Received: 2 May 2017 | Revised: 7 March 2018 | Accepted: 16 May 2018

DOI: 10.1111/ijal.12225

WILEY

SPECIAL ISSUE ARTICLE

Finding variation: assessing the development of syntactic complexity in ESL Speech

Mary Lou Vercellotti

Ball State University, Muncie, IN, USA

Correspondence
Mary Lou Vercellotti, Ball State University, Department of English, 297 Robert Bell Building, Muncie, IN 47306
Email: mlvercellott@bsu.edu

This paper examines the development and variation of syntactic complexity in the speech of 66 L2 learners over three academic semesters in an intensive English program. This investigation tracked development using hierarchical linear modeling with three commonly-used, recommended measures of productive complexity (i.e., length of AS-unit, clause length, subordination) and three exploratory measures of structural complexity (i.e., syntactic variety, weighted complexity scores, frequency of nonfinite clauses) to capture different aspects of syntactic complexity. All measures showed growth over time, suggesting that learners are not forced to prioritize certain aspects of the construct at the expense of others (i.e., no trade-off effects) across development. The unexplained significant variation found in these data differed among the measures reinforcing notions of multidimensionality of linguistic complexity.

KEYWORDS

L2 development, longitudinal, oral data, structural variety, trade-off effects

本文以66名英语二语学习者为对象,分析他们在三个学期强化学习期间的英语口语表达,对他们在英语口语中句法复杂性的发展和差异进行研究。研究采用分层线性模型进行跟踪调查,结合三种常用的产出复杂性推荐测量方法 (即单位长度、子句长度、主从关系),以及三种结构复杂性的探索性测量方法 (即句法多样化、加权复杂性评分、非限定小句的使用频率)。结果发现,随着时间的推移,学习者在以上各种复杂性句式表达方面均有所提高,并没有因为要追求某一复杂句式表达而忽略其它方面 (即没有权衡效果)。数据中关于学习

者在复杂句式表达方面的某些难以解释的重要差异,也进一步印证了语言复杂性的多维概念。.

关键词

结构多样化, 二语发展, 纵向, 口语语料, 权衡效果

1 | INTRODUCTION

Researchers studying second language (L2) learning have described language performance in terms of complexity, accuracy, and fluency. Of these, the definition of complexity seems to be most intertwined with language development. For instance, complex language has been described in L2 research as "more advanced" (e.g., Skehan, 2009). Operationalizing "more advanced" language, however, has been challenging, in part, because complexity has multiple meanings (Pallotti, 2009). Perhaps consequently, L2 researchers have chosen a variety of measures to study complexity in learner language performance, resulting in a patchwork of findings on the construct of syntactic complexity. Norris and Ortega (2009) reviewed measures of complexity in L2 research and proposed that complexity be assessed with multiple, complementary measures to fully capture the construct. In fact, both length and type of grammar construction are relevant to language development (Ellis, 1996), but most often, language researchers have used measures of length to represent complexity. Despite the interest in measuring syntactic complexity with additional measures (beyond length-based measures), few studies have included such measures. Further, Verspoor and Behrens (2011) have called for more studies looking of interrelated, overlapping, phenomena in order to better understand language development. By using multiple measures of the same construct, variation between measures can be explored in order to better understand syntactic complexity. Yet, few L2 studies have tracked linguistic complexity in learners' speech with multiple measures with longitudinal data. More research is needed to understand variation found in language performances, especially in speech, because the syntactic complexity found in oral language performance (with less planning time) more directly reflects the process of constructing complex language. This research investigates the development of syntactic complexity in the speech of English language learners using multiple measures of syntactic complexity, both length-based and structure-based. Even though all measures captured the growth of complexity across development, the measures revealed differences in remaining variation between learners' performances, which has theoretical insights and practical applications.

2 | BACKGROUND

2.1 | Complexity in oral production

Many factors influence language performance, such as task effects (e.g., Ferrari, 2012), limited cognitive resources (Ortega, 2009), dialogic communication (Ferrari, 2012), and modality (Biber, Grey, & Poonpon, 2011; Lambert & Kormos, 2014). Interactions between these language performance factors and the measure of complexity chosen by L2 researchers may cloud conclusions of the usefulness of the measure and implications from the findings. Research limited to data from a single task, however, avoids the variability which complicates identifying useful measures.

Production modality (written or spoken) in particular influences syntactic complexity because speakers must plan and monitor their language production simultaneously, and the additional cognitive load might hinder or delay learners from producing complex oral language (Trebits, 2014). The ephemeral nature of speech (for both the speaker and the listener) may also discourage integration of ideas into multi-clause utterances in speech (Schiffrin, 2014). In

other words, simple one-clause utterances might be preferred in oral communication, which would result in less subordination in speech than in texts. In contrast, Biber et al. (2011) have claimed that speech has more clausal subordination while written language has more complex phrases. Oral language performance certainly does not have the luxuries of written language (e.g., visual presence and time) to allow post-production review, revision, and potentially expansion of the basic message with the addition of modifying words (e.g., adjective and adverb phrases). Given all the possible interactions between modality and complexity measure, conclusions drawn about syntactic complexity based on written texts has limited generalizability for understanding syntactic complexity in oral language performance. And, although more research has been done with written data, speech more closely reflects the cognitive demands of L2 production, so oral language performance is more relevant for understanding the development of linguistic complexity. This paper will then focus solely on oral language production. Research with multiple measures could clarify how and how much L2 speakers complexify their speech, while also giving guidance in how to operationalize the construct of syntactic complexity in speech.

2.2 | Measuring syntactic complexity

Ellis (1996: 100) has stated that "(w)ith increasing competence … mean length of utterance and structural complexity increases." Fittingly, both utterance length and structures have occasionally been used to operationalize complexity (Arnold, Losongeo, Wasow, & Ginstrom, 2000). Complexity measured with length has been described as productive complexity (Foster, Tonkyn, & Wigglesworth, 2000), and productivity has been used extensively in first language (L1) research since Brown (1973) introduced mean length of utterance to measure language development in children, under the assumption that longer utterances are generally more complex than shorter ones. Empirical results with L1 data have supported this assumption. For instance, Nippold, Hesketh, Duthie, and Mansfield (2005) found that mean length of utterance captured differences in performance by age group. L2 researchers have adopted productivity measures for child and adult learners. One advantage of productivity measures is their practicality; length can easily be calculated. General complexity has often been measured as the length in words per base unit, usually a sentential unit (Norris & Ortega, 2009). This productive complexity measure captures lengthening at any level—phrasal or clausal.

Since sentence-level length measures cannot differentiate the source of the lengthening, such measures are "crude" (Rimmer, 2006), and another measure is needed to isolate increased complexity of non-clausal features, such as noun phrase expansion. Accordingly, phrasal complexity has been recommended as a complementary measure, which can be calculated as the mean clause length (Norris & Ortega, 2009). Despite the theoretical justification for phrasal complexity measures, some L2 researchers (e.g., De Clercq & Housen, 2017), have questioned its usefulness in oral data because, in part, phrasal expansion is challenging during oral production.

Measures of clause embedding are often considered a requirement of syntactic complexity because utterances with multiple clauses are considered more complex than single clause utterances, regardless of length. Thus, to capture embedding, a subordination measure (e.g., clause/sentential unit) has frequently been used to measure complexity in L2 research (Foster et al., 2000; Lambert & Kormos, 2014; Norris & Ortega, 2009; Tonkyn, 2012). In summary, the use of three complementary productivity measures quantitatively captures complexity by calculating the lengthening of utterances.

Despite the frequent use of productive complexity measures, reliance solely on length measures has been questioned. First, length-based measures may be useful early development but fail to capture development when length of unit plateaus (Bulté & Housen, 2012). Second, as both Ortega (2003) and Pallotti (2009) cautioned, "more" does not necessarily equate to more complex or even better language. Adult L2 learners may have the cognitive resources to create long utterances without structural complexity (Foster et al., 2000). Third, utterances of equal length or subordination may be structurally different. With these limitations of productivity measures of syntactic complexity, structural measures are needed, but it is unclear what measure(s) might be useful and practical.

Foster et al. (2000) offered "wide repertoire" as a second aspect of syntactic complexity. A wide repertoire could be captured with a measure of syntactic variety. A variety measure tallying all syntactic structures, however, is not feasible, and measures of structural variety have rarely been included in L2 research (De Clercq & Housen, 2017). The field has, however, attempted various exploratory measures of variety. For instance, De Clercq and Housen (2017: 322) created a measure based on the standard deviation of length, with the assumption that length is a "proxy of underlying structural differences." Given the main criticism of length measures, a variety measure based on length is also questionable. Calculating variety measures based on L1 theoretical abstractions is also problematic, with limited application to L2 language-learning theory, because some L2 frameworks question that an abstract grammar exists (Verspoor & Behrens, 2011) and/or that L2 learners have the same abstract categories and parameters as L1 speakers (e.g., White, 2000). One possibility to measure syntactic variety would be to identify a specific salient surface construction. For instance, Yuan and Ellis (2003) measured syntactic variety in English L2 speech with the verb form combinations. This measure has been criticized because verb form variety may not represent general syntactic variety (De Clercq & Housen, 2017). To capture the variety of structures, De Clercq and Housen (2017) introduced a measure with a sampling methodology which requires a minimum number of sentential units in the sample; this requirement, though, makes such methodology prohibitive with language performance data from lower proficiency learners.

A potential categorization for data from any proficiency level would be to label utterances as simple, compound, complex, compound-complex, based on written language conventions, (e.g., Spoelman & Verspoor, 2010). Although these labels capture some structural variety, they are still too coarse because a complex sentence, for example, could be created by a number of different clause types. Therefore, a study of structural complexity must use a finer-grain measure. The variety of clause types could give a full description of syntactic complexity (Rimmer, 2006). Both L1 (e.g., Nippold et al., 2005) and L2 researchers (e.g., Vercellotti & Packer, 2016) have coded speech by clause type (e.g., independent, adverbial, relative), so a wide repertoire of clause types is a possible useful measure.

Variety measures, too, may be criticized for ignoring the complexity differences among clause types (Lambert & Kormos, 2014), so a measure which includes a consideration of the relative complexity of each clause type might be useful. In fact, L2 researchers have assumed that language development means producing "more complex, less frequent and less similar" constructions (Verspoor & Behrens, 2011: 38). Developmental-based complexity measures have been created for L1 development (e.g., Covington, He, Brown, Naçi, & Brown, 2006), but L2 developmental metrics should be based on L2 data. Indeed, Norris and Ortega (2009: 574) called for "developmentally sensitive and interlanguage-based measures that tap complexity defined as structural variety, sophistication, and acquisitional timing" for L2 learning research.

Without such a complexity measure, L2 researchers have included measures of specific constructions (Norris & Ortega, 2009) which are aligned with proficiency, under the assumption that the frequency of a carefully chosen construction adequately represents the construct for that population. For instance, Nippold et al. (2005) found that the frequency of relative clauses increased with age in L1 oral data. L2 researchers may identify a development-related structure to serve as a proxy measure to represent structural complexity, and Norris and Ortega (2009) called for more research using specific constructions.

In summary, complementary measures of general complexity, phrasal complexity, and subordination have been recommended (Norris & Ortega, 2009) to capture productive complexity. Language learning research has largely relied on productivity measures despite the differing theoretical implications in the distinction between productive (length) and structural complexity. Therefore, measures of structural complexity must be explored to identify which capture the variety and/or the complexity of syntactic constructions, unrelated to length. A measure of clause variety, a weighted measure of difficulty, and a specific construction aligned with development seem to be most promising.

2.3 | Measuring complexity across development

Language performance can change in various linguistics aspects across development. For instance, Tonkyn (2012) found that L2 syntactic complexity significantly increased in multiple syntactic features (e.g., subordinate clauses,

adverbs). Language-learning researchers may suppose that any measure of complexity should increase with proficiency and that higher scores are better, but that assumption has been questioned (e.g., Pallotti, 2009) because excessively long utterances or utterances with excessive embedding hinder communication, especially oral communication (Schiffrin, 2014). Also, aspects of syntactic complexity might develop differently; some constructions may appear and decline with proficiency. Theoretically, patterns within development might be driven by limited cognitive resources during language learning, particularly during oral language. When learners develop in one aspect of language, fewer cognitive resources are available for development of another (Verspoor, Lowie, & van Dijk, 2008), which means that growth may wax and wane across development. Indeed, Ortega (2003) and Norris and Ortega (2009) outlined theory and research which suggest that learners complexify their language through first coordination, and then through subordination at intermediate proficiency, which would decline at higher proficiency when phrasal complexity increases. In order to be useful for language-learning studies, measures of complexity must capture change across stages of L2 development (Bulté & Housen, 2012).

Empirical research investigating the syntactic complexity in L2 speech has been mixed. Using cross-sectional data, De Clercq and Housen (2017) found an increase of subordination with increasing proficiency while Iwashita, Brown, McNamara, and O'Hagan (2008) concluded that subordination in English L2 speech did not increase linearly. Iwashita et al., however, was severely limited by extremely short language samples at the lowest proficiency level, which would have prevented statistically significant results between proficiency levels, and the authors explained that their data did indeed show more subordination at the higher levels. Based on their longitudinal case study, Polat and Kim (2014) concluded that general, phrasal, and subordination measures are all required to study syntactic complexity over time. Longitudinal research with larger data sets is required to identify useful measures of productive complexity across development.

Even less is known about the development of structural complexity. Some studies have found that certain clauses are produced earlier by ESL learners (Lambert & Kormos, 2014; Vercellotti & Packer, 2016), and are easier for ESL learners (e.g., Kazemi, 2011). Vercellotti and Packer (2016) reported that adult L2 learners in an intensive English program produced adverbial clauses early, followed by nonfinite clauses, while relative clauses and complement-taking predicate clauses emerged later. De Clercq and Housen's (2017) findings with adolescent English L2 learners, also found relative clauses were rarely produced at the earliest proficiency levels, but in their data, complement-taking predicate clauses were produced early and most frequently. Overall, recent research with cross-sectional data has shown that learners produce clause types at differing frequencies, but that largely each clause type increases (e.g., De Clercq & Housen, 2017; Vercellotti & Packer, 2016).

From taking snapshots of separate participant groups and calculating means, researchers (e.g., De Clercq & Housen, 2017; Nippold et al., 2005; Skehan & Foster, 1997) have drawn conclusions about how complexity develops. Conclusions based on cross-sectional data, however, can be misleading because means can be skewed by a subset of the participants, and different subsets might drive 'significant' differences at different time points. (Skehan, 2009). Besides, L2 cross-sectional designs often sort participants' proficiency by categorical variables (e.g., beginning, intermediate, advanced), ignoring the variation within each proficiency level, which may then obscure significant findings.

To investigate syntactic complexity and variation within development, repeated measures of a longitudinal study are necessary (Verspoor et al., 2008). Monologues avoid variation in performance in reaction to the interlocutor, which is especially important when investigating change over time. Longitudinal data are especially important when considering if there are any trade-off effects because trade-off effects must be found within individuals' performances. Existing longitudinal research with oral data, however, has been only singe case studies (e.g., Polat & Kim, 2014) or multiple case studies (e.g., Ferrari, 2012) which have limited generalizability. Research with longitudinal data from multiple learners with a continuous measure of proficiency can better investigate change over time. In addition, the measures should ideally capture development as well as variation, a well-known phenomenon in L2 learning.

Given the dearth of longitudinal oral performance studies, this paper investigates the development and variation of syntactic complexity in monologic speech of instructed English L2 learners using multiple, complementary measures of productivity and structural complexity in order to further explore the construct of complexity.

RQ 1:. Do the various measures of syntactic complexity

a. capture growth over time in the speech of instructed ESL learners and capture the expected variation from differences in initial proficiency?

b. reveal additional variation in these longitudinal data?

RQ 2:. What do the findings suggest about the development of complexity in ESL speech?

3 | METHOD

This descriptive study investigated the development of syntactic complexity in transcripts of monologues from an English L2 data set available at http://talkbank.org/access/SLABank/English/Vercellotti.html, which were produced during a classroom-based activity over a year in a pre-university Intensive English Program (IEP). This research focused on clausal structures and used multiple measures to capture any sub-constructs within the complex construct of syntactic complexity.

3.1 | Participants

Participants were ESL learners (n = 66) who entered an IEP in the United States in 2010. This IEP offered courses during three 13-week semesters per academic year at three proficiency levels: low-intermediate, high-intermediate, and low-advanced. Most students attended full-time (20 hours of instruction/week), being enrolled in speaking, listening, grammar, reading, and writing classes. Participants with enough data points for the longitudinal analysis (more than three data points or data from two semesters) were included in this study. The participants were adults aged 18–35 years (M = 25.3; SD = 4.5), both male (n = 34) and female (n = 32), from multiple L1 groups: Arabic (n = 43), Chinese (n = 16), and Korean (n = 7). At enrollment, students took written proficiency tests and were placed at the low-intermediate (n = 27) or high intermediate level (n = 39). For this study, one proficiency test score was chosen to represent initial proficiency so that initial proficiency was a continuous variable rather than categorical (i.e., low, high). The listening test (M = 19.3; SD = 4.5; range 9–27) was chosen as the measure of initial proficiency because listening requires phonological loop skills which are linked to language learning (Baddeley, Gathercole, & Papagno, 1998) and because this test's scores most highly correlated with placement level (r = 0.838; p < 0.001).

3.2 | Data

The speech transcripts were from two-minute, semi-spontaneous, topic-based monologues, after a one-minute planning without the use of notes or any support materials. The speaking task was part of the curriculum and was graded by the instructor. (See McCormick & Vercellotti, 2013 for a description of the activity.) The learners were not directed to attend to any linguistic focus, but in the context of an IEP speaking course, it is plausible that learners focused more on fluency and/or accuracy. Each monologue was on a different topic, chosen by the IEP instructors for pedagogical purposes. The monologues were given multiple times each semester, roughly one month apart within semesters but further apart across semesters. Not every participant remained in the IEP for the year. On average, there were 4.45 (SD = 1.25) speeches per participant, with a maximum of seven speeches.

For oral language data, researchers must first segment utterances into base units. Foster et al. (2000) defined AS-units as sentence-level utterances with minimally an independent clause with a finite verb and its dependent clauses. AS-units have been adopted as the base unit for oral language in L2 research (Norris & Ortega, 2009). These transcripts were coded into AS-units and clauses following Foster et al. (2000). In this coding, a nonfinite clause must have a complement or adjunct to be considered a clause, which at least partially addresses the concern about including nonfinite clauses in measures of complexity (e.g., Bulté & Housen, 2012). Each clause was coded as an

independent, coordinated verb, adverbial, relative, complement-taking predicate, or nonfinite clause. (See Vercellotti & Packer, 2016.) These data included ungrammatical attempts, like other L2 development research (e.g., Spinner, 2011).

3.3 | Data coding

Productive complexity was measured as mean length of AS-unit (general complexity), mean length of finite clause (phrasal complexity), and finite clauses per AS-unit (subordination), following the recommendations by Norris and Ortega (2009). These three length-based measures have been used in L2 research with oral data (e.g., De Clercq & Housen, 2017; Polat & Kim, 2014).

Three measures of structural complexity were included to explore whether such measures might find different patterns of development or variation than the length measures. To capture the notion of wide repertoire (Foster et al., 2000), each speech was coded for syntactic variety based on the number of different clause types within it. A speech consisting of only independent clauses was coded as 0.167 because it has only one of the six clause types (1/6) while a speech with at least one token of all six clause types (6/6) was coded as 1.0, reflecting the full repertoire of clause types. This variety score represents discourse-level complexity and was rooted in the assumption that speakers produce a variety of forms in a language sample but that each clause type is judiciously used. This exploratory measure, therefore, balanced the practical problem of using all structures while circumventing the "more is better" assumption, inherent in the calculation of productive complexity measures.

In response to Norris and Ortega's (2009) suggestion for measures of structural complexity which capture sophistication and acquisition, a weighted structural complexity scale was devised. Language learners normally start with more simple and frequent constructions and proceed to more complex and infrequent constructions (Verspoor & Behrens, 2011), so the weighted complexity scale assigned more points to more complex and less frequent constructions. Each AS-unit in the data set was copied into an Excel file to be coded. Basic AS-units with only an independent main clause, the minimal structure, were scored as 0. AS-units with a coordinated verb phrase or an adverbial clause were scored as 1 because these clauses represent early complexification (Diessel, 2013). AS-units with a nonfinite clause, as the next early-produced clause type in adult L2 speech (Vercellotti & Packer, 2016), were scored as 2. AS-units with a relative clause or a complement-taking predicate clause, both more difficult expansion clauses (Diessel, 2013), were coded as 3. AS-units with more than one type of dependent clause received the highest score of 4. Coding utterances with multiple clause types as the most complex has been done in L1 language research (e.g., Covington et al., 2006). Table 1 summarizes the weighted structural complexity scale and offers examples from the data. This basic five-level scale, based on theory and empirical findings of adult L2 oral production, acknowledges the impact of function and frequency, which is relevant in studying the development of complexity (Verspoor & Behrens, 2011). A subset of the data was independently coded, and inter-coder agreement was calculated as 0.991. The scores of the AS-units within the speech were then averaged using the Excel function to calculate the weighted complexity score for each speech.

TABLE 1 Weighted structural complexity scale

	Structure	Example
0	Independent (main) finite clause	next time I can pay them back
1	Independent clause with compound verb phrase OR with an adverbial clause	then maybe we can share or share our food together
2	Independent clause with a nonfinite clause	he insisted to pay the bills
3	Independent clause and a relative clause OR a complement-taking predicate clause	that's the only thing that I don't like it
4	Combination of 2+ types of dependent clauses	if I don't like this man and I don't want to have a next date I think they pay the bill first

TABLE 2 Measures of syntactic complexity

Productivity complexity	Structural complexity
mean length (words) of AS-unit	syntactic variety
mean length (words) of finite clause	weighted syntactic complexity
mean number of finite clause/AS-unit	frequency of nonfinite clause

Finally, since frequency of nonfinite clauses has been found to be developmentally-aligned (e.g., De Clercq & Housen, 2017; Vercellotti & Packer, 2016), it was chosen as a proxy measure for comparison to the newly designed measures of syntactic variety and weighted structural complexity. Importantly, frequency of nonfinite clauses is complementary (not redundant) to the subordination measure (finite clause/AS-unit) in the study. Table 2 summarizes the six measures of syntactic complexity considered in this study. Each measure can be calculated for speeches of different lengths, necessary for measures of development.

3.4 | Data analysis

The data were analyzed using hierarchical linear and non-linear modeling (HLM). HLM analysis captures the change in performance considering variation in the number of observations and in the distance between observations (Singer & Willett, 2003). The use of this analysis in L2 research has many advantages (See Cunnings, 2012). In particular, the analysis determines whether there is remaining significant variance in the model, which can potentially be explained by predictor variables, such as initial proficiency. Initial proficiency of each participant was converted to distance from the mean to be tested as a predictor variable. The inclusion of initial proficiency in the model is relative to any difference from the population's mean in initial proficiency, which means that the model predicts a larger effect of initial proficiency for participants who scored farther from the population mean. With a focus of identifying generally useful measures of complexity, this study did not include any other independent variables. To construct a parsimonious best-fitting model, the random effects were constrained to zero in any model that lacked statistically significant variation. Such a finding in HLM models does not mean that performances were uniform, only that the variation was not systematic among the participants. For each model, the full maximum likelihood method, which reports the goodness-of-fit for all parameters in the model (Singer & Willett, 2003), was used, and robust standard errors figures are reported.

4 | RESULTS AND DISCUSSION

For each measure, the HLM model results are presented in the following order: expected value at the first observation, effect of initial proficiency, rate of change, and remaining (unexplained) significant variation in the model, if any. Table 3 summarizes the conditioned linear HLM model of each measure.

4.1 | Productive complexity

The results of the general measure of productive complexity (words/AS-unit) found that the expected mean length of AS-unit at the first observation for participants with average initial proficiency was estimated to be 9.90 words. For each extra point of initial proficiency (distance from population mean), the mean number of words increased by 0.24. The mean linear growth rate was estimated to be 7.53, indicating a significantly positive average rate of increase in length of AS-unit over time. The mean intercept, initial proficiency, and growth rate were statistically significant ($p < 0.001$), indicating that all parameters are necessary for describing the mean growth trajectory. After adding initial proficiency as a predictor for initial scores, participants no longer varied significantly in the length of AS-unit at the first observation ($\chi^2 = 77.92$, $p = 0.113$), but participants did vary significantly in growth rate ($\chi^2 = 99.51$, $p = 0.004$).

TABLE 3 Conditioned linear growth model for each measure

	Fixed effects				Random effects			
	Coefficient	SE	t	p	Variance Component	df	χ^2	p
Length of AS-unit (in words)								
Mean score at 1st observation, β_{00}	9.90	0.21	46.76	< 0.001	0.647	64	77.92	0.113
Initial proficiency, β_{01}	0.024	0.04	6.73	< 0.001				
Mean growth rate, β_{10}	7.53	1.01	7.10	< 0.001	20.05	65	99.51	0.004
Length of Finite Clause (in words)								
Mean score at 1st observation, β_{00}	6.72	0.09	74.67	< 0.001	0.153	64	112.65	< 0.001
Initial proficiency, β_{01}	0.06	0.02	3.70	< 0.001				
Mean growth rate, β_{10}	1.93	0.35	5.45	< 0.001				
Finite clauses per AS-unit								
Mean score at 1st observation, β_{00}	1.50	0.028	54.08	< 0.001	0.003	64	70.29	0.275
Initial proficiency, β_{01}	0.02	0.006	3.54	< 0.001				
Mean growth rate, β_{10}	0.56	0.149	3.73	< 0.001				
Syntactic variety								
Mean score at 1st observation, β_{00}	0.729	0.016	45.46	< 0.001	0.003	64	97.51	0.005
Initial proficiency, β_{01}	0.014	0.003	4.617	< 0.001				
Mean growth rate, β_{10}	0.175	0.064	4.617	.007				
Weighted structural complexity								
Mean score at 1st observation, β_{00}	1.74	0.17	10.42	< 0.001	0.672	64	95.16	0.007
Initial proficiency, β_{01}	0.12	0.03	4.60	< .001				
Mean growth rate, β_{10}	1.73	0.58	2.98	0.004	3.05	65	86.92	0.036
Frequency of nonfinite clauses								
Mean score at 1st observation, β_{00}	2.94	0.21	13.48	< 0.001	0.315	64	81.18	0.072
Initial proficiency, β_{01}	0.21	0.06	5.90	< 0.001				
Mean growth rate, β_{10}	2.89	0.84	3.46	< 0.001				

This result suggests that AS-unit length does capture change in complexity across this proficiency range, which has been found with general measures of complexity in cross-sectional oral performance research (e.g., De Clercq & Housen, 2017; Iwashita et al., 2008). There was about a 70% increase in general complexity in these data. Variation in initial scores was explained by differences in initial proficiency, but variation remained in the change rate model, which means that the participants' performance diverged over time. Initial proficiency, however, was not a significant predictor for growth rate, so the variation in performance over time was not driven by differences in initial proficiency.

Phrasal complexity was measured as mean length of finite clause. The expected clause length at initial observation was estimated to be 6.72 words. With each additional point of initial proficiency, clause length increased by 0.06. The mean linear growth rate was estimated to be 1.93 words, a significantly positive average rate of increase. The mean intercept, initial proficiency, and growth rate were statistically significant, indicating that all parameters are necessary for describing the mean growth trajectory. After controlling for initial proficiency, participants still varied significantly in mean clause length at the first observation ($\chi^2 = 112.65$, $p < 0.001$).

Phrasal complexity increased over time, which is contrary to De Clercq and Housen's (2017) cross-sectional data results where clause length did not discriminate proficiency levels. Two central factors may explain this difference in results. First, De Clercq and Housen included nonfinite clauses in their phrasal complexity measure. Nonfinite clauses tend to be shorter than finite clauses, and when both the frequency of (shorter) nonfinite clauses and length of finite clauses increase, those developmental changes will mathematically off-set. Second, in the current study, the model

for phrasal complexity showed remaining variation in initial score even after controlling for initial proficiency. Hence, an additional explanation for the conflicting results may be a result of the variation in the proficiency of the learners. With De Clercq and Housen's cross-sectional data, the range of scores within their proficiency groups could obscure differences between the groups.

In the current study, phrasal complexity increased approximately 30%, but there was no significant variation in the slope, or change rate for phrasal complexity. The lack of variation in growth rate may be explained by clause length constraints in oral language (e.g., Chafe, 1988), in that learners must balance increasing syntactic complexity with conveying meaning without the message becoming lost on the listener (and speaker).

For subordination, the expected number of finite clauses/AS-unit at initial observation was estimated to be 1.50. For each point of initial proficiency, subordination increased by in 0.02. The mean linear growth rate of increase was estimated to be 0.56, a significantly positive average rate of increase. All parameters were necessary for describing the mean growth trajectory ($p < 0.001$). Participants did not vary significantly in subordination at the first observation ($\chi^2 = 70.29$, $p = 0.275$) after controlling for initial proficiency.

This finding of growth over time for subordination is consistent with De Clercq and Housen (2017) and Tonkyn (2012), and the broad trends reported in Iwashita et al. (2008). Despite the ephemeral nature of speech which may inhibit subordination (Schiffrin, 2014), these learners continually added more finite clauses across development. This (more coarse) measure had no significant variation at initial proficiency (after controlling for initial proficiency) nor in change rate.

4.2 | Structural complexity

Syntactic variety was calculated as the number of different clause types in the speech divided by the total number of possible clause types. Results of the best-fitting growth model for syntactic variety found that the expected syntactic variety score at initial observation was estimated to be 0.729. For each point of initial proficiency, the mean syntactic variety score increased by 0.014. The mean linear change rate was estimated to be 0.175, indicating a significantly positive average rate of increase in syntactic variety over time. The mean intercept, initial proficiency, and growth rate were statistically significant, indicating that all parameters are necessary for describing the mean change trajectory. Participants still varied significantly in syntactic variety at the first observation ($\chi^2 = 97.51$, $p = 0.005$) after controlling for initial proficiency.

The expected score at the initial observation for the average student was 72.9% of the clause types, suggesting that many speeches at the initial observation included at least four of the clause types. But, there was significant variation among the participants' variety scores at the first observation. This measure of discourse-level syntactic variety reached ceiling at later observations, evident in the raw data (which showed many scores of 1.0).

The second structural complexity measure captured complexity with consideration of the developmental order. The expected weighted complexity score at initial observation was estimated to be 1.74. For each point increase of initial proficiency, there was a corresponding 0.12 increase in weighted complexity score. The mean linear growth rate of increase was estimated to be 1.73, indicating a significantly positive average rate of increase over time. The mean intercept, initial proficiency, and growth rate were statistically significant; all parameters are necessary for describing the mean growth trajectory. Participants still varied significantly in complexity score at the first observation ($\chi^2 = 95.16$, $p = 0.007$) and in growth rate ($\chi^2 = 86.92$, $p = 0.036$).

The expected weighted complexity score at the initial observation indicates that for the average student, the speech included some AS-units with dependent clauses, and the growth rate indicated that the scores doubled, rising to nearly 3.5 after a year. The model for this measure of complexity had additional unexplained variation both in initial scores (even after controlling for initial proficiency) and in growth rate, which means that there was variation in the language performances and variation in the development among individuals.

For the third measure of structural complexity, the frequency of nonfinite clauses, the results of its best-fitting growth model found that the expected number of nonfinite clauses for students with average initial proficiency

was estimated to be 2.94 at the initial observation. For each point of initial proficiency, the mean number of nonfinite clauses increased by 0.21. The mean linear growth rate of increase was estimated to be 2.89, indicating a significantly positive average rate of increase in production of nonfinite clauses over time. The mean intercept, initial proficiency, and growth rate were statistically significant ($p < 0.001$). After controlling for initial proficiency, participants no longer varied significantly in the number of nonfinite clauses at the first observation ($\chi^2 = 81.18$, $p = 0.072$).

This longitudinal analysis confirms the nonfinite clause findings of previous studies (e.g., De Clercq & Housen, 2017; Vercellotti & Packer, 2016) based on means at different stages during instruction. These data showed a 100% increase in the number of nonfinite clauses in the speech samples. The result for this proxy measure was similar to the other measures of structural complexity used in this study in that all three measures found that initial proficiency contributed to initial scores, and all showed growth over time.

4.3 | General discussion and implications

Each measure of syntactic complexity was useful for capturing language development since the best-fitting model for each measure showed a significant and meaningful increase in complexity in the speeches over time. Complexity, particularly in speech, is expected to be retarded by limitations in cognitive load (Trebits, 2014); these results suggest that with increasing proficiency, cognitive resources are available for complexifying the language performance. There was no decrease or plateau in any type of complexity (such as subordination) in these data, echoing De Clercq and Housen's (2017) findings, despite expectations about the development of syntactic complexity (e.g., Lambert & Kormos, 2014; Norris & Ortega, 2009).

Initial proficiency was a significant predictor of initial scores for all measures, which validates each measure's ability to capture variation from the most expected and predicable individual difference in language learning. Some models, however, revealed more variation among the language performances (see Table 4). Additional variation between participants remained at the initial observation for phrasal complexity, syntactic variety, and weighted structural complexity. Unexplained variation early in development is expected in dynamic systems theory (Verspoor & Behrens, 2011). (See van Geert, 2008 for an introduction to dynamic systems theory and L2 development.) Additionally, individual differences based on L1 or individual style may be more pronounced at the beginning of L2 development. These three measures seem to be able to capture variation in language performance early in development, indicating that these measures, therefore, might be recommended for research investigating such variation.

Significant variation in growth trajectories (change rate) between participants was only found in the models for general complexity and weighted structural complexity. These measures might capture how learners diverge as development progresses, perhaps from known variables, such as motivation (Verspoor & Behrens, 2011) or linguistic style (Pallotti, 2009). Accordingly, general complexity and weighted structural complexity measures might be recommended for research interested in variation among learners across development or L2 learning outcomes.

Conversely, several models (i.e., phrasal complexity, subordination, syntactic variety, frequency of nonfinite clauses) did not have statistically significant variation even to include random effects in change rate model, which means that the differences among the performances were not systematic, perhaps indicating that these measures

TABLE 4 Summary of change over time by complexity measure

	Measure	Change trajectory	Variation explained by initial proficiency	Remaining variation
Productive	General–length of AS-unit	Linear increase	Initial scores	Change rate
	Phrasal -length of finite clause	Linear increase	Initial scores	Initial scores
	Subordination –finite clause/AS-unit	Linear increase	Initial scores	--
Structural	Syntactic variety	Linear increase	Initial scores	Initial scores
	Weighted structural Complexity	Linear increase	Initial scores	Initial scores Change rate
	Frequency of nonfinite clauses	Linear increase	Initial scores	--

reveal intra-individual differences rather than systematic effects. As Pallotti (2009) has stated, results showing variation and lack of variation are both valuable to our understanding.

The models *within* each type of syntactic complexity (productive and structural), revealed differences in the remaining variation (or lack thereof) in the data. The commonly-used length measures are expected to measure syntactic complexity at distinct linguistic levels (Norris & Ortega, 2009), and the differences between them support their use as complementary measures. The three structural measures in this study were mostly exploratory, and the remaining variation differences between those models may indicate that they, too, capture different aspects of structural complexity. Admittedly, using multiple measures of syntactic complexity may be impractical in some language learning research, such as those investigating complexity, accuracy, and fluency. Since the subordination measure captured growth and initial variation by initial proficiency without any remaining statistically significant variation, subordination might perform well as a broad measure of complexity in language performance studies. Yet, as a broad measure of a specific type of complexity, a subordination measure may be limited to instructed learners (where this type of complexity is explicitly taught) or to studies which expect larger changes in proficiency.

There were two interesting similarities across these two types of syntactic complexity. First, the model for phrasal complexity and the model for the syntactic variety produced similar results, with the measure of initial proficiency explaining some but not all of the variation in initial scores, and no systematic variation in the change rate of growth. The calculation of these measures were unrelated (mean length of finite clause and percentage of clause types) and thus captured unconnected aspects of syntactic complexity. The shared pattern, if meaningful, might reflect a higher-order construct, such as proficiency (and/or willingness for risk-taking), where both measures of complexity steadily increase with language development.

Second, the subordination measure offered similar results as the frequency of nonfinite clauses; both increased linearly with initial proficiency explaining variation at initial observation with no other significant variation. Since interactions between measures which have a meaningful relationship are relevant for development and its variability (Verspoor et al., 2008), this finding is extremely valuable because these measures capture two different types of subordination—subordination of finite clauses and subordination of nonfinite clauses. These types of subordination could plausibly be in competition during production, but these measures capturing distinct aspects of complexity both showed growth. Unlike cross-sectional data where high end-points in two measures could be explained as balancing out with proficiency, the longitudinal analysis modeled simultaneous growth in both measures. L2 studies which have concluded evidence of trade-off effects have been overwhelmingly with static data, and those findings should only be extrapolated to language development with caution (Vercellotti, 2015).

In fact, there was no evidence of trade-off effects within the construct of complexity in any of models of syntactic complexity across development. This finding supports a 'connected' and 'supportive' rather than competitive theory of development (Spoelman & Verspoor, 2010). The results support theories (e.g., dynamic systems theory) which view language as a complex, interrelated system where development does not necessarily hinder growth in another (de Bot, 2008), even closely related subsystems, such as within the construct of syntactic complexity as examined in this study.

5 | CONCLUSION

This paper describes the development of syntactic complexity, using productive (length) and structural measures, in the speech of adult instructed ESL learners. Employing commonly-used measures (i.e., length of AS-unit, clause length, subordination,) as well as designed measures of structural complexity (i.e., syntactic variety, weighted complexity) and a proxy measure (i.e., frequency of nonfinite clauses), this research sheds new light on the construct of complexity in English L2 oral language performance across development. The data showed that all measures increased over time informing theoretical implications about the development of the construct of complexity. Given that each measure, with each capturing distinct aspects of complexity, showed growth over time, there is no

evidence of trade-off effects within the construct of complexity in these data. It is also relevant to note that there was no trade-off between complexity by adding finite clauses and complexity by adding nonfinite clauses in these data; both increased with increasing proficiency. As this research revealed remaining variation in different parts of the models, the results also confirm that complexity is a multifaceted construct.

A number of limitations must be noted before continuing. These results reflect the performance of one population of ESL learners in one IEP during three consecutive semesters. These learners' proficiency ranged from low-intermediate to low-advanced, so any predicted decrease or plateau in subordination could possibly still occur beyond the range of this study. The results of this longitudinal study should be supplemented with similar research with less proficient and more advanced students to advance the understanding of the development of syntactic complexity in ESL oral performance. This research used existing learner data, with some factors (e.g., prompt, instructions) driven by pedagogy; this research could be replicated while controlling for classroom-based variables, such as prompt. Further, these results reflect the development of syntactic complexity only of ESL oral monologues elicited with topic prompts. A different pattern might be found with other task types, particularly interactive tasks, where increased complexity is less appropriate (Ferrari, 2012).

The results can inform the measurement choices and methodology for future English L2 research. As would be expected with language learning performance, there was substantial variation. L2 researchers likely want to use practical measures that capture the variation between individuals and across development. The variation in different parts of the measure's models suggest that the measures capture separate aspects of complexity, and some suggestions can be offered. Subordination may serve as a practical, broad measure of complexity in instructed contexts. The easily calculated phrasal complexity revealed variation early in development, as did the weighted structural complexity measure. Moreover, researchers may want to consider using the weighted complexity measure for research investigating individual differences in language performance. One possibility is to create a measure based on standard deviation (e.g., De Clercq & Housen, 2017) of the weighted complexity measure, if the study's purpose is to measure the variety of structural complexity in the language sample, rather than the growth of the developmentally-aligned structural complexity. When investigating differences in language learning outcomes, general complexity and the weighted structural complexity may be useful, given the additional variation found in the models. The unexplained significant remaining variation between individuals is fodder for future longitudinal research. For instance, future research might consider how production may be influenced by the frequency and function of constructions in learners' L1s, motivation (Verspoor & Behrens, 2011), or individual speaking style (Pallotti, 2009). Overall, this paper offers a unique comparison of syntactic complexity, both productive and structural complexity measures, advancing our understanding of this most complex construct of language performance.

ACKNOWLEDGEMENTS

Thanks to Bonnie D. Schwartz for inspiring this research by questioning my blind acceptance of length as a proxy for complexity, and to Yu (Daisy) Xiang for translating the abstract. Many thanks to the anonymous reviewers whose feedback greatly improved the manuscript. Any remaining imprecision or errors are entirely my own.

ORCID

Mary Lou Vercellotti http://orcid.org/0000-0002-8115-5711

REFERENCES

Arnold, J. E., Losongeo, A., Wasow, T., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, *76*(1), 28–55.

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*(1), 158–173.

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

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Bulté, B., & Housen, A. (2012). Defining and operationalizing L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency* (pp. 21–46). Philadelphia, PA: John Benjamins.

Chafe, W. (1988). Linking intonation units in spoken English. In J. Haiman, & S. A. Thompson (Eds.), *Clause combining in grammar and discourse* (pp. 1–27). Philadelphia, PA: John Benjamins.

Covington, M. A., He, C., Brown, C., Naçi, L. & Brown, J. (2006). How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-level Scale. CASPR Research Report 2006–01. The University of Georgia Artificial Intelligence Center.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, *28*, 368–382.

De Bot, K. (2008). Introduction: Second language development as a dynamic process. *Modern Language Journal*, *92*(2), 166–178.

De Clercq, B., & Housen, A. (2017). A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. *The Modern Language Journal*, *101*, 315–334. https://doi.org/10.1111/modl.12396

Diessel, H. (2013). Construction Grammar and first language acquisition. In T. Hoffmann, & G. Trousdale (Eds.), *The Oxford handbook of construction grammar* (pp. 347–364). New York: Oxford University Press.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, *18*, 91–126.

Ferrari, S. (2012). A longitudinal study of complexity, accuracy and fluency variation in second language development. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency* (pp. 277–297). Philadelphia, PA: John Benjamins Publishing Company.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*, 354–375.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*, 24–49.

Kazemi, A. (2011). An investigation into the relationship between the type of self-repair and structural complexity of utterance. *Journal of English and Literature*, *2*(4), 96–102.

Lambert, C., & Kormos, J. (2014). Complexity, accuracy, and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition. *Applied Linguistics*, *35*, 607–614. https://doi.org/10.1093/applin/amu047

McCormick, D. E., & Vercellotti, M. L. (2013). Examining the impact of self-correction notes on grammatical accuracy in speaking. *TESOL Quarterly*, *92*, 410–420.

Nippold, M. A., Hesketh, L. J., Duthie, J. K., & Mansfield, T. C. (2005). Conversational versus expository discourse: A study of syntactic development in children, adolescents, and adults. *Journal of Speech, Language, and Hearing Research*, *48*, 1048–1064.

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

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.

Ortega, L. (2009). *Understanding second language acquisition*. New York, NY: Routledge.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*, 590–601. https://doi.org/10.1093/applin/amp045

Polat, B., & Kim, Y. (2014). Dynamics of complexity and accuracy: A longitudinal case study of advanced untutored development. *Applied Linguistics*, *35*(2), 184–207.

Rimmer, W. (2006). Putting grammatical complexity in context. *Literacy*, *42*(1), 29–35.

Schiffrin, D. (2014). Discourse. In R. W. Fasold, & J. Connor-Linton (Eds.), *An Introduction to Languages and Linguistics* (pp. 169–203). Cambridge, UK: Cambridge University Press.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis*. New York, NY: Oxford University Press.

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

Skehan, P., & Foster, P. (1997). Task type and processing conditions as influences on foreign language performance. *Language Teaching Research*, *1*(3), 185–211.

Spinner, P. (2011). Second language assessment and morphosyntactic development. *Studies in Second Language Acquisition*, *33*, 529–561.

Spoelman, M., & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study of the acquisition of Finnish. *Applied Linguistics*, *31*, 523–533.

Tonkyn, A. (2012). Measuring and perceiving changes in oral complexity, accuracy and fluency: Examining instructed learners' short–term gains. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency* (pp. 221–245). Philadelphia, PA: John Benjamins.

Trebits, A. (2014). Sources of individual differences in L2 narrative production: The contribution of input, processing, and output anxiety. *Applied Linguistics*, *37*, 155–174.

Van Geert, P. (2008). The dynamic systems approach in the study of L1 and L2 acquisition: An introduction. *Modern Language Journal*, *92*, 179–199.

Vercellotti, M. L. (2015). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, *38*, 90–111. https://doi.org/10.1093/applin/amv002

Vercellotti, M. L., & Packer, J. (2016). Shifting structural complexity: The production of clause types in speeches given by English for academic purposes students. *Journal of English for Academic Purposes*, *22*, 179–190. https://doi.org/10.1016/j.jeap2016.04.004

Verspoor, M. H., & Behrens, H. (2011). Dynamic systems theory and a usage-based approach to Second Language Development. In M. H. Verspoor, K. de Bot, & W. Lowie (Eds.), *A dynamic approach to second language development* (pp. 25–38). Philadelphia, PA: John Benjamins.

Verspoor, M. H., Lowie, W., & van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *Modern Language Journal*, *92*, 214–231.

White, L. (2000). Second language acquisition: From initial to final state. In J. Archibald (Ed.), *Second language acquisition and linguistic theory* (pp. 130–155). Oxford: Blackwells.

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, *24*, 1–27.

How to cite this article: Vercellotti ML. Finding variation: assessing the development of syntactic complexity in ESL Speech. *Int J Appl Linguist*. 2018;1–15. https://doi.org/10.1111/ijal.12225