity measures using large-scale college-level ESL data from the WECCL. This corpus consists of 3,678 essays written by English majors aged 18–22 years from nine Chinese colleges. Each essay is annotated with a header that includes the following information: mode, genre, school level, year of admission, timing condition, institution, and length. Sixteen topics were used across the corpus. The prompts were generally brief, and prompts for the same genre followed a similar pattern. For example, prompts for argumentative essays presented either one view or two opposing views on an issue and asked the students to state their own opinions, e.g., “Some people think that education is a lifelong process, while others don’t agree. Write an essay to state your own opinion” (Wen et al., 2005, p. 111). Students in the same school level within the same institution wrote on the same topic, but topics varied from institution to institution.

With a script written to verify the integrity of the data, it was found that 124 of the 3,678 files were unusable, including 1 with no header, 1 with two headers, 4 with one sentence, 17 empty files, and 101 duplicates. This left 3,554 valid files. The corpus has a total of 1,119,510 words, and the essays range from 89 to 892 words in length (mean = 315, standard deviation = 87). Table 2 (adapted from Table 2, Lu, 2010) summarizes the distribution of the essays in terms of school level, genre, and timing condition. The number of essays from each institution ranges from 82 to 1031 (mean = 395, standard deviation = 266). However, only the institution coded ND is represented in all nonzero cells in Table 2. All expository essays were from this institution, as were all essays by fourth-year students.

Research Questions

Given the information available in the corpus, I conceptualized proficiency level using school level. Following Wolfe-Quintero et al. (1998), I assumed that if a measure progresses linearly in a way that is significantly related to school level, it is potentially a good candidate for a developmental index. With this conceptualization and assumption, I analyzed the syntactic complexity of the essays in the corpus using the 14 measures, with the aim to answer the following four research questions.

TABLE 2
Essay Distribution in the WECCL

	Argumentative		Narrative		Expository		
School Level	Timed	Untimed	Timed	Untimed	Timed	Untimed	Total
1	695	395	89	0	30	0	1,209
2	441	398	246	0	28	0	1,113
3	504	459	91	0	30	0	1,084
4	60	0	88	0	0	0	148
Total	1,700	1,252	514	0	88	0	3,554

Note. WECCL = Written English Corpus of Chinese Learners. Table adapted from Table 2, in Lu (2010), with kind permission from John Benjamins, Amsterdam/Philadelphia. www.benjamins.com.

1. What is the impact of sampling condition, including institution, genre, and timing condition, on the mean values of any given syntactic complexity measure?
2. Which measures show significant between-proficiency differences? What is the magnitude at which between-proficiency differences in each measure reach statistical significance?
3. What are the patterns of development for the measures that show significant between-proficiency differences?
4. What is the strength of the relationship between different pairs of syntactic complexity measures?

A note on the validity of the conceptualization of proficiency level using school level in this study is in order. With the goal to evaluate syntactic complexity measures as objective indices of ESL writers' language development, a conceptualization is needed that is valid developmentally, but not necessarily indicative of writing proficiency or quality. As discussed earlier, among the various conceptualizations of proficiency level, Wolfe-Quintero et al. (1998) considered program level to be the most valid developmentally. In the context of English major programs in Chinese universities, school level functions in essentially the same way as program level does in other ESL contexts for two reasons. First, students admitted into the same English major program can be expected to be at about the same proficiency level, because they have been exposed to the same national English curriculum in secondary school and have performed comparably on the National College Entrance Examination, including its spoken English test component. Second, the curriculum of English major programs in Chinese universities follows the national syllabus for English majors, and within the same program, students must pass the same set of required English courses to advance to the next level.

Analysis

The essays in the corpus are analyzed using a computational system designed to automate the measurement of syntactic complexity of college-level ESL writing samples (Lu, 2010). Given a sample as input, the system first utilizes the Stanford parser (Klein & Manning, 2003) to analyze the syntactic structure of each sentence, then queries the parsed sample with a set of syntactic patterns to retrieve the occurrences of the relevant production units and structures, and finally computes the 14 syntactic complexity indices of the sample using the frequency counts of those units and structures. The system was evaluated on 20 samples randomly selected from the WECCL. The occurrences of the relevant production units and structures retrieved by the system were compared against those manually identified by two human annotators.[2] For production unit and structure identification, the system achieved *F*-scores ranging from 0.830 for complex nominals to 1.000 for sentences (see Table 3, adapted from Table 6; Lu, 2010). Correlations between the complexity scores computed by the system and the annotators were significant ($p < 0.01$), ranging from 0.834 for CP/C to 1.000 for MLS (see Table 4, adapted from Table 7; Lu, 2010). These results indicate that the production units and structures that the system identifies and the syntactic complexity indices it generates are highly reliable.

Given the large number of measures involved, a Bonferroni correction was employed to adjust the *p*-values for each set of statistical tests of significance. This preserves simultaneous 95% confidence for all tests in each set to avoid false positive conclusions because of repeated use of the same test. The reported *p*-values reflect these adjustments.

RESULTS AND DISCUSSION

Research Question 1

Institution

Because the data were collected from nine institutions, and students from different institutions wrote on different topics, it is necessary to examine whether significant differences in mean syntactic complexity

[2] Lu (2010) reported interannotator and system-annotator agreement using precision, recall, and *F*-score. Let *X* and *Y* denote the number of occurrences of a structure identified in two annotations, and let *Z* denote the number of identical occurrences of the structure in *X* and *Y*; precision = *Z/X*, recall = *Z/Y*, and *F*-score = (2 × precision × recall)/(precision + recall). Interannotator agreement ranged from 0.907 for complex nominals to 1.000 for sentences (*F*-score). Correlations between the syntactic complexity scores computed by the two annotators ranged from 0.912 for CT/T to 1.000 for MLS. Interannotator disagreements were resolved through discussion.

TABLE 3

System Performance on Production Unit and Structure Identification

	Occurrences identified			System–annotator agreement		
Structure	System	Annotators	Identical	Precision	Recall	*F*-score
S	357	357	357	1.000	1.000	1.000
C	545	558	530	0.972	0.950	0.961
DC	170	178	161	0.947	0.904	0.925
T	376	380	369	0.981	0.971	0.976
CT	129	136	126	0.977	0.926	0.951
CP	138	135	125	0.906	0.926	0.916
CN	660	572	511	0.774	0.893	0.830
VP	750	758	698	0.931	0.921	0.926

Note. Table adapted from Table 6, in Lu (2010), with kind permission from John Benjamins, Amsterdam/Philadelphia. www.benjamins.com.

values exist among students from different institutions. Without controlling for genre or timing condition, a one-way analysis of variance (ANOVA) shows significant differences ($p < 0.05$) in the mean values of 13 measures (all but DC/C) among students from different institutions. The timed argumentative essays written by students in the first three levels are used to control for genre and timing condition, because all expository and narrative essays are untimed and all essays by fourth-year students are from one institution. A one-way ANOVA shows significant differences ($p < 0.05$) in the mean values of nine measures (all but C/S, C/T, DC/T, T/S, and CT/T) among students from different institutions.

Genre

The effect of genre using argumentative and narrative essays was investigated, because expository essays all come from a single institution and include no essays by fourth-year students. Without controlling for timing condition and institution, an independent-samples *t*-test shows

TABLE 4

Correlations Between System-Computed and Annotator-Computed Complexity Scores

Measure	Correlation	Measure	Correlation
MLC	0.932	DC/T	0.941
MLS	1.000	CP/C	0.834
MLT	0.987	CP/T	0.871
C/S	0.928	T/S	0.919
C/T	0.961	CN/C	0.867
CT/T	0.892	CN/T	0.896
DC/C	0.840	VP/T	0.858

Note. Table adapted from Table 7 in Lu (2010), with kind permission from John Benjamins, Amsterdam/Philadelphia. www.benjamins.com.

significant differences ($p < 0.05$) in the mean values of 13 measures (all but T/S) between argumentative and narrative essays (Samples 1 and 6 in Table 5). Argumentative essays generally exhibit higher syntactic complexity than narrative essays. The subset of data that comprises timed argumentative and narrative essays is used to control for timing condition. An independent-samples *t*-test shows significant differences ($p < 0.05$) in the mean values of 13 measures (all but C/S) between timed argumentative and narrative essays (Samples 2 and 6 in Table 5). Finally, the subset of data that includes timed argumentative and narrative essays from the institution coded ND is used to control for both timing condition and institution, because all essays by fourth-year students are from this institution. An independent-samples *t*-test shows significant differences ($p < 0.05$) in the mean values of 12 measures (all but C/S and C/T) between timed argumentative and narrative essays from the institution coded ND (Samples 4 and 7 in Table 5).

Timing Condition

Because all narrative and expository essays are timed, argumentative essays were used to investigate the effect of timing condition. Without controlling for institution, an independent-samples *t*-test shows significant differences ($p < 0.05$) in the mean values of 10 measures (all but C/S, C/T, T/S, and CP/C) between timed and untimed argumentative essays (Samples 2 and 3 in Table 5). Untimed argumentative essays

TABLE 5
Mean Complexity Values by Genre and Timing Condition

Sample	1	2	3	4	5	6	7
Genre	Arg.	Arg.	Arg.	Arg.	Arg.	Nar.	Nar.
Timing	All	Timed	Untimed	Timed	Untimed	Timed	Timed
Institution	All	All	All	ND	ND	All	ND
Size	2,952	1,700	1,252	422	169	514	352
MLC	9.293	9.194	9.428	9.889	10.577	8.343	8.578
MLS	15.417	15.057	15.906	15.916	17.463	13.283	13.716
MLT	13.910	13.663	14.245	14.546	15.981	11.752	12.161
C/S	1.677	1.658	1.703	1.627	1.663	1.611	1.610
C/T	1.510	1.500	1.524	1.483	1.521	1.423	1.429
CT/T	0.395	0.388	0.404	0.388	0.418	0.390	0.358
DC/C	0.332	0.327	0.338	0.330	0.352	0.304	0.305
DC/T	0.520	0.509	0.534	0.506	0.552	0.452	0.456
CP/C	0.250	0.245	0.257	0.261	0.334	0.178	0.190
CP/T	0.369	0.359	0.383	0.382	0.499	0.247	0.264
T/S	1.107	1.103	1.114	1.098	1.095	1.129	1.125
CN/C	1.034	1.013	1.063	1.122	1.292	0.789	0.825
CN/T	1.551	1.509	1.608	1.655	1.955	1.120	1.182
VP/T	2.057	2.031	2.093	2.083	2.173	1.881	1.921

Note. Arg. = argumentative; Nar. = narrative.

generally exhibit higher syntactic complexity than timed argumentative essays. Argumentative essays from the institution coded ND are used to control for institution, again because all essays by fourth-year students are from this institution. An independent-samples *t*-test shows significant differences ($p < 0.05$) in the mean values of seven measures (all but C/S, C/T, DC/C, DC/T, CT/T, T/S, and VP/T) between timed and untimed argumentative essays from the institution coded ND (Samples 4 and 5 in Table 5).

In summary, the results suggest that institution, genre, and timing condition have significant effects on the observed mean values of all or most measures. These results are consistent with previous findings on the effects of genre (Beers & Nagy, 2009; Way et al., 2000) and planning (Ellis & Yuan, 2004).

Research Question 2

Given the significant effects of institution, genre, and timing condition on the mean values of all or most of the measures, it is necessary to control for these variables in determining which measures significantly differentiate between school levels and what magnitude is required for between-proficiency differences in each measure to reach statistical significance. The subset of data that comprises all timed argumentative essays from the institution coded ND, the institution with the greatest number of essays and the only institution with essays by fourth-year students, was examined. Table 6 summarizes the mean complexity values and standard deviations (SD) for each school level in this subset. Table 7 summarizes the actual between-level differences found to be significant ($p < 0.05$) in a Bonferroni test, a one-way ANOVA post hoc multiple comparison test.

The three measures of length of production are shown to discriminate the first two adjacent levels as well as two or three pairs of nonadjacent levels. Statistical significance of between-level differences in these measures is reached at the magnitudes of 0.573, 1.658, and 1.650, respectively. These results are generally consistent with the empirical support for the length of production measures as an index of L2 proficiency, as summarized in Wolfe-Quintero et al. (1998). However, only MLT was previously reported to discriminate adjacent school levels (e.g., Hirano, 1991; Larsen-Freeman, 1983), whereas MLC and MLS were reported to discriminate nonadjacent school levels only (e.g., Cooper, 1976; Yau, 1991).

The sentence complexity measure, C/S, is shown to discriminate nonadjacent levels in a negative direction, with statistical significance of between-level differences reached at the magnitude of 0.118. Previous

TABLE 6
Mean Complexity Values of Timed Argumentative Essays From Institution ND

	1 (157)		2 (91)		3 (114)		4 (60)	
Level (Size)	Mean	SD	Mean	SD	Mean	SD	Mean	SD
MLC	8.958	1.362	9.995	1.609	10.568	1.839	10.874	1.681
MLS	14.717	2.847	16.375	3.378	16.861	3.407	16.568	2.595
MLT	13.276	2.467	14.926	2.804	15.561	3.185	15.368	2.543
C/S	1.657	0.308	1.651	0.298	1.612	0.298	1.539	0.236
C/T	1.492	0.242	1.504	0.230	1.483	0.240	1.426	0.219
CT/T	0.389	0.133	0.403	0.128	0.390	0.139	0.360	0.113
DC/C	0.334	0.092	0.348	0.098	0.325	0.091	0.301	0.082
DC/T	0.515	0.218	0.540	0.221	0.500	0.217	0.444	0.187
CP/C	0.226	0.119	0.266	0.125	0.292	0.149	0.292	0.158
CP/T	0.333	0.175	0.394	0.180	0.427	0.216	0.409	0.218
T/S	1.111	0.112	1.096	0.087	1.088	0.094	1.083	0.089
CN/C	0.941	0.291	1.140	0.299	1.273	0.363	1.281	0.324
CN/T	1.403	0.474	1.703	0.489	1.881	0.594	1.814	0.477
VP/T	2.043	0.377	2.088	0.336	2.130	0.346	2.090	0.275

studies have not examined its relationship to school level. In a longitudinal study, Ishikawa (1995) found a significant increase of C/S in one group over a three-month period. Because these results come from cross-sectional and longitudinal studies, respectively, further research is needed to interpret the discrepancy.

Among the four subordination measures, C/T and CT/T are not shown to discriminate school levels. DC/C and DC/T are shown to discriminate nonadjacent levels in a negative direction, with statistical significance of between-level differences reached at the magnitudes of 0.047 and 0.096, respectively. Despite the general progression in C/T

TABLE 7
Significant Between-Level Differences in Timed Argumentative Essays From Institution ND

Levels	1–2	2–3	3–4	1–3	2–4	1–4
MLC	1.037			1.610	0.879	1.916
MLS	1.658			2.144		1.851
MLT	1.650			2.285		2.092
C/S						−0.118
C/T						
CT/T						
DC/C					−0.047	
DC/T					−0.096	
CP/C				0.067		0.067
CP/T				0.094		
T/S						
CN/C	0.199	0.133		0.333	0.141	0.340
CN/T	0.300			0.478		0.411
VP/T						

relative to program or school level reported in several previous studies (Wolfe-Quintero et al., 1998), the results suggest otherwise and support the claim that C/T may be lower at advanced levels as a result of reduction from clauses to phrases (Ortega, 2003; Sharma, 1980). Previous studies on CT/T, DC/C, and DC/T are limited and focused primarily on their relationship to holistic ratings. Different from the present results, Hirano (1991) reported significant increases in CT/T between three program levels.

The two coordinate phrase measures, CP/C and CP/T, are shown to discriminate nonadjacent levels, with statistical significance of between-level differences reached at the magnitudes of 0.067 and 0.094, respectively. This contrasts with the result reported in Cooper (1976), who found only a trend toward a relationship between CP/T and school level. The results disconfirm the speculation by Wolfe-Quintero et al. (1998) that CP/C might perform better than CP/T. The third coordination measure, T/S, is not shown to discriminate school levels. This is consistent with the results reported in most previous studies, except for that of Monroe (1975), who found a significant relationship between T/S and school level.

The two complex nominal measures, CN/C and CN/T, show significant differences between three and two adjacent levels, respectively, with statistical significance of between-level differences reached at the magnitudes of 0.133 and 0.300, respectively. This is consistent with the moderate effect Cooper (1976) reported for CN/T. The results also confirm speculation by Wolfe-Quintero et al. (1998) that CN/C might outperform CN/T. VP/T is not found to discriminate school levels, disconfirming speculation by Wolfe-Quintero et al. that including both finite and nonfinite verb phrases in a VP/T measure may yield better results than the C/T measure, which includes finite verb phrases only.

Ortega (2003) proposed critical magnitudes for four measures for medium-sized samples with at least 10 participants per level: 4.5 for MLS, 2 for MLT, slightly over 1 for MLC, and at least 0.2 for C/T. She cautioned that these should be regarded as provisional, because of gaps in accumulated observations. Although C/T is not found to discriminate school levels, the magnitudes at which between-level differences in the first three measures reached statistical significance ($p < 0.05$) in the present study are much smaller. This supports Ortega's speculation that smaller magnitudes may be required for larger-size samples. As mentioned earlier, the samples used in the studies reviewed in Ortega (2003) are considerably smaller than samples in the present study. This may also explain why some measures that were not found to discriminate adjacent levels in previous studies are found to do so in the present study, such as MLC and MLS.

Research Question 3

The results reveal several patterns of development for the 10 measures that show significant between-level differences. These measures were grouped, first, according to whether the observed significant changes from lower to higher levels are positive or negative and, second, according to whether the observed pattern of development is linear or nonlinear across the four levels.

Seven measures show significant positive changes from lower to higher levels. These include the three measures of length of production: MLC, MLS, and MLT; the two coordinate phrase measures, CP/C and CP/T; and the two complex nominal measures, CN/C and CN/T. Among these, MLC and CN/C progress linearly across all four levels. CP/C progresses linearly across the first three levels and stays at the same level in Level 4. MLS, MLT, CP/T, and CN/T also increase linearly across the first three levels but show an insignificant decline from Level 3 to Level 4.

Three measures show significant negative changes from lower to higher levels, including the sentence complexity measure C/S and two subordination measures, DC/C and DC/T. C/S declines linearly across all four levels, with significant decreases found between levels one and four as well as between levels two and four, but not between adjacent levels. DC/C and DC/T show the same nonlinear pattern of development, i.e., they increase from Level 1 to Level 2 by an insignificant margin but then decline linearly from Level 2 to Level 4. Significant declines are not found between adjacent levels but are found between Levels 2 and 4 for both measures.

The nonlinear pattern observed for DC/C and DC/T supports the developmental prediction (Ortega, 2003; Sharma, 1980; Wolfe-Quintero et al., 1998), which argues for "non-linear complexification as far as subordination is concerned" (Ortega, 2003, p. 514), based on the expectation that advanced proficiency groups should capitalize on complexification at the phrasal, rather than the clausal level.

Research Question 4

Table 8 summarizes the correlations between the 14 measures. These reveal several interesting patterns of their relationships. First, measures generally correlate strongly with other measures of the same type or involving the same structure, as evidenced in the very high correlations[3] between MLS and MLT, C/T and DC/T, CN/C and CN/T, and CP/C

[3] Following Wolfe-Quintero et al. (1998), correlations are characterized as high ($r \geqslant 0.650$), moderate ($0.450 \leq r < 0.650$), and weak ($0.250 \leq r < 0.450$).

TABLE 8

Correlations Between Complexity Measures

	MLC	MLS	MLT	C/S	C/T	CT/T	DC/C	DC/T	CP/C	CP/T	T/S	CN/C	CN/T	VP/T
MLC	1	0.571**	0.657**	−0.330**	−0.281**	−0.231**	−0.194**	−0.228**	0.632**	0.547**	−0.160**	0.842**	0.640**	0.098*
MLS	0.571**	1	0.907**	0.572**	0.510**	0.438**	0.429**	0.496**	0.344**	0.501**	0.256**	0.561**	0.774**	0.641**
MLT	0.657**	0.907**	1	0.373**	0.526**	0.422**	0.451**	0.514**	0.401**	0.566**	−0.167**	0.654**	0.872**	0.681**
C/S	−0.330**	0.572**	0.373**	1	0.865**	0.739**	0.674**	0.794**	−0.219**	0.030	0.476**	−0.183**	0.243**	0.639**
C/T	−0.281**	0.510**	0.526**	0.865**	1	0.809**	0.789**	0.922**	−0.190**	0.103*	−0.022	−0.103*	0.386**	0.766**
CT/T	−0.231**	0.438**	0.422**	0.739**	0.809**	1	0.846**	0.865**	−0.124*	0.114*	0.051	−0.083	0.311**	0.652**
DC/C	−0.194**	0.429**	0.451**	0.674**	0.789**	0.846**	1	0.952**	−0.070	0.166**	−0.043	−0.088	0.306**	0.690**
DC/T	−0.228**	0.496**	0.514**	0.794**	0.922**	0.865**	0.952**	1	−0.118*	0.152**	−0.030	−0.087	0.367**	0.760**
CP/C	0.632**	0.344**	0.401**	−0.219**	−0.190**	−0.124*	−0.070	−0.118*	1	0.945**	−0.103*	0.390**	0.266**	0.059
CP/T	0.547**	0.501**	0.566**	0.030	0.103*	0.114*	0.166**	0.152**	0.945**	1	−0.116*	0.358**	0.387**	0.287**
T/S	−0.160**	0.256**	−0.167**	0.476**	−0.022	0.051	−0.043	−0.030	−0.103*	−0.116*	1	−0.181**	−0.186**	−0.062
CN/C	0.842**	0.561**	0.654**	−0.183**	−0.103*	−0.083	−0.088	−0.087	0.390**	0.358**	−0.181**	1	0.867**	0.105*
CN/T	0.640**	0.774**	0.872**	0.243**	0.386**	0.311**	0.306**	0.367**	0.266**	0.387**	−0.186**	0.867**	1	0.467**
VP/T	0.098*	0.641**	0.681**	0.639**	0.766**	0.652**	0.690**	0.760**	0.059	0.287**	−0.062	0.105*	0.467**	1

Note. ** Correlation is significant at the 0.01 level (2-tailed). * Correlation is significant at the 0.05 level (2-tailed).

and CP/T. Second, among the three measures of length of production, MLS and MLT are more strongly correlated with each other than with MLC. Both of them show weak to moderate positive correlations with C/S and the four subordination measures, whereas MLC exhibits very low to weak negative correlations with those measures. MLS and MLT also demonstrate much stronger correlations with VP/T than MLC. Nevertheless, all three show very low to weak correlations with T/S, moderate to high correlations with the complex nominal measures CN/C and CN/T, and weak to moderate correlations with the coordinate phrase measures CP/C and CP/T. Third, the sentence complexity measure C/S and the subordination measures display high correlations with each other. They all exhibit very low negative correlations with CP/C and CN/C, and very low to weak positive correlations with CP/T and CN/T. All but C/S show trivial correlations with T/S, but all show moderate to strong correlations with VP/T. Finally, the coordinate phrase and complex nominal measures are weakly correlated with each other. They generally show very low correlations with T/S and VP/T, except that CP/T and CN/T are weakly and moderately correlated with VP/T, respectively. T/S and VP/T are trivially correlated.

These correlations help to understand why certain measures exhibit similar patterns of development. For example, MLC is most strongly correlated with CN/C, and both measures not only progress linearly across all four levels but also discriminate adjacent school levels. Similarly, the high correlations between the subordination measures are in line with the similar pattern of development they exhibit. These correlations also help to determine which measures should be considered together in describing a learner’s proficiency. As Mendelsohn (1983) suggested, a low correlation between two measures suggests that both should be considered, because they capture different aspects of development. For example, it is advisable to consider MLC along with one of the subordination measures, given the low correlations between them and their ability to discriminate school levels.

CONCLUSIONS AND IMPLICATIONS

This corpus-based evaluation of the 14 syntactic complexity measures has allowed important questions to be answered on how the measures perform as objective indices of college-level ESL writers’ language development, how they relate to each other, and how their relationships to proficiency are affected by external factors. A major advantage of this research over previous studies is the investigation of a large set of measures using one large sample. This makes the results more reliable, because it avoids the inconsistency and variability found among previous studies in terms of choice and definition of measures, writing task used, sample size, and corpus length.

Summary of Findings and Implications for L2 Writing Development Research

This research yielded several substantive findings with important implications for L2 writing development research. First, I showed that institution, genre, and timing condition significantly affect the relationship between syntactic complexity and proficiency. The institution variable has not been considered in previous studies, which usually analyzed samples from a single institution. Consistent with previous findings on the effects of genre (Beers & Nagy, 2009; Way et al., 2000) and planning (Ellis & Yuan, 2004), argumentative essays showed higher syntactic complexity than narrative essays, and untimed essays showed higher syntactic complexity than timed essays. These results confirm the importance of controlling the effects of relevant learner-, task-, and context-related factors in interpreting between-proficiency differences in syntactic complexity. Methodologically, the results demonstrated how the computational system described in Lu (2010) can be used for this purpose, provided adequate information is encoded in the dataset.

Second, I identified good and poor candidates for developmental indices, based on the effects for proficiency and the patterns of development observed for the measures. It was assumed that good candidates are measures that progress linearly in a way that is significantly related to school level. Whereas all 14 measures were claimed or speculated to discriminate proficiency levels in previous studies (see, e.g., Ortega, 2003; Wolfe-Quintero et al., 1998), only 10 measures were shown to do so in this study, among which 7 progressed linearly from Level 1 to Level 3, with an insignificant change from Level 3 to Level 4, including 3 measures of length of production, 2 complex nominal measures, and 2 coordinate phrase measures. These seven measures are recommended to L2 writing researchers as good candidates for developmental indices.

The best candidates are complex nominals per clause (CN/C) and mean length of clause (MLC), both of which not only discriminated two or more adjacent levels but also increased linearly across all four levels. The next group includes complex nominals per T-unit (CN/T), mean length of sentence (MLS), and mean length of T-unit (MLT), all of which discriminated two adjacent levels and generally progressed relative to school level. The final group includes the two coordinate phrase measures, CP/C and CP/T, both of which discriminated nonadjacent levels, with significant increases from lower to higher levels. The present results on the three measures of length of production concur with previous findings that they constitute useful indices of L2 proficiency (Henry, 1996; Ishikawa, 1995; Larsen-Freeman, 1978; Ortega, 2003). However, different from previous findings that only MLT discriminates adjacent school levels (Cooper, 1976; Hirano, 1991; Larsen-Freeman, 1983; Yau, 1991), all three measures were found to discriminate adjacent school levels, with MLC demonstrating the strongest effect instead of MLT. The complex nominal and coordinate phrase measures have received little attention in previous studies. The only study that examined the relationship of CP/T and CN/T to school level was that of Cooper (1976), who reported insignificant and moderate effects for CP/T and CN/T, respectively. The present results indicated stronger discriminative power of the complex nominal and coordinate phrase measures than most other measures. This finding suggests that complexity at the phrasal level deserves closer attention in future research.

The other three measures that discriminated school levels, including clauses per sentence (C/S), dependent clauses per clause (DC/C), and dependent clauses per T-unit (DC/T), decreased significantly from lower to higher levels. Previous results on these measures were mixed, with moderate to strong effect reported for all of them by some researchers (Hirano, 1991; Homburg, 1984; Ishikawa, 1995) and insignificant effect reported by others (Ishikawa, 1995; Kameen, 1979). This discrepancy could be partially attributed to the lack of a consistent definition of dependent clause among previous studies. The present results on these measures concur with the linear progression observed for the complex nominal and coordinate phrase measures discussed earlier and support the developmental prediction that, as students advance to higher levels of proficiency, they learn to capitalize on complexification more at the phrasal level and less at the clausal level (Ortega, 2003; Sharma, 1980; Wolfe-Quintero et al., 1998).

The other four measures showed no significant differences between school levels, including clauses per T-unit (C/T), complex T-units per T-unit (CT/T), T-units per sentence (T/S), and verb phrases per T-unit (VP/T). Some previous studies reported significant effects for the first three of these (Casanave, 1994; Hirano, 1991; Monroe, 1975), and Wolfe-Quintero et al. (1998) speculated that the VP/T measure could be useful. These previous findings and speculations were not supported by the present results. I consider these four measures poor candidates for developmental indices, because they provide no useful information about a learner's proficiency level.

Third, I reported magnitudes at which between-level differences in each measure reach statistical significance. The thresholds observed were smaller than those reported in previous studies that used smaller samples (see, e.g., Ortega, 2003). These differences offer empirical support for Ortega's speculation that larger-size samples would likely yield smaller magnitudes. Methodologically, the sample size effect is a relevant issue that L2 writing development researchers should consider. The fact that the present study found effects for some of the measures, such as MLC and MLS, different from those reported in previous studies may also be partially attributed to the use of a larger sample.

Fourth, I discussed the patterns of relationship among the 14 measures. Previous studies have been unable to systematically examine relationships among as large a set of measures as evaluated in this study. The patterns observed provide useful guidance for selecting multiple measures that are good developmental indices and that are not highly correlated with each other, for example, MLC and one of the subordination measures, for use in future research.

Finally, the present results provide several pieces of evidence in support of the clause as a potentially more informative unit of analysis than the T-unit. Between MLC and MLT, MLC showed perfect linear progression across all four levels, whereas MLT progressed linearly across the first three levels only. Between the two complex nominal measures, CN/C discriminated the first three adjacent levels and progressed linearly across all four levels, whereas CN/T discriminated the first two adjacent levels and progressed linearly across the first three levels only. Similarly, between the two coordinate phrase measures, CP/C discriminated one more pair of nonadjacent levels than CP/T. In previous research, Gaies (1980) challenged the validity and usefulness of the T-unit analysis, and Bardovi-Harlig (1992) argued for the superiority of a sentence analysis to a T-unit analysis. The present findings suggest that the clause may serve as an informative unit of analysis for L2 writing researchers.

Implications for ESL Assessment and Pedagogy

Wolfe-Quintero et al. (1998) pointed out several potential applications of developmental measures in language assessment and pedagogy, including test validation, program placement, end-of-course assessment, and trait analysis of holistic ratings, among others. These applications necessitate a solid understanding of how developmental measures work. The present investigation of which measures significantly differentiate between proficiency levels, what magnitudes are required for between-proficiency differences to reach statistical significance, what patterns of development are exhibited for these measures, and how different measures relate to each other contributes to achieving this understanding.

Some ESL assessment studies used developmental measures to validate placement tests (e.g., Arnaud, 1992) and noted the need for more studies of the relations between objective measures of textual features and ratings (e.g., Connor-Linton, 1995). Other studies examined the syntactic features of ESL writing by students at different L2 proficiency levels (e.g., Ferris, 1994) and of ESL academic texts in comparison to L1 academic texts (e.g., Hinkel, 2003). As Wolfe-Quintero et al. (1998) argued, these types of studies could inform placement and promotion decisions. At the same time, objective measures, including developmental measures of syntactic complexity, are often used by ESL instructors to assess student level at a given time or progress over time. The present findings can be used to aid such assessment studies and practices in determining what factors should be taken into account in using syntactic complexity measures, which measures are especially pertinent to focus on, and how to interpret the differences observed in these measures.

Syntactic complexity has also been shown to be a relevant construct in materials development and syllabus design. For example, Gaies (1979) examined the relationship between syntactic complexity and readability of ESL reading materials and reported some degree of correlation between the overall syntactic complexity of an ESL reader and its claimed target audience. Pica (1984) argued that syllabus design can benefit from attention to both issues of L1 transfer and target language complexity, where the latter pertains to the notion that a syllabus that presents structures to the learner in order of increasing linguistic complexity reduces learning difficulty. The present results on the nature of a wide range of syntactic complexity measures will prove useful to studies on these aspects of language pedagogy as well as materials development and syllabus design practices that take advantage of insights from such studies.

Finally, for ESL instruction, the present findings indicate the importance for ESL instructors to be aware of the developmental patterns of the syntactic complexity measures. For example, the significant negative changes from lower to higher levels shown by clauses per sentence (C/S) and dependent clauses per clause (DC/C) and per T-unit (DC/T) on the one hand and the significant positive changes from lower to higher levels shown by the complex nominal and coordinate phrase measures on the other imply that instructors may wish to help students improve the ability to engage complexity at the phrasal level as they progress to advanced proficiency levels.

Limitations and Future Research

Given the scope of this research and the information available in the WECCL, several important issues were not taken up in this study. These will be pursued in future research. First, as discussed earlier, the use of school level in this study to conceptualize proficiency level obtains the same validity as the use of program level in other ESL contexts. However, the construct validity of the developmental measures of syntactic complexity can be further tested in future replications of this study using multiple datasets that facilitate other conceptualizations of proficiency level, such as holistic ratings.

Second, the WECCL contains samples produced by L1 Chinese learners only. To determine which aspects of syntactic complexity are developmental across heterogeneous L1 groups and which aspects are affected by the learner's L1, one needs to compare results from different L1 groups. Note that this is very different from simply analyzing data from one single group with learners of mixed L1 backgrounds. This latter approach renders the reliability of the results obtained contingent on the untested assumption that the learner's L1 does not significantly affect the relationship between syntactic complexity and language development. In relation to the first issue mentioned earlier, future research in this direction calls for efforts to compile large-scale learner corpora that not only include adequate subgroups of data from learners of diverse L1 backgrounds but also facilitate coherent conceptualizations of proficiency level across those subgroups.

Finally, the insights obtained from this study can be used to inform future investigations on how developmental measures of fluency, accuracy, and complexity at various linguistic levels, including the lexicon, morphology, syntax, semantics, pragmatics, and discourse, relate to and interact with each other as indices of language development in L2 writing.

THE AUTHOR

Xiaofei Lu is an assistant professor of applied linguistics at The Pennsylvania State University. His research interests are primarily in computational linguistics, corpus linguistics, and intelligent computer-assisted language learning.

REFERENCES

Arnaud, P. J. L. (1992). Objective lexical and grammatical characteristics of L2 written compositions and the validity of separate-component tests. In P. J. L. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 133–145). London, England: Macmillan.

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly*, *26*, 390–395. doi:10.2307/3587016.

Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition*, *11*, 17–34. doi:10.1017/S0272263100007816.

Beers, S. F., & Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre? *Reading and Writing, 22*, 185–200. doi:10.1007/s11145-007-9107-5.

Breiner-Sanders, K. E., Swender, E., & Terry, R. M. (2001). *ACTFL proficiency guidelines—Writing* (Revised 2001). Alexandria, VA: American Council on the Teaching of Foreign Languages.

Buckingham, T. (1979). The goals of advanced composition instruction. *TESOL Quarterly, 13*, 241–254. doi:10.2307/3586213.

Casanave, C. (1994). Language development in students' journals. *Journal of Second Language Writing, 3*, 179–201. doi:10.1016/1060-3743(94)90016-7.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly, 29*, 762–765. doi:10.2307/3588174.

Cooper, T. C. (1976). Measuring written syntactic patterns of second language learners of German. *The Journal of Educational Research, 69*, 176–183.

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26*, 59–84. doi:10.1017/S0272263104261034.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28*, 414–420. doi:10.2307/3587446.

Gaies, S. J. (1979). Linguistic imput in formal second language learning: The issues of syntactic gradation and readability in ESL materials. *TESOL Quarterly, 13*, 41–50. doi:10.2307/3585974.

Gaies, S. J. (1980). T-unit analysis in second language research: Applications, problems and limitations. *TESOL Quarterly, 14*, 53–60. doi:10.2307/3586808.

Henry, K. (1996). Early L2 writing development: A study of autobiographical essays by university-level students of Russian. *The Modern Language Journal, 80*, 309–326. doi:10.2307/329438.

Hinkel, G. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly, 37*, 275–301. doi:10.2307/3588505.

Hirano, K. (1991). The effect of audience on the efficacy of objective measures of EFL proficiency in Japanese university students. *Annual Review of English Language Education in Japan, 2*, 21–30.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly, 18*, 87–107. doi:10.2307/3586337.

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

Hunt, K. W. (1970). Do sentences in the second language grow like those in the first? *TESOL Quarterly, 4*, 195–202. doi:10.2307/3585720.

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing, 4*, 51–69. doi:10.1016/1060-3743(95)90023-3.

Kameen, P. (1979). Syntactic skill and ESL writing quality. In C. Yorio, K. Perkins. & J. Schachter (Eds.), *On TESOL ' 79: The learner in focus* (pp. 343–364). Washington, DC: TESOL.

Klein, D., & Manning, C. D. (2003). Fast exact inference with a factored model for natural language parsing. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 3–10). Cambridge, MA: MIT Press.

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly, 12*, 439–448. doi:10.2307/3586142.

Larsen-Freeman, D. (1983). Assessing global second language proficiency. In H. W. Seliger & M. Long (Eds.), *Classroom-oriented research in second language acquisition* (pp. 287–305). Rowley, MA: Newbury House.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*, 474–496. doi:10.1075/ijcl.15.4.02lu.

Mendelsohn, D. J. (1983). The case for considering syntactic maturity in ESL and EFL. *International Review of Applied Linguistics, 21*, 299–311. doi:10.1515/iral.1983.21.4.299.

Monroe, J. H. (1975). Measuring and enhancing syntactic fluency in French. *The French Review, 48*, 1023–1031.

Ortega, L. (2000). *Understanding syntactic complexity: The measurement of change in the syntax of instructed L2 Spanish learners* (Unpublished doctoral dissertation). University of Hawaii, Manoa, HI.

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. doi:10.1093/applin/24.4.492.

Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly, 11*, 651–671. doi:10.2307/3586618.

Pica, T. (1984). L1 transfer and L2 complexity as factors in syllabus design. *TESOL Quarterly, 18*, 689–704. doi:10.2307/3586583.

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47*, 101–143. doi:10.1111/0023-8333.31997003.

Sharma, A. (1980). Syntactic maturity: Assessing writing proficiency in a second language. In R. Silverstein (Ed.), *Occasional papers in linguistics, No. 6* (pp. 318–325). Carbondale, IL: Southern Illinois University.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly, 27*, 657–677. doi:10.2307/3587400.

Sotillo, S. M. (2000). Discourse functions and syntactic complexity in synchronous and asynchronous communication. *Language Learning and Technology, 4*, 82–119.

Stockwell, G., & Harrington, M. (2003). The incidental development of L2 proficiency in NS-NNS email interactions. *CALICO Journal, 20*, 337–359.

Way, D. P., Joiner, E. G., & Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of French. *The Modern Language Journal, 84*, 171–184. doi:10.1111/0026-7902.00060.

Wen, Q., Wang, L., & Liang, M. (2005). *Spoken and written English corpus of Chinese learners.* Beijing, China: Foreign Language Teaching and Research Press.

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

Yau, M. (1991). The role of language factors in second language writing. In L. Malave & G. Duquette (Eds.), *Language, culture and cognition: A collection of studies in first and second language acquisition* (pp. 266–283). Clevedon, England: Multilingual Matters.