Available online at www.sciencedirect.com

ScienceDirect

Journal of Second Language Writing 26 (2014) 66–79

JOURNAL OF SECOND LANGUAGE WRITING

Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners

Scott A. Crossley [a,*], Danielle S. McNamara [b,1]

[a] Department of Applied Linguistics/ESL, Georgia State University, 34 Peachtree Street, Suite 1200, One Park Tower Building, Atlanta, GA 30303, United States

[b] Department of Psychology, Arizona State University, Tempe, AZ 85287, United States

Abstract

This study examines second language (L2) syntactic development in conjunction with the effects such development has on human judgments of writing quality (i.e., judgments of both overall writing proficiency and more fine-grained judgments of syntactic proficiency). Essays collected from 57 L2 learners in a longitudinal study were analyzed for growth and scoring patterns using syntactic complexity indices calculated by the computational tool Coh-Metrix. The analyses demonstrate that significant growth in syntactic complexity occurred in the L2 writers as a function of time spent studying English. However, only one of the syntactic features that demonstrated growth in the L2 learners was also predictive of human judgments of L2 writing quality. Interpretation of the findings suggest that over the course of a semester, L2 writers produced texts that were increasingly aligned with academic writing (i.e., texts that contain more nouns and phrasal complexity), but that human raters assessed text quality based on structures aligned with spoken discourse (i.e., clausal complexity). Thus, this study finds that the syntactic features that develop in L2 learners may not be the same syntactic features that will assist them in receiving higher evaluations of essay quality.

Keywords: Computational linguistics; L2 writing; Writing development; Writing quality; Syntactic complexity

Introduction

Syntactic development is an important component of second language (L2) acquisition and one that has received considerable attention in previous research (Hawkins, 2001; Lu, 2010) in both longitudinal and cross-sectional studies. Researchers have focused on L2 syntactic development under the notion that the ability to arrange words syntactically into phrases and phrases into clauses demonstrates the capacity to manipulate a language's combinatorial properties, which is argued to be a strong indicator of general language acquisition. One of the primary questions addressed by syntactic research is how syntactic knowledge develops over time and, more specifically, what syntactic features develop early and which develop later for L2 learners (Hawkins, 2001).

* Corresponding author. Tel.: +1 404 413 5179.
E-mail addresses: sacrossley@gmail.com (S.A. Crossley), dsmcnamara1@gmail.com (D.S. McNamara).
[1] Tel.: +1 480 727 5690.

http://dx.doi.org/10.1016/j.jslw.2014.09.006
1060-3743/

Examinations into the development of syntactic features often focus on the variation and sophistication of the phrases and clauses produced by L2 learners. The basic premise underlying such examinations is that syntactic complexity can be used to directly measure L2 learner proficiency (Foster & Skehan, 1996; Lu, 2011; Ortega, 2003; Wolfe-Quintero, Inagaki, & Kim, 1998).

While a number of studies have examined longitudinal growth in L2 learners using both spoken and written corpora, few studies have examined L2 syntactic development in conjunction with the relationships such developments have with human judgments of writing quality (both judgments of overall writing proficiency and more fine-grained judgments of syntactic proficiency). That is to say, while past research has focused on L2 learner development, it has rarely linked the effects of such development to assessments of language proficiency. However, such an approach is important because it can afford an opportunity to examine not only syntactic growth, but also the relations of such growth with the judgments of expert raters. To address this research gap, this study examines L2 writing samples using computational indices of syntactic complexity to understand how syntactic complexity changes over time in L2 writers (i.e., longitudinal growth) and to understand how changes in syntactic complexity are related to human ratings of language use in L2 writing.

Syntactic complexity

As mentioned earlier, syntactic complexity refers to the sophistication of syntactic forms produced by a speaker or writer and the range or variety of syntactic forms produced (Lu, 2011; Ortega, 2003). Analysis of L2 output in terms of its syntactic complexity is a common means to L2 growth because language development in L2 learners is argued to entail the acquisition and production of less frequent syntactic features along with the use of a greater variety of syntactic features. Many features related to syntactic complexity are relatively easy to investigate using both hand- and automated-coding of texts which allows for the sampling of a variety, but by no means all, of available syntactic features.

The traditional method of measuring syntactic complexity is with T-units (Biber, Gray, & Poonpon, 2011), which can be defined as the shortest allowable grammatical units that can be punctuated at the sentence level (i.e., the main clause plus additional, embedded subordinated clauses; Street, 1971 as cited in Larsen-Freeman, 1978, p. 441). T-units were initially used to assess writing development in first language (L1) writers (Hunt, 1965) and were later adopted for use by the L2 research community (Casanave, 1994; Henry, 1996; Lu, 2011; Ortega, 2003; Stockwell & Harrington, 2003). The use of T-units as measures of syntactic complexity for L2 learners has provided mixed results, with some studies demonstrating no links between classic T-unit measures such as mean length of T-unit and measures of L2 syntactic growth (Bardovi-Harlig, 1992; Casanave, 1994; Ishikawa, 1995) and other studies finding strong links (Ortega, 2003; Stockwell & Harrington, 2003).

The most promising T-unit indices are error-free T-units (Larsen-Freeman, 1978), but such indices are not strictly syntactic and focus more on accuracy than T-units. Additionally, such indices are difficult, if not impossible, to implement computationally and require expert hand coding, which is prone to subjectivity and error. The use of T-units to investigate L2 writing has also been called into question recently by Biber et al. (2011). They found that the clausal subordination measured by T-unit indices is more common in conversation whereas academic writing is characterized syntactically by the use of noun phrase constituents and complex phrases.

Other measures of syntactic complexity that are not specifically based on T-units but are commonly used in L2 writing studies include indices that measure the length of syntactic structures, the types and incidence of embeddings, the types and number of coordinations between clauses, the range and types of phrasal units produced, and the frequency of clauses and phrases used (Ortega, 2003). Such indices can be accessed in computational tools such as the Biber tagger (Biber, 1988) and Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, Graesser, McCarthy, & Cai, 2014).

Syntactic development in L2 learners

Previous research into L2 syntactic acquisition has focused on syntactic development in both spoken and written L2 language samples and has demonstrated that L2 learners follow general patterns of syntactic development that occur in identifiable stages. For instance, English speakers learning French must acquire the rule that direct and indirect object pronouns come before the verb (as compared to after the verb in English). When learning such a rule, L2 learners

generally first produce postverbal pronouns, followed by preverbal pronouns. However, when preverbal pronouns do occur, they compete with omitted objects (Selinker, Swain, & Dumas, 1975; White, 1996). L2 learners of English also generally follow the accessibility hierarchy with respect to the acquisition of relative clauses (Gass, 1979) in which L2 learners first acquire subject relative clauses followed by direct-object, indirect-object, and object-of-a-preposition relative clauses. Other syntactic patterns demonstrated by L2 learners include the development of question formations (from wh-fronting, to auxiliary verb before the subject, to the subject verb inversion found in yes/no questions; Eckman, Moravcsik, & Wirth, 1989) and negation formations (from *no*, to *don't*, to *not*, to auxiliary verbs plus *not*; Schumann, 1979).

Patterns in syntactic development have also been noted in numerous longitudinal studies of L2 writing (e.g., Casanave, 1994; Ishikawa, 1995; Stockwell & Harrington, 2003). Casanave (1994) examined growth in syntactic complexity by examining the journal writing of intermediate Japanese English learners over the course of three semesters of instruction. Casanave found that as L2 learners developed over time, they began to produce longer and more complex syntactic clauses (as measured by T-unit indices) that were also more accurate. Ishikawa (1995) examined two groups of low proficiency L2 English learners at the beginning and at the end of a semester of instruction. Ishikawa found that two accuracy indices (total words in error-free clauses and error-free clauses per composition) best discriminated between writings produced at the beginning of the semester and end of the semester. Lastly, Stockwell and Harrington (2003) investigated L2 syntactic growth in e-mail exchanges over a five-week period. Syntactic complexity was measured using T-unit indices and human judgments of quality (but links were not made between the two). Stockwell and Harrington found that L2 learners showed differences in the average number of words per T-unit, the average number of words per error-free T-unit, and the percentage of error-free T-units as a function of time spent writing. They also reported that human ratings of syntactic complexity increased over the same five-week period.

Another approach to investigating syntactic development in L2 learners is through cross-sectional studies, which can be used to investigate differences between proficiency levels in L2 writers (e.g., Ferris, 1994; Larsen-Freeman, 1978; Lu, 2011; Ortega, 2003). Larsen-Freeman (1978) used T-unit indices to discriminate between essays based on the placement levels of L2 learners (212 learners placed into five proficiency levels). The results demonstrated that the percentage of error-free T-units and the average length of error-free T-units were the best discriminators of proficiency. Ferris (1994) examined essays written by 160 L2 learners that were divided into high and low proficiency groups. Using a variety of lexical and syntactic indices, Ferris found that high proficiency L2 writers differed from low proficiency L2 writers in their more frequent production of passives, nominalizations, conjuncts, and prepositions. More proficient L2 writers also produced a greater number of relative and adverbial clauses. Ortega (2003), in a synthesis study, found that length and T-unit syntactic indices such as mean length of sentence, mean length of T-unit, mean length of clause, and clauses per T-unit were reliable indicators of proficiency level differences for L2 writers. More recently, Lu (2011) investigated the performance of 14 T-unit indices to distinguish between grade levels for essays written by university level L2 learners. Lu found that 10 of the 14 indices discriminated between grade level, but only seven of the 10 indices progressed linearly across proficiency levels. These indices included three indices of length production, two indices of complex nominals, and two indices of coordinated phrases.

Syntactic features and human judgments of writing quality

Another approach to assessing writing development is to examine how linguistic features in a text can predict human ratings of essay quality. Such an approach is built on the notion that syntactic features of texts are prime indicators of syntactic development because the presence of more sophisticated syntactic features will lead to higher ratings of essay quality.

Such predictions have been borne out in studies of both L2 and L1 writing. For instance, studies have indicated that higher rated L2 essays contain greater subordination (Grant & Ginther, 2000), use of passive voice (Ferris, 1994; Grant & Ginther, 2000), and while containing fewer present tense forms (Reppen, 1994), and base verb forms (Crossley & McNamara, 2012). Similar findings have been reported in L1 studies of writing quality with higher quality L1 essays containing greater syntactic complexity (as measured by the number of words before the main verb; McNamara, Crossley, & McCarthy, 2010) and a greater incidence of verb base forms (Crossley, Roscoe, McNamara, & Graesser, 2011).

Method

The purpose of this study is to assess syntactic development in L2 writers as a function of time spent in a writing course. To this end, we use a number of automated syntactic complexity indices to assess syntactic differences in descriptive essays written by L2 learners at the beginning and at the end of a semester-long writing course. We complement this analysis by assessing how well the same syntactic indices are able to predict the variance in human ratings of essay quality for essays written throughout the course. In doing so, we address two key questions: (1) Do L2 writers demonstrate syntactic development over the course of a semester (i.e., longitudinal growth) and (2) Does this growth correspond to syntactic features that predict human ratings of writing proficiency.

Corpus

The data for this analysis were collected from 70 university-aged L2 writers at Michigan State University during a single semester of instruction in an intensive writing class. The participants were from the two highest levels at a university ESL program and from one level of an English for Academic Purposes (EAP) program (see Connor-Linton & Polio, 2014 this volume for additional information about the dataset used in this study). From this dataset, we selected writing samples from the 57 participants who completed all three writing assignments collected at the beginning, middle, and end of the semester. These essays were timed descriptive essays written in 30 minutes. The essays averaged 335.4 words (SD = 97.5) and 5.4 paragraphs (SD = 4.165) in length. Prior to analysis the corpus was cleaned to eliminate formatting and spelling errors.

Human ratings

Two expert raters assessed the quality of each essay using a composition grading scale that required the raters to rate each essay on five different analytical features: content, organization, vocabulary, language use, and mechanics (see Connor-Linton & Polio, 2014 this volume for additional information about the grading scale). These analytic ratings were combined into an overall rating for each essay. Of interest for this study is the combined rating for each essay and the Language Use rating, which includes assessments of syntactic properties. Briefly, the Language Use rating equates higher writing proficiency with no errors that interfere with comprehension, few morphological errors, no major errors in word or structure, the use of more complex sentences, and excellent sentence variety. The latter three properties are strongly related to syntactic complexity while the former two are linked to syntactic complexity, but are not exclusively syntactic (i.e., they also have links to grammar, morphology, and the lexicon). Interrater reliability between the two raters for the essays written by the 57 participants in this study was strong: $r = .767$ for Language Use ratings and $r = .880$ for overall ratings. These two ratings also demonstrated strong multicollinearity, $r = .914$.

Selected syntactic indices

We selected syntactic indices from Coh-Metrix, an advanced computational tool that measures cohesion and linguistic sophistication at various levels of language, discourse and conceptual analysis (Graesser et al., 2004; McNamara & Graesser, 2012; McNamara et al., 2014). To ensure we were assessing syntactic complexity, we selected only those syntactic indices that measure clausal and phrasal level syntactic features. These indices include incidence counts taken from the Charniak (2000) part of speech tagger/parser (normed for text length) in addition to ratio scores, raw scores, and length counts.

We used an automatic approach to assessing syntactic complexity because it affords speed, flexibility, and reliability. In addition, human raters are prone to subjectivity and require training, time to score, and monitoring, all of which consume resources (Higgins, Xi, Zchner, & Williamson, 2011). However, one potential problem with using a parser to investigate syntactic complexity is accuracy. For texts written by L1 speakers, the Charniak parser reports an average accuracy of 89% for expository and narrative texts (with greater accuracy reported for narrative texts; Hempelmann, Rus, Graesser, & McNamara, 2006). No studies that we know of have investigated the accuracy of the Charniak parser or similar parsers for L2 writing, but it can be presumed that the accuracy would decrease. Thus, a question remains about the degree to which parser accuracy is affected by L2 writing and, more importantly, how this accuracy compares with hand-coded ratings of syntactic complexity, which are also subject to accuracy limitations.

In total, we selected 11 Coh-Metrix indices that measure clausal and phrasal features of language. These indices include measurements of syntactic variety, syntactic transformations (e.g., negations and questions), syntactic embeddings, incidence of phrase types, and phrase length. Each index in Coh-Metrix is computed using the output produced by the Charniak (2000) parser for both lexical (i.e., part of speech tags) and syntactic categories (phrasal and clausal components). These indices selected are similar to Bulté and Housen (2014), but are calculated automatically as compared to manually. The selected indices also target clausal, phrasal, and sentential elements, while Bulté and Housen's indices focus more on sentential elements. The selected indices are discussed in greater detail in the following section.

Sentence variety. Sentence variety is assessed in Coh-Metrix by measuring the consistency and uniformity of the clausal, phrasal, and part of speech (POS) constructions located in the text (i.e., syntactic similarity). The syntactic similarity indices in Coh-Metrix assess syntactic similarity by comparing adjacent sentences for similar clausal, phrasal, and POS constructions. More uniform syntactic constructions result in less complex syntax that is easier for the reader to process (Crossley, Greenfield, & McNamara, 2008). However, less syntactic similarity is a hallmark of advanced writers (Crossley, Weston, McClain-Sullivan, & McNamara, 2011).

Syntactic transformations. Coh-Metrix measures a number of syntactic elements related to syntactic transformations. These include negations and wh-questions. Such transformations represent a syntactic complexity beyond the use of simple declarative sentences. These indices are computed using normalized incidences of occurrences.

Syntactic embeddings. Coh-Metrix also calculates syntactic embeddings as calculated by the Charniak parser. The embeddings reported by Coh-Metrix are in the form of normalized incidence counts and include counts for all clauses (including matrix clauses, coordinated clauses, and embedded clauses), infinitive clauses, S-bar counts (i.e., embedded sentences that can be marked with complementizers such as *that*, *for*, *who*, and *when*, prepositions such as *after* and *before*, conditionals such as *if* and *then*, and subordinating conjunctions such as *because* or *however*), 'that' verb complements, and relative clauses. Such embeddings generally indicate greater syntactic complexity.

Phrase types. Coh-Metrix computes incidence scores for a variety of phrase types. These phrase types include noun phrases (NP: related to density of propositions), verb phrases (VP: related to the number of clauses in a sentence), and preposition phrases (PP: related to the number of phrases that provide adjectival and adverbial information). In the sentence *The boy eats the pepperoni pizza under the tree*, the phrasal count for NP would be two (i.e., *The boy* and *the pepperoni pizza*), the phrasal count for VP would be one (i.e., *eats the pepperoni pizza under the tree*), and the phrasal count for PP would be one (i.e., *under the tree*). An example of multiple clauses in one sentence is found in the example *She sees that the boy is eating pepperoni pizza under the tree* in which the phrasal count for NP would increase by one (i.e., with the inclusion of *she*) and the phrasal count for VP would increase by one (i.e., with the inclusion of *sees that*. . .).

Phrase length. Coh-Metrix reports a variety of indices related to syntactic complexity that result from phrase length calculations. These include the length of noun phrases and verb phrases (under the hypothesis that longer phrases are more difficult to process) and the number of words before the main verb (under the hypothesis that the main verb controls the arguments in the sentence and the longer it takes to access the main verb, the more complex the sentence is; McNamara et al., 2010).

Statistical analysis

Our statistical analyses address two principal questions. Our first question is whether growth in the syntactic patterns of L2 learners' is evident in their writing. For this analysis, we first conducted within-subjects Analysis of Variance (ANOVA) using the selected Coh-Metrix indices focusing on the first and the last essays written over the course of the semester ($n = 114$). We did not focus on the middle essays because we did not expect syntactic growth to occur within an 8-week period. The ANOVA analysis provided us with information about which syntactic indices demonstrated significant growth patterns. Those indices that demonstrated significant growth patterns were then entered into a Naïve Bayes classifier to assess how well the indices predicted if an essay was written at the beginning of the semester or at the end of the semester. A Naïve Bayes classifier produces a statistical learning model that assigns a probability to each text given a number of instances (i.e., what is the probability, based on the syntactic features found within that text, that a text was written at the beginning or the end of the semester).

Our second question addressed whether syntactic indices were predictive of human ratings of essay quality. To answer this question, we conducted regression analyses to examine if the selected Coh-Metrix indices were predictive of human ratings of essay quality (both language use and combined score ratings). For this analysis we used all the rated essays in the analysis ($N = 171$). We first conducted Pearson Product Moment Correlations between the human ratings for the essay and the syntactic indices. Those indices that demonstrated significant correlations were then included in a stepwise regression analysis to examine how well the indices could predict the variance in the human ratings.

For the Naïve Bayes analysis, we tested the predictive strength of the indices using a leave-one-out-cross-validation (LOOCV) analysis (Witten, Frank, & Hall, 2011). In this analysis, we chose a fixed number of folds that equaled the number of observations (i.e., 114 essays). In LOOCV, one observation in turn is left out for testing and the remaining instances are used as the training set (i.e., in the case of the Naïve Bayes analysis, the 113 remaining essays). We assess the accuracy of the model by testing its ability to predict the classification (the human rating) of the omitted instance. Such an approach affords us the opportunity to test the models generated by the Naïve Bayes classifier on an independent data set (i.e., on essays that are not used to train the model). If the LOOCV results demonstrate significant classification results as reported by a Chi-squared test, our level of confidence in the model increases supporting the extension of the analysis to external data sets.

For the regression analysis, we used training and test sets to assess the generalizability of the regression model to an outside corpus. We divided the corpus of 171 essays into training and test sets following a 67/33 split (Witten et al., 2011). For the training set, we first conducted Pearson correlations to assess relationships between the selected variables and the human ratings. Those variables that demonstrated significant correlations with the human ratings were retained as predictors in a subsequent regression analysis. We next conducted a stepwise regression analysis using the essays in the training set only. The model from this regression analysis was then applied to the held back essays in the test set to predict their ratings.

For each analysis, we also control for multicollinearity by examining correlations between indices and ensuring that the indices were not strongly related (i.e., $r > .070$). In addition, we control for overfitting in the models by ensuring a 15/1 item to predictor ratio. Such controls allow us to use only variables that contribute uniquely to the models and to verify that the findings of the analysis are not the result of random noise in the data.

Results

Longitudinal growth

We first conducted repeated measure ANOVAs on the human ratings assigned to the essay to examine if, according to expert raters, there was an improvement between essays written at the beginning of the semester (1st essays) and at the end of the semester (3rd essays). Language use ratings increased significantly from the first essay ($M = 9.9$, SD = 2.1) to the third essay ($M = 11.1$, SD = 1.9), $F(1, 56) = 27.815$, $p < .001$, η_p^2 (partial eta squared) = .332, and the combined ratings increased from the first essay ($M = 47.4$, SD = 10.3) to the third essays ($M = 57.1$, SD = 8.3), $F(1, 56) = 47.378$, $p < .001$, $\eta_p^2 = .458$.

Repeated measure ANOVAs were then conducted on the selected Coh-Metrix indices to examine if significant differences in syntactic features existed between essays written at the beginning of the semester (1st essays) and at the end of the semester (3rd essays). Those indices that showed significant differences were then used in a confirmatory Naïve Bayes classifier algorithm to predict whether the essays were written at the beginning or at the end of the semester. Of the 11 indices, six demonstrated significant differences between the 1st and 3rd essays (see Table 1 for details): the incidence of all clauses, the number of modifiers per noun phrase, syntactic similarity, number of verb phrases, number of words before the main verb, and incidence of the negation word "not."

The Naïve Bayes classifier using the six significant syntactic indices and LOOCV correctly allocated 71 of the 114 essays in the total set, χ^2 (df = 1, $n = 114$) = 7.737, $p < .010$, for an accuracy of 62.28% (the chance level for this analysis is 50%). The results from the Naïve Bayes classifier are reported in the confusion matrix found in Table 2. The measure of agreement between the actual essay number and that assigned by the model produced a Cohen's Kappa of 0.246, demonstrating a fair agreement.

To illustrate the Naïve Bayes classifier results, we provide precision, recall, and F1 scores (see Table 3). Recall scores represent the number of true positive hits (i.e., correct predictions) over the number of hits + false negatives

Table 1
Within subjects ANOVA: Differences between beginning and end of semester.

Index	1st essay, mean (SD)	3rd essay, mean (SD)	F	p	η^2
Incidence of all clauses (matrix, coordinating, and embedded clauses)	171.40 (25.12)	158.26 (23.31)	9.405	0.003	0.144
Number of modifiers per noun phrase	0.65 (0.14)	0.73 (0.20)	7.813	0.007	0.122
Syntactic similarity score	0.14 (0.04)	0.13 (0.03)	6.888	0.011	0.110
Number of verb phrases	0.77 (.0.04)	0.75 (0.05)	6.429	0.014	0.103
Number of words before main verb	3.25 (1.16)	3.74 (1.37)	5.420	0.024	0.088
Incidence of negation "not"	1.54 (1.65)	2.30 (2.59)	4.050	0.049	0.067
Incidence of prepositional phrases	94.89 (21.59)	102.26 (20.96)	3.401	0.070	0.057
Incidence of subject relative clauses	0.60 (0.92)	0.98 (1.55)	2.563	0.115	0.044
Incidence of "that" verb complements	12.72 (5.63)	12.30 (5.76)	0.267	0.607	0.005
Incidence of S-bars	47.67 (14.71)	46.57 (14.96)	0.189	0.665	0.003
Incidence of infinitives	6.04 (4.80)	6.18 (4.04)	0.038	0.846	0.001

(e.g., the number of first essays that were misclassified as last essays) while precision scores represent the number of hits divided by the number of hits + false positives (e.g., the number of last essays that were classified as first essays). Thus, for the 45 essays out of the 57 essays written at the beginning of the semester that were correctly classified, the recall score is (45/(45 + 12)) = 79%. For the same essays, the precision score is (45/(45+31)) = 59%. The F1 score is a weighted average of the precision and recall results. The model performed best at classifying essays written at the beginning of the semester. The overall accuracy of the model was .606 (the average F1 score).

Regression analyses: language use

Correlations training set. Correlations were conducted between the syntactic indices and the human ratings of language use for the 113 essays in the training set. Six Coh-Metrix indices demonstrated significant correlations with the human ratings while not demonstrating multicollinearity with one another (see Table 4).

Regression analysis training set. A stepwise regression analysis using the six indices as the independent variables to predict the human ratings of language use yielded a significant model, $F(3, 109) = 13.013$, $p < .001$, $r = .514$, $r^2 = .264$. Three syntactic indices were included as significant predictors of the human ratings: *Incidence of all clauses*, *infinitives*, and "*that*" *verb complements*. The model demonstrated that the three indices explained 26.4% of the variance in the human ratings of language use for the essays in the training set (see Table 5 for additional information).

Regression analysis test set. We used the model reported for the training set to predict the human ratings of Language Use in the test set. To determine the predictive power of the three variables retained in the regression model, we computed an estimated rating for each essay in the test set using the B weights and the constant from the training set regression analysis. A Pearson's correlation was then conducted between the estimated rating and the actual rating of each of the essays in the test set. This correlation together with its r^2 was then calculated to determine the predictive accuracy of the training set regression model on the independent data set.

The regression model, when applied to the test set, reported $r = .457$, $r^2 = .208$. The results from the test set model demonstrated that the combination of the three syntactic indices accounted for 20.8% of the variance in the language use ratings of the essays in the test set.

Table 2
Confusion matrix for Naïve Bayes classification.

		First essay	Last essay
LOOCV set	First essay	**45**	12
	Last essay	31	**26**

Table 3
Precision and recall findings for Naïve Bayes classification.

LOOCV set			
Text set	Recall	Precision	F1
First essay	0.79	0.59	0.68
Last essay	0.46	0.68	0.55

Regression analyses: combined ratings

Correlations training set. Correlations were conducted between the syntactic indices and the combined human ratings of the 113 essays in the training set. Seven Coh-Metrix indices demonstrated significant correlations with the human ratings while not demonstrating multicollinearity with one another (see Table 6).

Regression analysis training set. A stepwise regression analysis using the seven indices as the independent variables to predict the combined human ratings yielded a significant model, $F(3, 109) = 19.659$, $p < .001$, $r = .593$, $r^2 = .350$. Three syntactic indices were included as significant predictors of the essay ratings: *Incidence of all clauses*, *infinitives*, and ''*that*'' *verb complements*. The model demonstrated that the three indices explained 35.0% of the variance in the combined human ratings of the essays in the training set (see Table 7 for additional information).

Regression analysis test set. The regression model, when applied to the test set, reported $r = .562$, $r^2 = .316$. The results from the test set model demonstrated that the combination of the three syntactic indices accounted for 31.6% of the variance in the combined ratings of the essays in the test set.

Discussion

This analysis has demonstrated that significant growth in syntactic complexity occurs in L2 writers as a function of time spent in a writing class. From the beginning of the semester until the end of the semester, L2 writers in this study produced fewer incidences of all clauses, longer noun phrases, less syntactic similarity between sentences, fewer verb phrases, more words before the main verb, and more negation. However, only one of these syntactic features was also predictive of human judgments of L2 writing quality: fewer incidences of all clauses. The other two predictors of writing quality, incidence of infinitives and incidence of ''that'' verb complements did not demonstrate significant growth patterns between the first and last essays for the L2 writers in this analysis. Thus, while L2 learners' writing does become more syntactically complex, most of the syntactic features demonstrating growth are not predictive of human judgments of writing quality.

In reference to L2 syntactic development, this study has shown that a number of syntactic complexity indices demonstrate growth in predicted directions from the first to the final essay of the semester. The strongest growth, as indicated by the effect size, was for the use of all clauses. Over the course of the semester, the L2 writers in this study produced fewer clauses overall (i.e., fewer matrix and embedded clauses). L2 writers also showed changes in their use

Table 4
Pearson correlations: Syntactic values to language use scores.

Index	*r*	*p*
Incidence of all clauses (matrix, coordinating, and embedded clauses)	−0.295	0.002
Incidence of ''that'' verb complements	0.249	0.008
Incidence of infinitives	0.234	0.013
Number of verbs phrases	−0.229	0.015
Average number of modifiers per noun phrase	0.211	0.025
Incidence of negation ''not''	0.194	0.039
Incidence of preposition phrases	0.179	0.058
Sentence syntax similarity score	−0.169	0.074
Mean number of words before the main verb	0.120	0.204
Incidence of S-bars	−0.093	0.327
Incidence of subject relative clauses	0.084	0.376

Table 5
Linear regression analysis to predict language use scores: Training set.

Entry	Variable added	r	r^2	B	SE	Beta
Entry 1	Incidence of all clauses (matrix, coordinating, and embedded clauses)	0.295	0.087	−0.031	0.006	−0.403
Entry 2	Incidence of infinitives	0.429	0.184	0.145	0.039	0.317
Entry 3	Incidence of "that" verb complements	0.514	0.264	0.103	0.030	0.285

Notes: Estimated constant term is 13.408; B is unstandardized Beta; SE is standard error; Beta is standardized Beta.

of phrasal complexity, producing longer noun phrases and more words before the main verb. Conversely, while producing more complex noun phrases, L2 writers began to produce fewer verb phrases, which is indicative of fewer embedded clauses. At both the clausal and phrasal level, L2 writers produced sentences that demonstrated less syntactic similarity with each other over time, indicating the production of a greater variety of syntactic constructions. Lastly, L2 writers produced a greater number of sentences containing "not" negations. Overall, the longitudinal analysis demonstrates that over the course of a semester L2 writers produced text that depended more on noun phrases than verb phrases and text that contained greater phrasal modifications (similar, supporting results are presented in Bulté & Housen, 2014). Knowing that nouns are more important than verbs in academic writing (i.e., academic writing has a nominal style; Fang, Schleppegrell, & Cox, 2006; Halliday, 1989; Halliday & Matthiessen, 1999; Wells, 1960), such a finding provides evidence that advancing L2 writers move toward the production of text that better aligns with an academic writing style. The longitudinal analysis also demonstrates that L2 writers move toward the development of more phrasal components as compared to clausal components as they advance in writing. This notion is evidenced by the movement away from a production of a greater number of matrix and embedded clauses to denser phrasal components such as noun phrases and an increase in the number of words before the main verb. These findings, taken in conjunction with the notion that academic writing relies more on phrasal modification (as compared to dependent clauses, which are typical in speech; Biber, 1985, 1986, 1988, 2006; Biber et al., 2011), provides further evidence that developing L2 writers move toward academic writing as they advance. Lastly, L2 writers began to produce sentences that demonstrate greater variety of structure and more transformations (e.g., negations).

We find a different pattern of results when we analyze the syntactic features that are most predictive of human judgments of syntactic properties and combined writing scores. Similar to our longitudinal analysis, we find that the overall production of clauses is the strongest predictor of both judgments of syntactic proficiency and combined writing proficiency. The negative correlation with human ratings signifies that fewer matrix and embedded clauses are indicative of increased writing quality. However, unlike longitudinal growth patterns, human judgments of writing proficiency are not strongly predicted by nominal style (i.e., a greater emphasis on nouns) or complex phrasal elements. Rather, higher ratings of writing quality are predicted by dependent clause features such as the incidence of infinitives and "that" verb complements (see Table 8 for a comparison of the predictive indices found in each of the analyses). This is similar to the findings of Friginal and Weigle (2014) in which they reported that higher rated

Table 6
Pearson correlations: Syntactic values to combined scores.

Index	r	p
Incidence of all clauses (matrix, coordinating, and embedded clauses)	−0.350	0.001
Infinitives	0.321	0.001
Negation (not)	0.263	0.005
Number of verbs phrases	−0.237	0.012
Incidence of preposition phrases	0.229	0.015
Average number of modifiers per noun phrase	0.213	0.023
"that" verb complements	0.199	0.034
Mean number of words before the main verb	0.174	0.065
Sentence syntax similarity	−0.157	0.097
Incidence of S-bars	−0.130	0.168
Subject relative clauses	0.099	0.297

Table 7
Linear regression analysis to predict combined scores: Training set.

Entry	Variable added	r	r^2	B	SE	Beta
Entry 1	Incidence of all clauses (matrix, coordinating, and embedded clauses)	0.350	0.123	−0.172	0.029	−0.478
Entry 2	Incidence of infinitives	0.543	0.295	0.909	0.170	0.424
Entry 3	Incidence of ''that'' verb complements	0.593	0.351	0.405	0.132	0.239

Notes: Estimated constant term is 71.534; B is unstandardized Beta; SE is standard error; Beta is standardized Beta.

essays and essays written at the end of the semester contained a greater number of complex syntactic structures including syntactic structures related to clause complexity (''that'' clauses and ''to'' clauses).

What is remarkable about this contrast is that increased dependent clauses are argued to be characteristic of speech (i.e., interpersonal spoken registers) and not academic writing (Biber, 1985, 1986, 1988, 2006; Biber et al., 2011). Thus, we are left with the finding that whereas L2 writing develops to more closely match academic writing, human judgments of writing quality (at least in these essays) are not predicted by most of the syntactic features that develop longitudinally in L2 writers. A similar finding is reported in Bulté and Housen (2014). In fact, it appears that the expert raters in this study did not evaluate L2 essays based on the syntactic features that are common in academic writing at all (i.e., nouns and phrases), but rather they evaluated writing samples syntactically based on the use and complexity of clausal components. Therefore, we find a disassociation between L2 syntactic development and judgments of L2 writing quality that leads to the conclusion that the syntactic features that develop in L2 learners are not the same syntactic features that will assist them in receiving higher evaluations for essay quality.

In Appendix A, we provide examples of this disassociation with two essays written by one participant (participant 15) at the beginning of the semester and at the end of the semester. Table 9 provides the syntactic complexity scores reported by Coh-Metrix for these two essays. The essays do not demonstrate differences in the use of the negation ''not.'' However, all indices related to nominal style demonstrate growth in the predicted directions, as do all phrasal indices. The samples also demonstrate differences in syntactic similarity. Conversely, the samples show no differences in reference to dependent clause indices (''that'' verb complements and incidence of infinitives). Likewise, the samples demonstrate no differences in the human scores for Language Use and the combined scores. Thus, while participant 15 shows developmental patterns in both nominal and phrasal elements in writing, this development appears to have little influence on human judgments of essay quality.

The question is then, if academic writing is defined by a nominal style and the use of complex phrasal elements, why are the human ratings in this study predicted by clausal elements that are more indicative of speech? The answer likely lies in the genre of descriptive writing, which may not be a prototypical academic genre, especially for intermediate proficiency writers. Describing homes, campuses, teachers, family, and friends may lead to writing that is characteristic of a more interpersonal register (Biber, 1992). Such a register would likely influence how the human raters judge the quality of the writing. In the case of descriptive writing, it is likely that raters do not expect features of academic writing and thus evaluate the writing based on features more common in spoken discourse. As a result, essays containing more dependent clauses are judged to be of higher quality. In contrast, our findings indicate that lower quality essays contain more clauses in general, including matrix clauses, coordinating clauses, and dependent clauses.

Table 8
Differences between judgments and development.

Predictive of human judgments	Classifiers of development
Incidence of all clauses (matrix, coordinating, and embedded clauses)	Incidence of all clauses (matrix, coordinating, and embedded clauses)
Infinitives	Number of modifiers per noun phrase
''that'' verb complements	Syntactic similarity
	Number of verb phrases
	Number of words before main verb

Table 9
Comparison between 1st and 3rd essays for participant 15.

Syntactic complexity index	1st essay	3rd essay
Syntactic similarity score	0.20	0.09
Number of words before main verb	2.09	4.79
Incidence of all clauses (matrix, coordinating, and embedded clauses)	223.40	150.69
Number of modifiers per noun phrase	0.51	1.04
Incidence of negation "not"	0.00	0.00
Number of verb phrases	0.78	0.69
Incidence of infinitives	5.00	5.00
Incidence of "that" verb complements	4.00	5.00
Language use score	11.00	11.00
Combined score	57.00	56.50

From a practical perspective, such a finding suggests that descriptive writing tasks may not best assess writing development for intermediate level L2 writers. While evaluations of L2 writing development may center on clausal features rather than nominal and phrasal features (Biber et al., 2011), awareness of how different writing tasks may influence human judgments of quality should be an important pedagogical consideration undertaken before assignments are developed and assigned. If the goal of an L2 writing class is to transition L2 writers toward academic writing, then assignments that revolve around comparing and contrasting ideas, producing persuasive arguments, or integrating outside information into an essay may be better suited to evaluate developing syntactic proficiency in L2 writers than descriptive writing. Such proficiency seems to develop relatively quickly and as a result of instruction. Ortega (2003) preliminarily found that two to three months of university level instruction would result in "negligible to small-sized change" (p. 511) and that the rate of change may be greater for L2 learners studying in an English as a second language (ESL) environment as compared to an English as a foreign language (EFL) environment. Our findings support this notion, with small effect sizes reported for learners' syntactic development over a four-month semester in an ESL environment. Knowing that syntactic proficiency can develop in such a small window of time, it seems imperative that assessments of proficiency mirror this development.

Conclusion

This study has provided further evidence of syntactic development in L2 writers. Key to this study was the attempt to link such growth to human ratings of writing and syntactic proficiency. We found that, in most cases, features of syntactic complexity that demonstrated growth patterns in L2 writers were not the same features that predicted judgments of proficiency. In fact, there was a disassociation between L2 syntactic development and judgments of proficiency such that L2 learner growth was associated with greater nominal style and phrasal complexity whereas human judgments were predicted by clausal features (with the exception of the incidence of all clauses, which was an important indicator of L2 growth and human judgments of proficiency).

Some caution should be taken in interpreting these findings. While we investigated a number of indices related to syntactic complexity, we could not examine all potential syntactic complexity indices and not all features of syntactic growth are easily automated within computational tools. Thus, we analyzed a number of local metrics, which have strong links to global metrics of syntactic knowledge, but may not be inclusive of syntactic proficiency (i.e., we did not examine all elements of syntactic complexity). Additionally, we examined a small sample of writers over a relatively brief span of time in an ESL environment only. Future studies may benefit from a larger sample population that is investigated over the course of a year-long program of study. Future studies may also benefit from comparing ESL to EFL learners and instructed versus uninstructed learners. Lastly, this study focused on timed, descriptive writing. Future studies should consider assessing proficiency using a variety of different speaking and writing tasks to test the effects of genre and task. Such methodological changes would allow for falsification studies that could provide additional evidence in relation to syntactic development and its effects on human judgments of proficiency.

Acknowledgments

This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES.

Appendix A

A.1. First essay for participant 15

I am going to tell about my family. My family consists of my parents, my dog and me. I am only child; I don't have any brother or sister. My father is a politician. My mother is an English teacher. My dog is a kind of traditional Japanese dog.

My father works hard. I guess he loves his job. He has many kind of job and his is always busy. In addition, he lives in another house which is located near my house. Therefore, I rarely meet him; I meet him once or twice a month. Because of that, I am used to spend my time and eat meals without my farther. However, when he makes his vacation, he always takes me foreign countries.

My mother was an English teacher. She taught many junior and high school students in my house. However, now, she passed exam of a local University, so she stopped to teach English. I think she is great! I didn't know when she was studying. She is going to go the University from this April.

I have a dog. His name is ''little''. I named him when he was puppy. He was so tiny and small that time, but now he grow up and become a big dog. He is sixteen years old. He is an old dog. When I was kindergarten student, he came my house. Therefore, I live with him for 15 years. Because of that, I became to like dogs. He is very cute and smart, so I love him.

These are my family. I like my family, and I miss them. Moreover I could realize how family is important and they support me.

A.2. Third essay for participant 15

The high school that you have attended is still as same system as you were high school student? Most school have long history and keep tradition; however a few years ago, my high school decided to make a change started to ready to it and my high school changed system and uniform dramatically after I graduated the school. The big change of my high school make people confuse, but it brought some benefits.

The biggest change is how to take and chose classes. Before the change, all students had to take same subjects and class, and they didn't have choice. However, after the change, students can choose classes, based on what kind of University does he want to enter, and they don't have to take all subject and they can make their own schedule of classes like college student.

The second change is student's uniform. Before the change, students wear the kind of traditional uniform and other high school students also wear the same uniform. However, after the change, my high school made their own uniform and the uniform is totally different design from the past uniform. Most current student and the people who graduated the high school objected against the change, but today most people agree with the change of uniform and people think the high school can emphasize the character of the renewed high-school in other way.

In conclusion, my high school made the biggest change in the high school history such as change in system of taking class and changing the traditional uniform into the newly designed uniform. These changes are made some confusing, but it made some benefit. This big change will work with the future of the high school positively and make new traditions.

References

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

Biber, D. (1985). Investigating macroscopic textual variation through multi-feature/multi-dimensional analyses. *Linguistics, 23*, 337–360. http://dx.doi.org/10.1515/ling.1985.23.2.337

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language, 62*, 384–414. http://dx.doi.org/10.2307/414678

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

Biber, D. (1992). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes, 15*, 133–163.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

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

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

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics* (pp. 132–139). San Francisco: Morgan Kauffmann.

Connor-Linton, J., & Polio, C. (2014). Comparing perspectives on L2 writing: Multiple analyses of a common corpus. *Journal of Second Language Writing, 26*, 1–9.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly, 42*(3), 475–493.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading, 35*(2), 115–135.

Crossley, S. A., Roscoe, R. D., McNamara, D. S., & Graesser, A. (2011). Predicting human scores of essay quality using computational indices of linguistic and textual features. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), In *Proceedings of the 15th international conference on artificial intelligence in education* (pp. 438–440). New York: Springer.

Crossley, S. A., Weston, J., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28*, 282–311.

Eckman, F., Moravcsik, E., & Wirth, J. (1989). Implicational universals and interrogative structures in the interlanguage of ESL learners. *Language Learning, 39*, 173–205.

Fang, Z., Schleppegrell, M. J., & Cox, B. (2006). Understanding the language demands of schooling: Nouns in academic registers. *Journal of Literacy Research, 38*, 247–273. http://dx.doi.org/10.1207/s15548430jlr3803_1

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

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*, 299–323.

Gass, S. (1979). Language transfer and universal grammatical relations. *Language Learning, 29*, 327–344.

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

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing, 9*, 123–145.

Halliday, M. A. K. (1989). *Spoken and written language*. Oxford, England: Oxford University Press.

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

Hawkins, R. (2001). *Second language syntax: A generative introduction*. Oxford, UK: Wiley-Blackwell.

Hempelmann, C. F., Rus, V., Graesser, A. C., & McNamara, D. S. (2006). Evaluating state-of-the-art treebank-style parsers for Coh-Metrix and other learning technology environments. *Natural Language Engineering, 12*, 131–144.

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. http://dx.doi.org/10.2307/329438

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

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

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing, 4*, 51–69.

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

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*(1), 36–62.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57–86. http://dx.doi.org/10.1177/0741088309351547

McNamara, D., & Graesser, A. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution*. Hershey, PA: IGI Global.

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

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. http://dx.doi.org/10.1093/applin/24.4.492

Reppen, R. (1994). *Variation in elementary student language: A multi-dimensional perspective.* (Unpublished doctoral dissertation) Flagstaff: Northern Arizona University.
Schumann, J. H. (1979). The acquisition of English negation by speakers of Spanish: A review of the literature. In R. W. Andersen (Ed.), *The acquisition and use of Spanish and English as first and second languages* (pp. 3–32). Washington, DC: TESOL.
Selinker, L., Swain, M., & Dumas, G. (1975). The interlanguage hypothesis extended to children. *Language Learning, 25*, 139–152.
Stockwell, G., & Harrington, M. (2003). The incidental development of L2 proficiency in NS-NNS email interactions. *CALICO Journal, 20*, 337–359.
Street, J. H. (1971). *Readability of UCLA materials used for foreign students.* (Unpublished M.A. thesis) Los Angeles: University of California.
Wells, R. (1960). Nominal and verbal style. In T. A. Sebeok (Ed.), *Style in language* (pp. 213–220). Cambridge, MA: MIT Press.
White, L. (1996). Clitics in child L2 French. In H. Clahsen (Ed.), *Generative perspectives on language acquisition: Empirical findings, theoretical considerations and crosslinguistic comparisons* (pp. 335–368). Amsterdam: John Benjamins.
Witten, I. A., Frank, E., & Hall, M. A. (2011). *Data mining*. San Francisco, CA: Elsevier.
Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity.* Honolulu, HI: University of Hawaii Press.

Scott Crossley is an Associate Professor at Georgia State University. His work involves the development, design, and application of natural language processing in language assessment. His current interests include computational linguistics, corpus linguistics, discourse processing, and discourse analysis.

Danielle McNamara is a Professor at Arizona State University. Her work involves the theoretical study of cognitive processes as well as the application of cognitive principles to educational practice. Her current research ranges a variety of topics including text comprehension, writing strategies, building tutoring technologies, and developing natural language algorithms.