Studies in Second Language Acquisition, **43** (2021), 781–812
doi:10.1017/S0272263120000546

Research Article

MEASURING LONGITUDINAL WRITING DEVELOPMENT USING INDICES OF SYNTACTIC COMPLEXITY AND SOPHISTICATION

Kristopher Kyle*

University of Oregon and Yonsei University

Scott Crossley

Georgia State University

Marjolijn Verspoor

University of Groningen

Abstract

Measures of syntactic complexity such as mean length of T-unit have been common measures of language proficiency in studies of second language acquisition. Despite the ubiquity and usefulness of such structure-based measures, they could be complemented with measures based on usage-based theories, which focus on the development of not just syntactic forms but also form-meaning pairs, called constructions (Ellis, 2002). Recent cross-sectional research (Kyle & Crossley, 2017) has indicated that indices related to usage-based characteristics of verb argument construction (VAC) use may be better indicators of writing proficiency than structure-based indices of syntactic complexity. However, because cross-sectional studies can only show general trends across proficiency benchmarks, it is important to test these findings in individuals over time (Lowie & Verspoor, 2019). Thus, this study investigates the developmental trajectories of second language learners of English across two academic years with regard to syntactic complexity and VAC sophistication.

INTRODUCTION

Syntactic complexity has been an important measure of proficiency and development in second language acquisition (SLA) research in general and second language writing

We would like to thank the four anonymous reviewers and the handling editor, Andrea Revesz, for their detailed comments on previous versions of this manuscript.

* Correspondence concerning this article should be addressed to Kristopher Kyle, Straub Hall 161, 1290 University of Oregon, Eugene, Oregon 97403-1290. E-mail: kkyle2@uoregon.edu

(SLW) in particular over the past 45 years (Crymes, 1971; Larsen-Freeman, 1978; Lu, 2011; Ortega, 2003; Wolfe-Quintero et al., 1998). Although a variety of indices have been used, syntactic complexity has predominantly been operationalized as the mean length of T-unit (MLT), mean length of clause (MLC), or the proportion of clauses in a text that are dependent clauses (DC/C) (Lu, 2011; Ortega, 2003). The default hypothesis has been that proficient learners will produce more linguistically complex structures (e.g., more dependent clauses and/or longer T-units) as a function of language proficiency (broadly construed). Recently, however, a number of researchers have commented on some of the limitations of complexity indices such as MLT, MLC, and DC/C for the purpose of modeling differences across proficiency levels and/or developmental trends (Biber et al., 2011; Kyle & Crossley, 2017; Norris & Ortega, 2009). One limitation is that traditional syntactic complexity measures are based solely on the number of elements in a syntactic form and do not take into account the relative frequency of the syntactic forms or the relationship between the syntactic forms and the lexical items with which they are used.

A usage-based perspective, however, considers syntactic forms and lexical items to be inseparable: Language consists of form-meaning pairs, called *constructions*, that may differ in the level of specificity and schematicity and may range from morphemes to words, phraseological units, and verb argument constructions (VACs). Furthermore, because the main learning mechanism in both L1 and L2 is association, learning of constructions is sensitive to frequency, saliency, and contingency effects (e.g., Ellis, 2002). In English, for example, VACs such as *make—direct object* and *subject—have—direct object* are particularly frequent and are likely to be learned earlier than less frequent VACs.

In response to some of the limitations mentioned in the preceding text, Kyle and Crossley (2017) introduced a number of usage-based VAC indices related to frequency and contingency, which we will refer to as indices of VAC sophistication.[1] Kyle and Crossley (2017) compared the relationships between indices of VAC sophistication and syntactic complexity (e.g., MLT, MLC, and DC/C) and holistic scores of writing quality. Their findings indicated that indices of VAC sophistication were better predictors of holistic writing quality scores than indices of syntactic complexity. In the current study, we build on Kyle and Crossley (2017) by investigating whether the trends observed in their earlier cross-sectional study are also representative of developmental trends over time. This is particularly important given that studies have repeatedly found different trends for cross-sectional and longitudinal data (Bestgen & Granger, 2014; Bulté & Housen, 2014, 2018; Crossley & McNamara, 2014; Murakami & Alexopoulou, 2016). The mismatch between some cross-sectional and longitudinal studies suggests that although linguistic features evident at particular proficiency benchmarks may be relatively stable, the paths that individual learners take to reach those benchmarks may be diverse (e.g., de Bot et al., 2013; Larsen-Freeman & Cameron, 2008). Moreover, Lowie and Verspoor (2019) have made a strong case for complementing cross-sectional studies with individual longitudinal case studies because of the ergodicity problem (i.e., group means may not represent individual trajectories). In other words, to be sure developmental indices are useful, they need to be tested in both cross-sectional and longitudinal studies.

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

SYNTACTIC COMPLEXITY

For much of the time that SLA has been a field of study, one measure of productive language proficiency and/or development has been the complexity of the syntactic structures used in second language writing or speaking samples. Commonly, measures such as the MLT have been used as broad measures of complexity (e.g., Norris & Ortega, 2009; Ortega, 2003). Hunt (1965) introduced the T-unit (an independent clause and all associated clauses) and the associated index MLT to measure writing development in first language (L1) school age children. Hunt found a positive relationship between average T-unit length and grade level, suggesting that more advanced students tended to write longer T-units. Following Hunt's study, MLT also gained traction as an index of L2 development (Crymes, 1971; Larsen-Freeman, 1978; Larsen-Freeman & Strom, 1977; Thornhill, 1969). Other indices of syntactic complexity such as clauses per T-unit (C/T), MLC, and DC/C were also soon added to researchers' repertoire (Lu, 2011; Ortega, 2003; Wolfe-Quintero et al., 1998), though MLT has been the most popular (e.g., Ortega, 2003).

Although results have been mixed, a number of cross-sectional L2 writing studies have indicated a positive relationship between MLT and writing proficiency and/or grade level. Wolfe-Quintero et al. (1998), for example, report that 23 studies between 1970 and the mid-1990s found significant differences in MLT between proficiency levels (though 17 studies also found nonsignificant differences). More recently, Cumming (2005) found that more proficient writers (as assigned by holistic scores of writing quality) tended to write longer T-units. In an investigation of writing differences across university levels, Lu (2011) found that MLT increased after the first year of studying English in an EFL university context. Longitudinal studies have also indicated a generally positive relationship between indices of syntactic complexity and time spent studying a second language. Byrnes (2009), for example, found that L2 German writers used more words per T-unit as a function of time spent studying German. Vyatkina (2012) found similar results in mean length of sentence (MLS) scores both cross-sectionally and with two focal participants who were studied longitudinally. Bulté and Housen (2014) also found longitudinal growth in MLT scores during a semester-long English for academic purposes (EAP) L2 writing course. However, these gains are not always large (or statistically significant), even over longer periods (e.g., Knoch et al., 2014).

Although large-grained indices of syntactic complexity such as MLT have been (and continue to be) used in a number of L2 studies, a number of limitations have been noted in the literature regarding their use (Biber et al., 2011; Kyle & Crossley, 2017; Norris & Ortega, 2009). First, such measures may obscure the particular linguistic features that account for longer T-units (Norris & Ortega, 2009). Second, Biber et al. (2011) indicated that clausal subordination (which tends to be strongly correlated with MLT [Lu, 2011]) is characteristic of spoken texts, while a characteristic feature of academic writing is the use of elaborated noun phrases, suggesting that different measures may be needed for different modes and at different stages in L2 development. Subsequent cross-sectional and longitudinal research (Biber et al., 2014; Kyle & Crossley, 2018; Penris & Verspoor, 2017) has indicated that more proficient L2 writers will produce a higher proportion of elaborated phrases, and that phrasal complexity is a better predictor of writing proficiency scores than clausal subordination. A third issue with indices of syntactic complexity (and particularly with large-grained indices such as MLT) is that it is difficult to motivate or

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

interpret these indices from a theoretical perspective (Biber et al., 2011; Kyle & Crossley, 2017; Norris & Ortega, 2009). While research findings seem to indicate a fairly clear model of a positive relationship between syntactic elaboration and L2 proficiency, this model is based on syntactic structure only and is difficult to link to the frequency of occurrence of form-meaning mappings found in usage-based perspectives on language learning.

VAC SOPHISTICATION

Usage-based theories of language learning posit that learning occurs as a function of language use (Ellis, 2002; Tomasello, 2003; Verspoor, 2017). Language items that are heard and/or used more frequently are more strongly associated and are more salient, and will be learned earlier and/or more easily than those that are encountered less frequently (Ellis & Ferreira-Junior, 2009), regardless of their structural complexity. From a usage-based perspective then, language development can be modeled using the frequency, strength of association, and saliency of linguistic constructions in an individual's input. The roles of frequency and saliency in language learning are widely accepted (regardless of theoretical orientation) when dealing with lexis (e.g., O'Grady, 2008). Lexical sophistication studies have demonstrated, for example, that more proficient L2 users tend to use a higher proportion of low-frequency words than less proficient L2 users (e.g., Crossley et al., 2010; Kyle & Crossley, 2015; Laufer & Nation, 1995). Further, indices such as concreteness have been increasingly used to measure lexical salience and have shown that L2 users use words that are less concrete as a function of time spent studying an L2 (e.g., Crossley et al., 2016).

In a similar vein, there are a growing number of studies that have investigated the relationship between the frequency and strength of association of multiword utterances and proficiency (e.g., Bestgen & Granger, 2014; Garner et al., 2019; Verspoor & Smiskova, 2012). These studies of phrasal sophistication[2] have primarily investigated two-word combinations (either contiguous or with intervening words). In contrast to lexical studies, these studies have found that more proficient L2 users tend to use more frequent and more strongly associated (i.e., more targetlike) multiword utterances (c.f. Durrant & Schmitt, 2009). The fact that beginners do not use these word combinations to begin with may be explained through the principle of overgeneralization (e.g., Ninio, 1999; Verspoor & Behrens, 2011). In usage-based theories, overgeneralization refers to phenomena wherein learners discover that a particular construction has schematicity (i.e., that a seemingly fixed construction includes a variable slot) and subsequently use a wide range of items in this variable slot. This results in the production of a wide range of constructions (e.g., multiword utterances) that are not (necessarily) as strongly associated as in the target language yet. Through additional language exposure, however, the items learners use in a particular construction tend to represent their input more closely (and therefore are more strongly associated).

Studies of lexical and phrasal sophistication have indicated that more proficient L2 users are more likely to use less frequent words, but are more likely to use those words in more targetlike multiword constructions. Usage-based perspectives also extend to the levels of linguistic abstraction that have been traditionally regarded as syntax[3] (e.g., Ellis, 2002; Goldberg, 2006; Tomasello, 2003). Recent work related to L2 development from a

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

usage-based perspective has focused on the characteristics of VACs used by learners over time and/or at varying proficiency levels (Ellis & Ferreira-Junior, 2009; Eskildsen, 2009; Kyle & Crossley, 2017). Due to the difficulties involved with manually annotating VACs and in obtaining large samples of direct language input, most studies (e.g., Ellis & Ferreira-Junior, 2009; Eskildsen, 2009) have only examined a small subset of particularly frequent VACs and have only used small snippets of interlocutor data to represent a learner's linguistic input. However, recent computational advances have allowed researchers to automatically extract frequency and strength of association data from large corpora such as the British National Corpus (BNC; BNC Consortium, 2007; O'Donnell & Ellis, 2010) and the Corpus of Contemporary American English (COCA; Davies, 2009; Kyle & Crossley, 2017). Much like work in lexical and phraseological development, reference corpus norms for VACs can be used to represent language input. Research by Römer and colleagues, for example, show that corpus frequencies closely model the verb choices made by L1 (Römer et al., 2015) and advanced L2 (Römer et al., 2014) users for a variety of VACs.

Recently, Kyle and Crossley (2017) introduced a number of indices related to VAC frequency and strength of association, which we will refer to as indices of VAC sophistication (owing to similarities with frequency and strength of association-based indices of lexical sophistication and phraseology). They then compared the relationship between indices of VAC sophistication and syntactic complexity (e.g., MLT, MLC) and holistic scores of writing quality in TOEFL essays. They found that a model consisting of indices of VAC sophistication explained significantly greater variance in writing quality scores ($R^2 = .140$) than a model consisting of indices of syntactic complexity ($R^2 = .058$). The results indicated that more proficient writers (as measured by holistic writing quality scores) tended to use verb-VAC combinations that were on average less frequent but more strongly associated, suggesting that more proficient L2 users had learned more verb-VAC combinations and also used verb-VAC combinations that are more targetlike. With regard to syntactic complexity, only two indices were found to have a meaningful relationship with holistic writing score, and only one (MLC) was included in the final regression model. The findings of Kyle and Crossley (2017) provide some cross-sectional evidence that indices of VAC sophistication may be better indicators of writing proficiency at advanced levels than indices of syntactic complexity. However, it is unclear whether these cross-sectional models also represent developmental trends (Bestgen & Granger, 2014; Bulté & Housen, 2014; Crossley & McNamara, 2014; Lowie & Verspoor, 2019). Another open question is the degree to which indices of syntactic complexity provide unique predictive power when used in conjunction with VAC sophistication indices to predict proficiency levels and/or points along a developmental trajectory. Kyle and Crossley (2017) found differences in the predictive power of indices of syntactic complexity and VAC sophistication (using separate models), but did not directly examine the relationship between the two types of indices to determine the degree to which changes in VAC sophistication and changes in syntactic complexity are distinct.

CURRENT STUDY

In the current study, we investigate longitudinal development in L2 learners using syntactic indices related to complexity and VAC sophistication. First, we use the

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

commonly used indices of syntactic complexity (e.g., MLT, MLC, and DC/C) available in Lu's (2011) second language syntactic complexity analyzer (L2SCA) as a baseline and to allow comparisons with previous studies that have depended on large-grained indices. We then use indices of VAC sophistication that have previously been found to be important predictors of proficiency including three VAC sophistication indices related to VAC frequency (frequency of VAC + verb, frequency of VAC irrespective of verb, and frequency of verb irrespective of VAC) and three indices of verb-VAC strength of association (each of which uses a different method of calculating strength of association). Finally, we construct a model that uses both indices of syntactic complexity and VAC sophistication to predict the time at which a particular essay was written (e.g., near the beginning or end of the study). The goal of the final model is to determine the degree to which indices of syntactic complexity and VAC sophistication provide unique information about developmental trajectories.

This study is guided by the following research questions:

1. What is the relationship between indices of syntactic complexity and time spent studying English?
2. What is the relationship between VAC sophistication and time spent studying English?
3. To what extent do indices of syntactic complexity and VAC sophistication uniquely predict the time at which a particular essay was produced?

METHOD

LEARNER CORPUS

The longitudinal EFL essay corpus used in this study is derived from a larger corpus of essays ($n = 20$) collected from Dutch students at a competitive secondary school in the Netherlands over two academic years (Verspoor & Smiskova, 2012) and includes essays from the students ($n = 9$) who completed all assignments. Part of this corpus was also used for the cross-sectional study reported on by Verspoor et al. (2012). These students aged 12–13 had a high scholastic aptitude as determined by the Dutch CITO test, which most children take around age 11 or 12 (cf. Verspoor et al., 2015). The students had studied some English during elementary school. In addition, because of the high media exposure to English in the Netherlands, they were low intermediate learners at the beginning of the study. In terms of the Common European Framework of Reference (CEFR), the learners were at an A2 level based on scores assigned to their essays by certified CEFR raters.

Essays were collected three times per year, in October, February, and June of the academic year, for a total of six essays per student. The subset of the corpus used for this study includes essays from the nine participants who wrote essays at each collection point. Essays were completed at school using a computer and were untimed but limited to 1,000 characters. No dictionaries or other external resources were allowed. All participants wrote on the same prompt at a particular collection point. The prompts were relevant to the students at time of writing (i.e., they had ecological validity) and were designed in a manner that avoided the need for specialized language. Table 1 contains a list of the topics used.

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

TABLE 1. Essay topics in the longitudinal learner corpus

Essay	Prompt
1	Write a short story about your new school, friends, and teachers.
2	Pretend you have a foreign pen pal. Tell him/her about your favorite holiday and explain what you find so special about it.
3	Write about the most awful (or best) thing that happened to you at school so far. It does not have to be truthful.
4	Write a short story about the most awful (or best) thing that happened to you during summer vacation. It does not have to be truthful.
5	Pretend you have just won 1,000 euros. Write a short story about what you would do with the money.
6	Pretend your school principal has stated that from now on anyone should wear a school uniform. Write him/her a short letter to explain why you agree/do not agree with this new rule.

TABLE 2. Overview of the longitudinal learner corpus data

Participant	Gender	Average score	Number of words collected	Average words per essay
EFL_1	Female	3.833	1,030	171.667
EFL_2	Male	4.500	1,057	176.167
EFL_3	Female	4.333	1,239	206.500
EFL_4	Female	4.167	1,257	209.500
EFL_5	Female	4.000	1,001	166.833
EFL_6	Female	4.000	974	162.333
EFL_7	Female	4.167	1,087	181.167
EFL_8	Female	4.333	966	161.000
EFL_9	Male	3.833	1,202	200.333
Average	**N/A**	**4.130**	**1,090.333**	**181.722**

As reported in Verspoor et al. (2012), each essay was scored by four raters using a holistic proficiency rubric, which ranged from 0 to 7. The rubric emerged from lengthy discussions between a group of experienced ESL teachers. The final rubric descriptors referred to a range of features including text length, sentence length, sentence complexity, syntactic variation, use of tense, aspect, mood, range of vocabulary, use of L1, use of idiomatic language, and accuracy (see the Appendix for the full rubric). Raters included eight experienced ESL teachers, who were split into two groups. If three of four raters agreed on a score, the score was kept. If fewer than three raters agreed on an essay score, the score was then adjudicated by the raters until sufficient agreement was reached (Verspoor et al., 2012, 2015). Table 2 includes an overview of the nine participants' writing from the longitudinal learner corpus included in this analysis.

LINGUISTIC INDICES

All texts were analyzed using the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC; Kyle, 2016). Each of the selected indices is described in the following text.

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

Indices of Syntactic Complexity

To compute syntactic complexity measures, we used Lu's (2011) L2 complexity analyzer (L2SCA) indices, which have been shown to be highly reliable with L2 data.[4] We use the version of L2SCA that is included as part of TAASSC. L2SCA includes 14 indices of syntactic complexity drawn from Ortega (2003) and Wolfe-Quintero et al. (1998). For this study, five indices of syntactic complexity outlined in Lu (2011) were considered. Two large-grained indices, MLT and MLC, were selected based on comparability with the bulk of previous studies (Ortega, 2003). Three more fine-grained indices were also selected to cover the use of a range of particular structures including complex nominals per clause (CN/C), coordinate phrases per clause (CP/C), and DC/C. Table 3 includes a description of the structures counted by L2SCA to calculate the variables used in this study, and Table 4 comprises a list of the five L2SCA indices used including a short description of each. For further information, refer to Lu (2011).

TABLE 3. A description of syntactic structures counted by L2SCA for variables used in this study

Structure	Description	Examples
Word	a sequence of letters that are bounded by white space	*I* *Ate*
Complex nominal	i. nouns with modifiers ii. nominal clauses iii. gerunds and infinitives that function as subjects	i. *red car* ii. I know *that she is hungry* iii. *Running* is invigorating
Coordinate phrase	adjective, adverb, noun and verb phrases connected by a coordinating conjunction	She *eats pizza and smiles*
Clause	a syntactic structure with a subject and a finite verb	*I ate pizza* *because I was hungry*
Dependent clause	a finite clause that is a nominal, adverbial, or adjective clause	I ate pizza *because I was hungry*
T-unit	an independent clause and any clauses dependent on it	*I ate pizza* *I ate pizza because I was hungry*

Note: As reported in Kyle and Crossley (2017, p. 521).

TABLE 4. A description of L2SCA variables considered

Index abbreviation	Index name	Index description
MLT	mean length of T-unit	number of words per T-unit
MLC	mean length of clause	number of words per clause
DC/C	dependent clauses per clause	number of dependent clauses per clause
CP/C	coordinate phrases per clause	number of coordinate phrases per clause
CN/C	complex nominals per clause	number of complex nominals per clause

Note: As reported in Kyle and Crossley (2017, p. 521).

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

VAC Sophistication Indices

TAASSC calculates a number of indices related to VAC frequency and strength of association derived from the corpus of contemporary American English (COCA; Davies, 2009). Identification of VAC features in TAASSC is based on the Stanford Neural Network Dependency Parser (Chen & Manning, 2014), a state-of-the-art dependency parser.[5] For this study, six indices related to frequency and strength of association were calculated using the ~360-million-word written sections of COCA (academic, fiction, magazine, and newspaper). The calculation of frequency scores and three strength of association scores (delta P VAC to verb, delta P verb to VAC, and a bidirectional measure) are described in the following text.

Frequency

Frequency indices related to main verb lemmas, unfilled VAC frames (e.g., nominal subject—verb slot—direct object), and VAC frames with particular main verb lemmas (e.g., nominal subject—*have*—direct object) were used. See Tables 5–7 for examples of the most frequent main verb lemmas (Table 5), VACs (Table 6), and verb-VAC combinations (Table 7) in COCA. Mean frequency scores for main verbs, VACs and verb-VAC combinations were calculated for each target text. If a particular target structure (e.g., a VAC) that occurs in a text does not occur in the reference corpus, it is not counted toward the index score.

Delta P

Delta P calculates the probability of an outcome (e.g., a VAC) given a cue (e.g., a particular verb) minus the probability of the outcome without the cue (e.g., with any other verb). Delta P is calculated with both VACs as cues and with verbs as cues. To calculate delta P with a VAC as the outcome and a verb as the cue we use the following formula (see Table 8): $\textit{delta p} = \left(\frac{\mathrm{a}}{\mathrm{a+b}}\right) - \left(\frac{\mathrm{c}}{\mathrm{c+d}}\right)$. The delta P value for the outcome of the SVO given the cue *have* (see Table 8) is calculated as: $\textit{delta p} = \left(\left(\frac{212,970}{212,970+991,685}\right) = .177\right) - \left(\left(\frac{1,733,964}{1,733,964+30,909,494}\right) = .053\right) = .124$. The probability

TABLE 5. Main verb lemma frequencies in the written sections of COCA

Rank	Frequency (per million VACs)	Main verb lemma
1	160,994.48	be
2	35,711.92	say
3	30,182.42	have
4	15,366.25	make
5	15,229.85	do
6	14,840.33	go
7	13,306.69	get
8	11,900.44	see
9	11,644.46	take
10	11,608.18	know

Note: As reported in Kyle and Crossley (2017, p. 522).

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

TABLE 6. Verb argument construction frequencies in the written sections of COCA

Rank	Frequency (per million VACs)	Verb argument construction	Most frequent co-occurring main verb lemma for each VAC
1	64,733.43	verb—direct object	make
2	48,780.10	subject—verb—direct object	have
3	34,540.26	subject—verb—nominal complement	be
4	33,315.86	subject—verb—adjective complement	be
5	21,321.88	subject—verb	say
6	20,297.22	subject—verb—clausal complement	say
7	15,960.63	subject—verb—external complement	have
8	11,788.37	verb—clausal complement	say
9	11,117.08	Verb	base
10	9,879.52	subordinator—subject—verb—direct object	have

Note: As reported in Kyle and Crossley (2017, p. 523).

TABLE 7. Most common verb argument construction-main verb lemma combinations in the written sections of COCA

Rank	Frequency (per million VACs)	Main Verb lemma	verb argument construction	Example (register)
1	34,517.41	Be	subject—verb—nominal complement	***It is*** *also an* ***indication*** *of the ways* … (academic)
2	33,287.74	Be	subject—verb—adjective complement	***They are*** *very* ***discerning*** … (news)
3	6,843.83	Be	subordinator—subject—verb—adjective complement	She hears ***that he is arrogant***. (news)
4	6,318.98	Say	clausal complement—subject—verb	***[“Andy is an amalgamation of all the douchebags that I ’ve dealt with in my life”], Helms says***. (magazine)
5	5,335.93	Have	subject—verb—direct object	***Iran has*** *obvious* ***interests*** *in Iraq*. (magazine)
6	5,124.34	Be	verb—nominal complement	That’s what’s great about ***being a teen***. (news)
7	4,986.51	Be	subordinator—subject—verb—nominal complement	Even before the man reached the car, she knew ***that it was Frank***. (fiction)
8	4,258.04	Be	verb—adjective complement	This is the reason I have found life ***to be harder than fiction*** … (fiction)
9	3,865.16	Say	subject—verb—clausal complement	***He said [that health decisions should be made by patients and doctors]*** (magazine)
10	3,516.17	Say	clausal complement—verb—subject	***[“We have an all-new situation now”], says Europol’s Storbeck*** (magazine)

Note: As reported in Kyle and Crossley (2017, p. 523).

TABLE 8. Contingency table used to calculate various indices of association strength

	Construction C	Not construction C	
	(*nsubj-v-dobj*)	(not *nsubj-v-dobj*)	Totals
Verb V	**a**	**b**	**a + b** = frequency of V
(*have*)	(212,970)	(991,685)	(1,204,655)
Not verb V	**c**	**d**	**c + d =** combinations that are not V + C
(not *have*)	(1,733,964)	(30,909,494)	(32,643,458)
	a + c	**b + d**	**(a + b) + (c + d)** = N (total number of VAC tokens in the corpus) = (33,848,113)
Totals	(1,946,934) frequency of C	(31,901,179)	

Note: Adapted from Gries et al. (2005, p. 644) and Kyle and Crossley (2017, p. 524).

of the outcome SVO given the cue *have* (.177) is larger than the probability that the SVO will be the outcome given another verb cue (.053), resulting in a positive delta P value (.124). In the current study, we use two indices based on the mean delta P score for tokens in the target text, one using the verb as the cue, and the other using the construction as the cue. Mean delta P scores are calculated for each target text (separate indices are included for verb-cued and VAC-cued scores) based on the written sections of COCA.

Bidirectional Association Strength

In studies that measure the strength of association between VACs and the verbs that fill them, both unidirectional measures (such as delta p) and bidirectional measures have been used. The bidirectional association strength measure of choice has been a transformation of the Fisher–Yates exact test[6] (Fisher, 1934; Yates, 1934), which refers to the joint probability that a VAC and a verb will co-occur (Gries et al., 2005). One issue in calculating the Fisher–Yates exact test with large corpora is that the resulting probability values are so small that most programs round them to $p = 0$ ($-\log_{10}(0) =$ infinity; for strongly associated verb-VAC combinations such as *have* + SVO) or $p = 1$ ($-\log_{10}(1) = 0$; for less strongly associated verb-VAC combinations such as *impute* + SVO). This can result in issues related to precision, where a large percentage of observed verb-VAC combinations are assigned the same association score due to rounding. One solution to this issue is to calculate a related measure that correlates almost perfectly with the original method (S. T. Gries, personal communication, 2014), but is much easier to compute with large frequency values (such as those found in COCA). This method is calculated as follows: $\textit{bidirectional association strength} = \left(\left(\frac{a}{a+b}\right) - \left(\frac{c}{c+d}\right)\right) * (a+b)$. Mean bidirectional association strength scores are calculated for each target text based on the written sections of COCA.

ANALYSES

To determine whether a systematic relationship existed between indices of syntactic complexity and VAC sophistication and time spent studying English, a series of linear mixed-effects (LME) models were developed. LME models can examine development over time while controlling for individual participant behaviors (Gries, 2015). The R (R Core Team, 2016) package *lme4* (Bates, 2010) was used to perform LME analyses. In each analysis, the linguistic variable (e.g., MLT or verb frequency) was set as the dependent variable, Time (which refers to collection points 1–6) was set as a fixed effect, and Participant was set as a random effect (with random intercepts). The LME analysis presumed that while participants may have different starting points (e.g., initial MLT values), their development would follow similar trends. Thus, the following equation was used for all analyses: lmer(Syntactic_Index ~ Time + (1|Participant)). The R package *lmerTest* (Kuznetsova et al., 2015) was used to estimate *p* values with the default Satterthwaite method (Satterthwaite, 1941). In cases in which the Satterthwaite method failed to converge, the Kenward–Roger method (Kenward & Roger, 1997) was used. The R package *MuMIn* (Barton, 2013) was used to generate effect sizes and to estimate the relative importance of fixed factors. We report both the marginal R^2 values (R^2m), which indicates the amount of variance explained by only the fixed effects, and the conditional R^2 values (R^2c), which indicate the amount of variance explained by the fixed effects and the random effects. In cases in which the model was not able to fit the random effects, only the R^2m values are reported.[7]

RESULTS AND DISCUSSION

PRELIMINARY ANALYSIS

Before addressing the research questions, we examined whether the participants become more proficient writers of English as a function of time. A LME model indicated a significant positive relationship between time and holistic essay scores ($p < .001$, R^2m = .576). A summary of the model can be found in Table 9. The results indicate that the participants' writing proficiency increased over the 2-year period, at a rate of approximately .5 points per collection point. Figure 1 includes plots of the holistic scores given to each participant's essays at each collection point with the regression line (indicated by the dashed line) produced by the model.

RQ1 Results: Syntactic Complexity Indices

Five indices of syntactic complexity outlined in Lu (2011) and discussed in the preceding text were used to examine longitudinal growth in the learner corpus (i.e., MLT and MLC,

TABLE 9. LME model predicting holistic writing score

Random effects	Variance	Standard deviation		
Participant	<0.001	<0.001		
Residual	0.630	0.794		
Fixed effects	**Estimate**	**Standard error**	***T***	***p***
(Intercept)	2.252	0.246	9.143	<.001
Time	0.537	0.063	8.483	<.001

Note: The fit was singular.

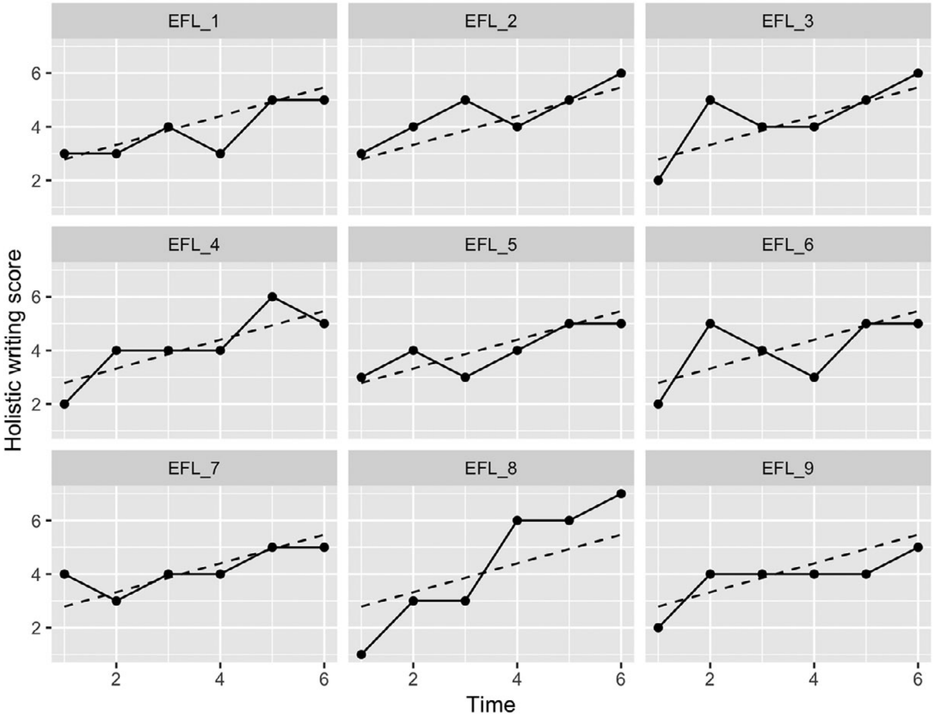

FIGURE 1. Holistic scores at each time point for each participant.

TABLE 10. LME model predicting complex nominals per clause

Random effects	Variance	Standard deviation		
Participant	<0.001	<0.001		
Residual	0.021	0.144		
Fixed effects	**Estimate**	**Standard error**	***t***	***p***
(Intercept)	0.454	0.045	10.210	<.001
Time	0.010	0.011	0.871	0.388

Note: The fit was singular.

CN/C, CP/C, and DC/C). An initial check for assumptions indicated that CP/C failed to meet the assumption of normality (due to rare occurrence in the corpus) and was removed from further consideration. The results for each of the remaining four indices are reported in the following text.

CN/C. A LME model was run using Time to predict the number of CN/C in the participant essays. The model did not identify a clear relationship between Time and CN/C in the participant essays ($p = .388$, $R^2m = .014$). A summary of the model can be found in Table 10. Figure 2 includes line plots for the CN/C scores for each participant at each collection point with the regression lines (indicated by the dashed line) produced by the model.

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

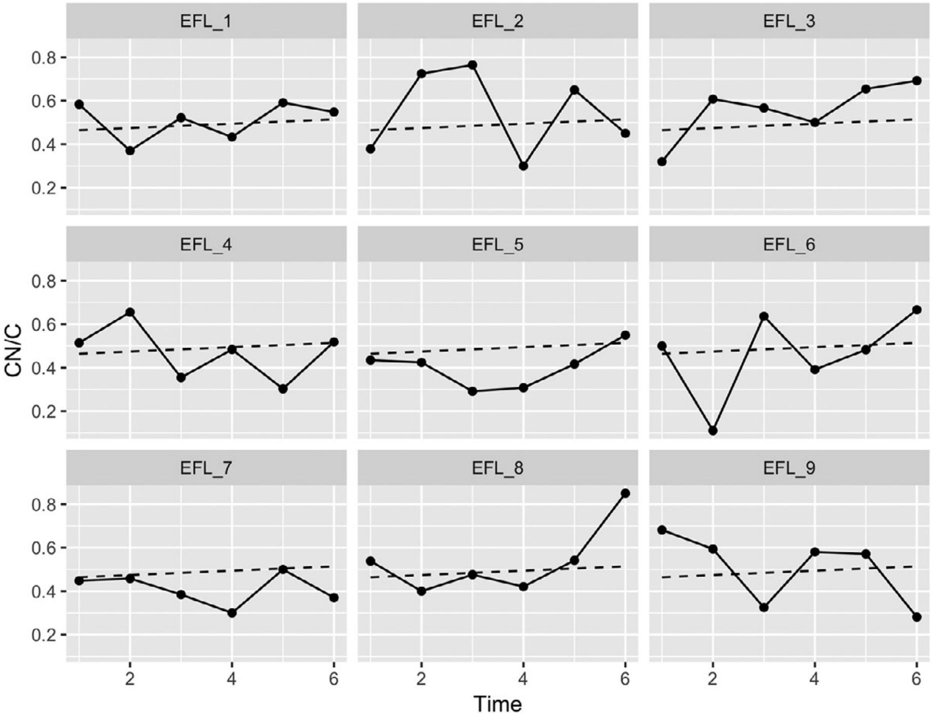

FIGURE 2. Complex nominals per sentence at each time point for each participant.

TABLE 11. LME model predicting dependent clauses per clause

Random effects	Variance	Standard deviation		
Participant	0.001	0.026		
Residual	0.010	0.101		
Fixed effects	Estimate	Standard error	*t*	*p*
(Intercept)	0.076	0.032	2.341	0.023
Time	0.049	0.008	6.154	<.001

DC/C. A LME model was run using Time to predict the number of dependent clauses per clause. The results of the analysis indicated that there was a significant positive relationship between Time and DC/C in the participant essays ($p < .001$, $R^2m = .401$, $R^2c = .440$). A summary of the model can be found in Table 11. The results indicate that the participants used a higher proportion of dependent clauses at each progressive collection point over the 2-year period. At time point one, dependent clauses on average account for approximately 12.5% of the clauses in a participant text. At each time point, the proportion of dependent clauses increases by 4.9%, indicating that by the final time point, dependent clauses account for approximately 37% of the clauses in each text. Figure 3 includes plots of the DC/C scores for each participant at each collection point with the regression lines (indicated by the dashed line) produced by the model.

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

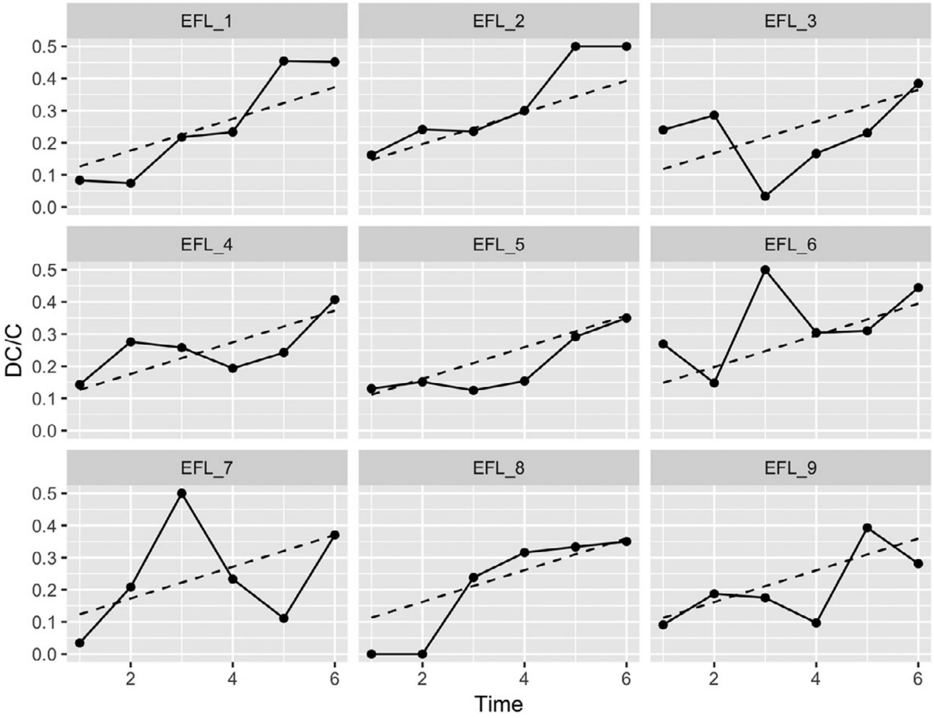

FIGURE 3. Dependent clauses per clause at each time point for each participant.

TABLE 12. LME model predicting mean length of clause

Random effects	Variance	Standard deviation		
Participant	0.357	0.598		
Residual	1.093	1.046		
Fixed effects	Estimate	Standard error	*t*	*p*
(Intercept)	6.760	0.381	17.754	<.001
Time	0.081	0.083	0.973	0.336

MLC. A LME model was run using Time to predict the MLC in the participant essays. The model did not identify a clear relationship between Time and MLC in the participant essays (p = .336, R^2m = .014, R^2c = .256). A summary of the model can be found in Table 12. Figure 4 includes plots of the MLC scores for each participant at each collection point with the regression lines (indicated by the dashed line) produced by the model.

MLT. A LME model was run using Time to predict the MLT in the participant essays. The results of the analysis indicated that there was a significant positive relationship between Time and MLT in the participant essays (p < .001, R^2m = .191, R^2c = .355). A summary of the model can be found in Table 13. The results indicate that the participants wrote longer T-units over the 2-year period, at a rate of approximately .7 words per T-unit

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

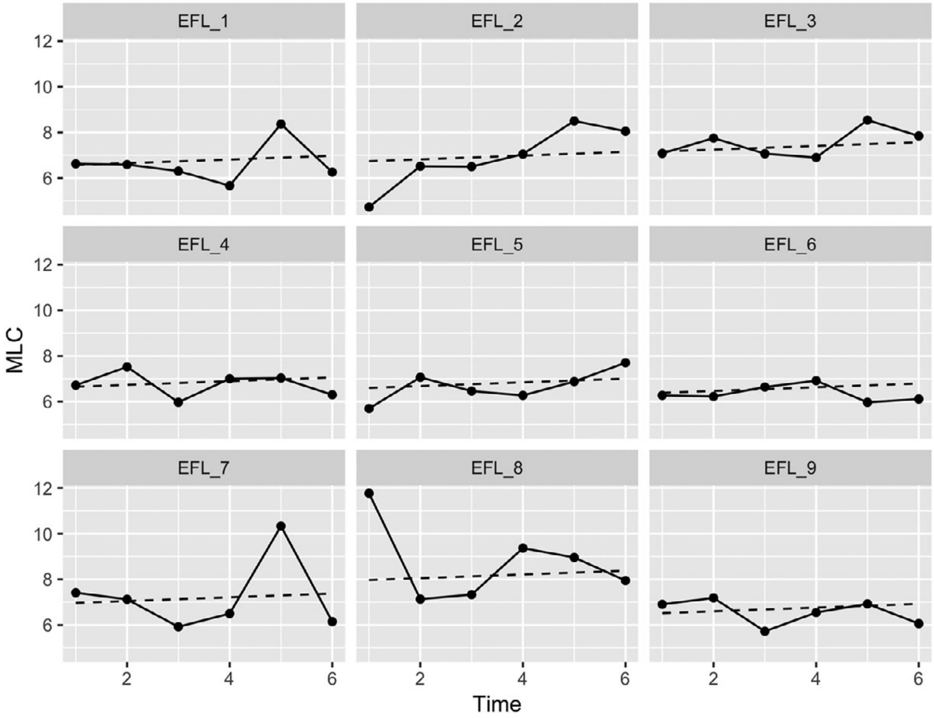

FIGURE 4. Mean length of clause at each time point for each participant.

TABLE 13. LME model predicting mean length of T-unit

Random effects	Variance	Standard deviation		
Participant	1.256	1.121		
Residual	4.934	2.221		
Fixed effects	Estimate	Standard error	*t*	*p*
(Intercept)	9.064	0.784	11.56	<.001
Time	0.701	0.177	3.96	<.001

at progressive collection points. Figure 5 includes plots of the MLT scores for each participant at each collection point with the regression lines (indicated by the dashed line) produced by the model.

RQ1 Discussion: Syntactic Complexity Indices

The results of the four analyses indicated that significant positive linear trends existed between Time and DC/C and MLT with moderate to large effect sizes. As participants spent time studying English, they tended to write longer T-units and produced a higher proportion of dependent clauses. No significant trends were observed between Time and MLC and CN/C, and only small effect sizes ($r = .118$, $R^2m = .014$) were observed. T-units

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

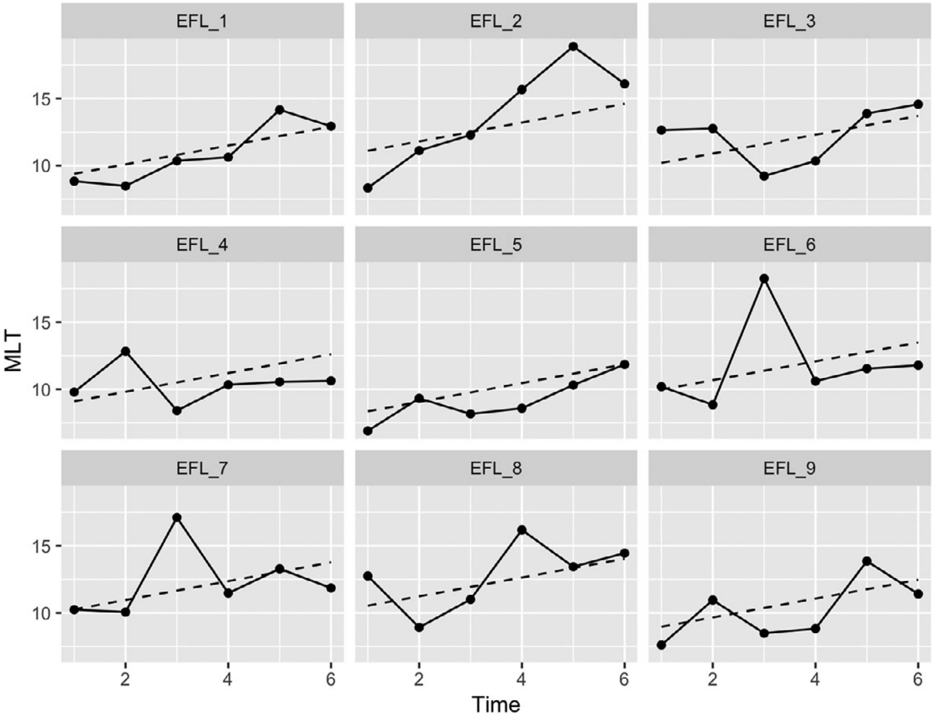

FIGURE 5. Mean length of T-unit values at each time point for each participant.

TABLE 14. Correlations between MLT, DC/C, and Time

	r	p
Time and DC/C	0.638	<0.001
Time and MLT	0.443	<0.001
MLT and DC/C	0.732	<0.001
Time and DC/C while controlling for MLT	0.513	<0.001
Time and MLT while controlling for DC/C	−0.046	0.744

can be lengthened through the addition of a number of linguistic constructions such as phrasal modifiers and/or subordinated clauses. Given that there was not a clear relationship between Time and CN/C, it is likely that the relationship between MLT and Time was primarily attributable to the addition of subordinate clauses. To test this hypothesis, post-hoc partial correlations were calculated using the ppcor package (Kim & Kim, 2015) in R. The results, which are summarized in Table 14, indicate a moderate relationship (r = .443) between Time and MLT. However, when we control for the influence of DC/C on the relationship between Time and MLT, the relationship between Time and MLT is negligible (r = −.046). Conversely, when we control for the influence of MLT on the relationship between Time and DC/C, the relationship between Time and DC/C is still

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

meaningful (r = .513). These results suggest that DC/C is the clearest and most accurate indicator of longitudinal growth in syntactic complexity in this dataset.

Overall, these results align with the majority of previous studies that have found increases in MLT as a function of time and/or proficiency (Bulté & Housen, 2014; Lu, 2011; Ortega, 2003). The results with regard to DC/C support the general hypothesis (at least for lower-proficiency learners) that more proficient writers will use more subordinated clauses in writing and is in line with Verspoor et al.'s (2012) finding that there was a strong correlation between scores and number of dependent clauses. As this trend has not been consistently demonstrated in the L2 writing literature (e.g., Bulté & Housen, 2014; Ortega, 2003), it is important to consider the effects of writing task (narrative and expository writing), age (middle school students), L2 writing proficiency (low), and context (bilingual EFL secondary school). Bulté and Housen (2014), for example, found that university-level ESL students demonstrated no meaningful development in DC/C over the course of a semester, but did demonstrate increases in noun phrase complexity, and similar results were reported in Kyle and Crossley (2018) as predicted by Biber et al. (2011). Further, Verspoor and colleagues (e.g., Penris & Verspoor, 2017; Verspoor et al., 2012) have argued that different linguistic features, both syntactic and lexical, will be prominent at different proficiency levels and stages of development. Dependent clause use may increase between some levels (such as the lower levels represented in this study), but then may become asymptotic and/or may be replaced by the use of other structures (e.g., features of noun phrase complexity).

RQ2 Results: VAC Sophistication Indices

Six indices related to usage-based theories of language development were considered. Three frequency measures were considered, including main verb frequency, VAC frequency, and verb-VAC frequency. Log transformations were used to adjust for the Zipfian nature of frequency data. Additionally, three indices that measured the strength of association between VACs and the verbs that filled them were considered, including delta p (verb as cue), delta p (VAC as the cue), and a bidirectional measure of association strength. All indices were calculated based on frequencies from the written portions of the COCA (Davies, 2009). The results of each of the six LME analyses are included in the following text.

Main Verb Frequency. A LME model was run using Time to predict the mean main verb frequency score in the participant essays. The results of the analysis indicated that there was a negative relationship between Time and main verb frequency in the participant essays (p < .001, R^2m = .457, R^2c = .467). A summary of the model can be found in Table 15. The results indicate that the participants used less frequent main verbs as a function of time. Figure 6 includes line plots for the main verb frequency scores for each participant at each collection point with the regression lines (indicated by the dashed line) produced by the model.

VAC Frequency. A LME model was run using Time to predict the mean VAC frequency score in the participant essays. The model did not identify a clear relationship between Time and VAC frequency in the participant essays (p = .425, R^2m = .012). A summary of the model can be found in Table 16. Figure 7 includes plots of the VAC frequency scores

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

TABLE 15. LME model predicting mean main verb frequency score

Random effects	Variance	Standard deviation		
Participant	0.001	0.030		
Residual	0.044	0.210		
Fixed effects	**Estimate**	**Standard error**	***T***	***p***
(Intercept)	6.065	0.066	92.160	<.001
Time	−0.113	0.017	−6.741	<.001

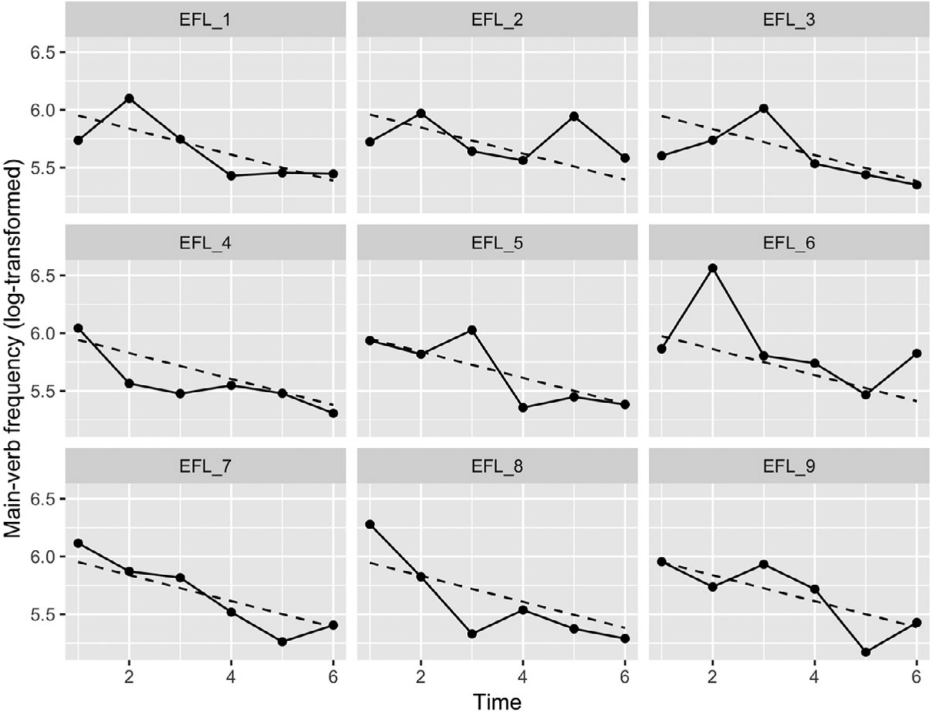

FIGURE 6. Average main verb frequency at each time point for each participant.

TABLE 16. LME model predicting mean VAC frequency score

Random effects	Variance	Standard deviation		
Participant	<0.001	<0.001		
Residual	0.133	0.365		
Fixed effects	**Estimate**	**Standard error**	***T***	***p***
(Intercept)	4.735	0.113	41.822	<0.001
Time	−0.023	0.029	−0.804	0.425

Note: The fit was singular.

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

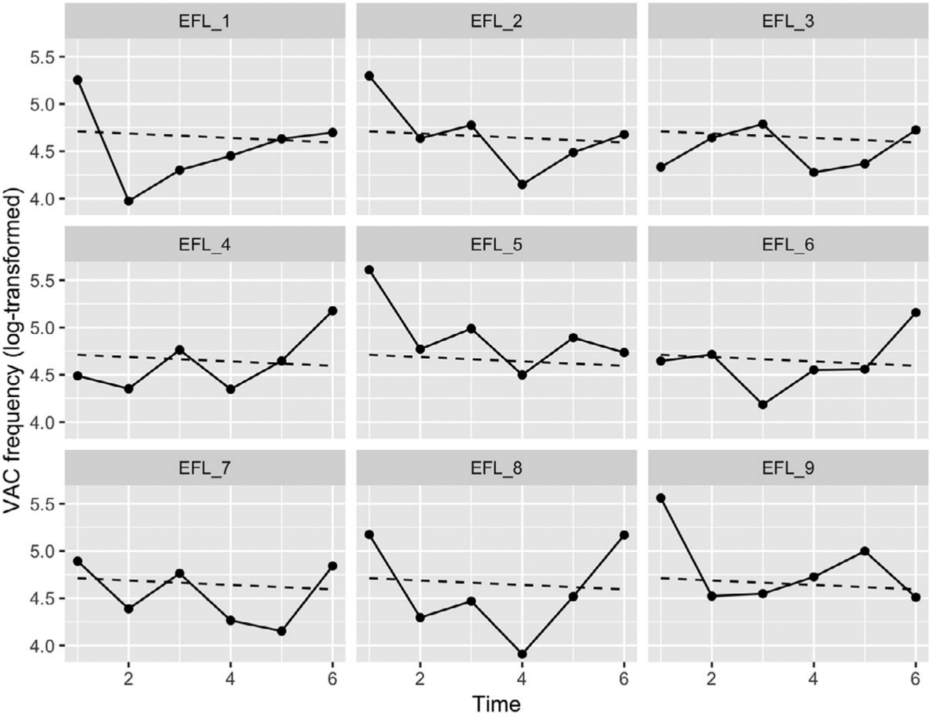

FIGURE 7. Average VAC frequency at each time point for each participant.

TABLE 17. LME model predicting mean verb-VAC frequency score

Random effects	Variance	Standard deviation		
Participant	<0.001	<0.001		
Residual	0.219	0.468		
Fixed effects	Estimate	Standard error	*t*	*p*
(Intercept)	4.080	0.145	28.096	<0.001
Time	−0.153	0.037	−4.105	<0.001

Note: The fit was singular.

for each participant at each collection point with the regression lines (indicated by the dashed line) produced by the model.

Verb-VAC Frequency. A LME model was run using Time to predict the mean verb-VAC frequency score in the participant essays. The results of the analysis indicated that there was a negative relationship between Time and verb-VAC frequency in the participant essays ($p < .001$, $R^2m = .241$). A summary of the model can be found in Table 17. Figure 8 includes plots of the verb-VAC frequency scores for each participant at each collection point with the regression lines (indicated by the dashed line) produced by the model.

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

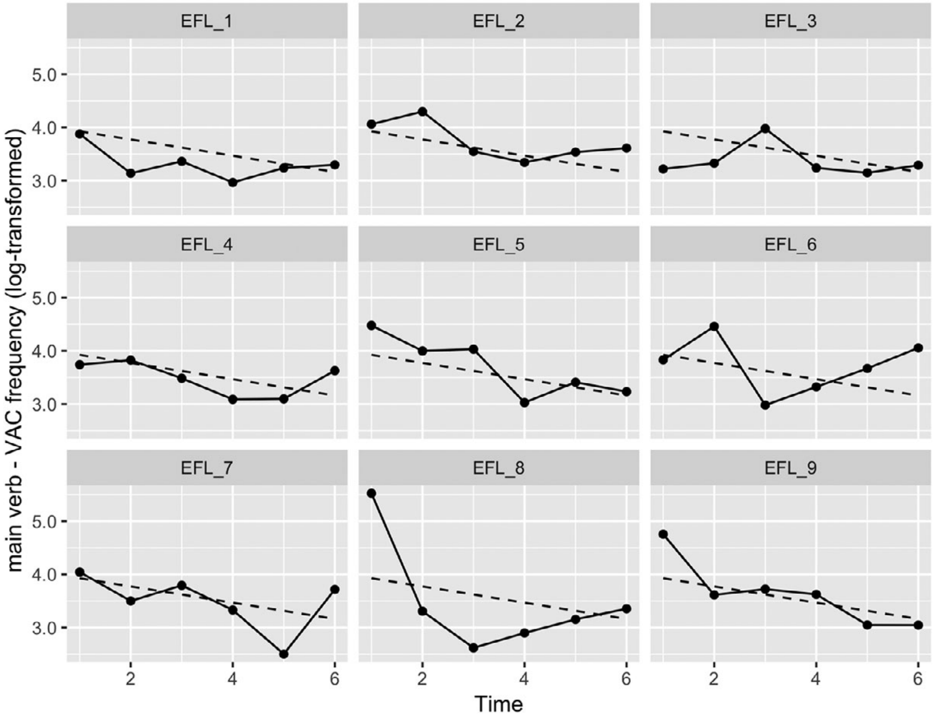

FIGURE 8. Average verb-VAC frequency at each time point for each participant.

TABLE 18. LME model predicting mean strength of association score (delta P verb as cue)

Random effects	Variance	Standard deviation		
Participant	<0.001	<0.001		
Residual	0.000	0.015		
Fixed effects	**Estimate**	**Standard error**	***t***	***p***
(Intercept)	0.021	0.005	4.457	<0.001
Time	0.000	0.001	−0.385	0.702

Note: The fit was singular.

Delta p (Verb as Cue). A LME model was run using Time to predict the mean strength of association score (delta p, verb as cue) in the participant essays. The model did not identify a clear relationship between Time and strength of association score as measured by delta p (verb as cue) in the participant essays ($p = .702$, $R^2m = .003$). A summary of the model can be found in Table 18. Figure 9 includes line plots for the delta p (verb as cue) scores for each participant at each collection point with the regression lines (indicated by the dashed line) produced by the model.

Delta p (VAC as Cue). A LME model was run using Time to predict the mean strength of association score (delta p, VAC as cue) in the participant essays. The model did not

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

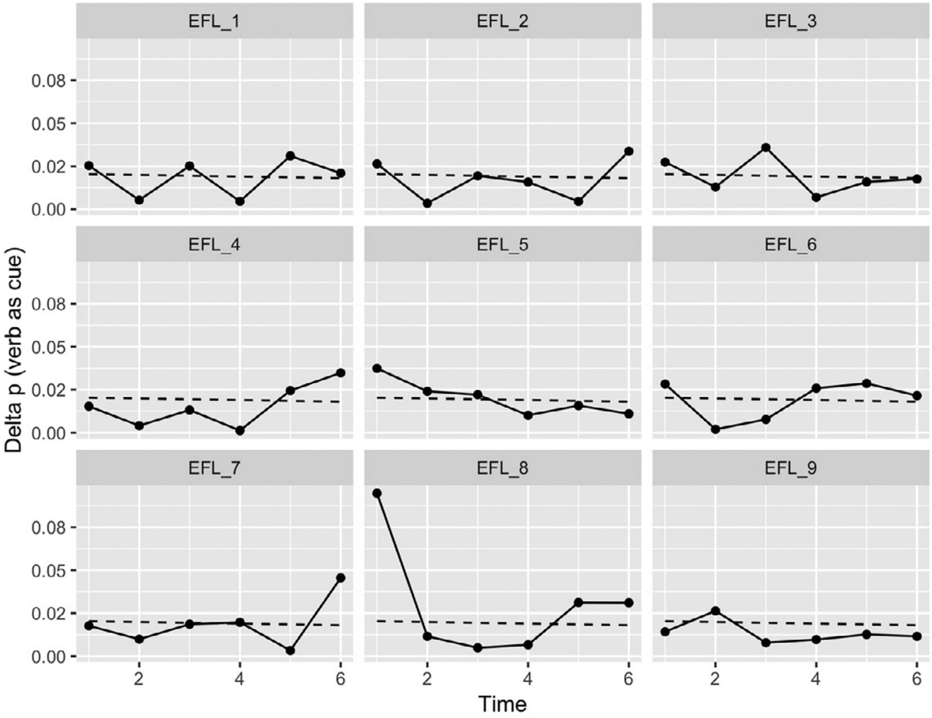

FIGURE 9. Average delta P (verb as cue) score at each time point for each participant.

TABLE 19. LME model predicting mean strength of association score (delta P VAC as cue)

Random effects	Variance	Standard deviation		
Participant	<0.001	<0.001		
Residual	0.000	0.002		
Fixed effects	**Estimate**	**Standard error**	***t***	***p***
(Intercept)	0.035	0.007	5.246	<0.001
Time	0.002	0.002	1.332	0.189

Note: The fit was singular.

identify a clear relationship between Time and strength of association score as measured by delta p (VAC as cue) in the participant essays ($p = .189$, $R^2m = .032$). A summary of the model can be found in Table 19. Figure 10 includes line plots for the delta p (VAC as cue) scores for each participant at each collection point with the regression lines (indicated by the dashed line) produced by the model.

Bidirectional Association Strength. A LME model was run using Time to predict the mean strength of association score (bidirectional association strength) in the participant essays. The model did not identify a clear relationship between Time and strength of association score as measured by bidirectional association strength in the participant

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

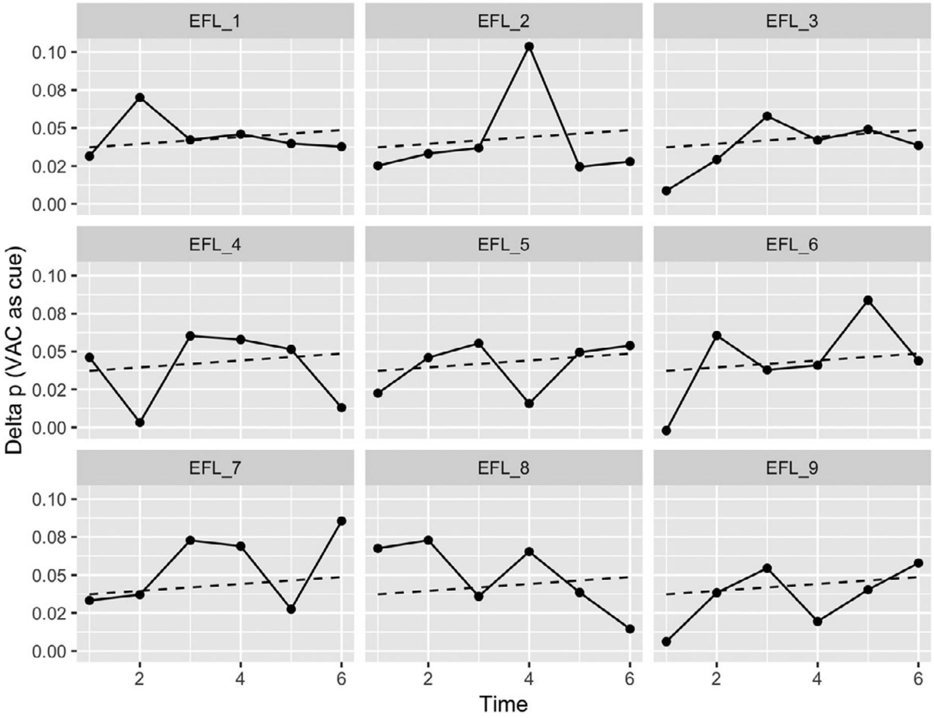

FIGURE 10. Average delta p (VAC as cue) at each time point for each participant.

TABLE 20. LME model predicting mean strength of association score (collostructional strength)

Random effects	Variance	Standard deviation		
Participant	<0.001	<0.001		
Residual	755,500,000.000	27,490.000		
Fixed effects	**Estimate**	**Standard error**	***T***	***p***
(Intercept)	20080.700	8529.500	2.354	0.022
Time	–432.900	2190.200	–0.198	0.844

Note: The fit was singular.

essays (p = .472, R^2m < .001). A summary of the model can be found in Table 20. Figure 11 includes line plots for the bidirectional association strength scores for each participant at each collection point with the regression lines (indicated by the dashed line) produced by the model.

RQ2 Discussion: VAC Sophistication Indices

Usage-based theories of language learning suggest that constructions that are frequently encountered and used will be easier to learn (and will be more likely to be produced) than

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

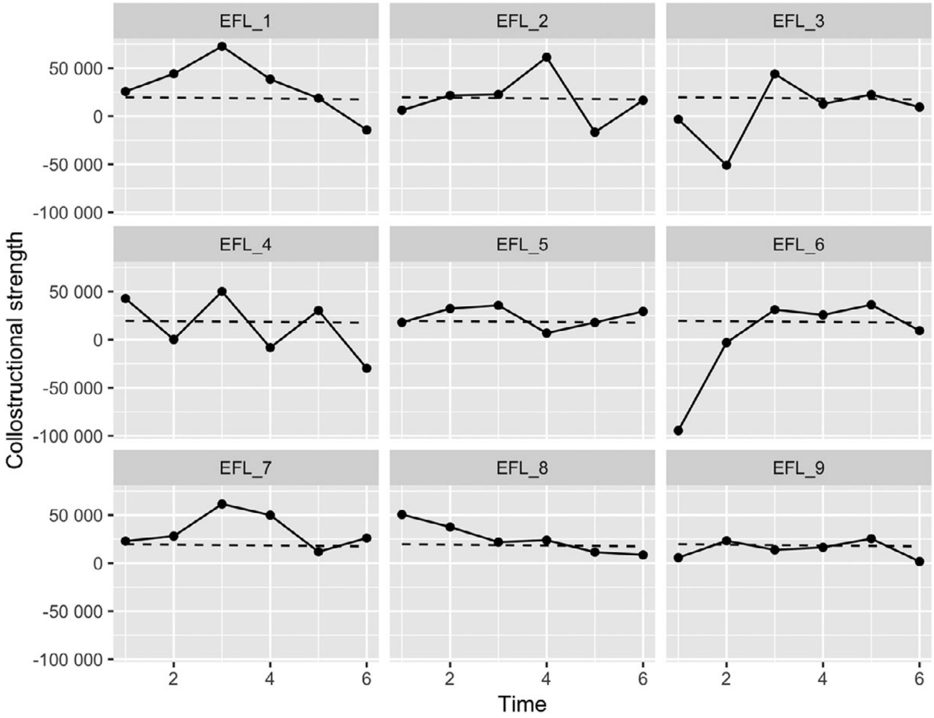

FIGURE 11 Average bidirectional association strength score at each time point for each participant.

those that are less frequently encountered and used. Further, verb-VAC combinations that are more strongly associated will also be easier to learn than those that are less strongly associated (e.g., Ellis & Ferreira-Junior, 2009). Less proficient learners may, however, produce less strongly associated verb-VAC combinations when they overgeneralize the use of verbs with newly learned VACs. A previous cross-sectional study based on holistic ratings of writing quality (Kyle & Crossley, 2017) found that as holistic scores increased, the frequency of verb-VAC combination decreased, but strength of association scores increased. This preliminarily finding suggested that writers of higher-quality essays not only had a larger VAC repertoire but also had learned which verbs tend to co-occur with particular VACs (and vice versa).

This longitudinal study partially replicated the findings of the cross-sectional study of Kyle and Crossley (2017). Much like Kyle and Crossley (2017), a meaningful relationship was found between verb-VAC frequency and time spent studying English ($R^2_m = .241$, $R^2_c = .241$), suggesting that as the participants spent time learning English, they learned (and used) less frequent verb-VAC combinations, as would be expected by usage-based accounts. However, there are at least two points of departure between the results of the present study and the previous study. First, in the previous study, the strongest predictor of proficiency was the strength of association between a VAC and the verb that fills it ($r = .251$, $R^2 = .063$). In this study, however, there was no meaningful relationship between any of the strength of association indices and time spent studying English. The most reasonable explanations may rest in the proficiency level of the participants in the current

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

study, which were relatively low. As Verspoor et al. (2012) showed in their cross-sectional study on the same learners, some linguistic constructions (such as simple sentences, dependent clauses, different types of chunks) seemed to develop at different rates and word frequency was a strong predictor of text score at all stages. However, longer collocational sequences such as fixed phrases, the use of particles, and compounds did not show significant increases until the end of the study (between stage 4 and 5). In other words, for these low-proficiency learners, early frequency effects are seen more strongly in isolated words than in groups of words, which would make sense as single words may be remembered more easily than combinations of words. Another potential explanation is that the method of measuring verb-VAC association strength (i.e., mean association strength score) may have obscured change that was taking place in specific VACs. It is possible that development occurred in particular VACs (e.g., double-object VACs), but not (or differently) in others. It may be useful, therefore, to look at the trajectories of individual VACs (much like some previous studies have done) instead of looking at mean scores (as has been common in studies of lexical sophistication and phraseology). Finally, if we consider the examples in Table 7, we can see that the most frequent VACs in the corpus are from a rather academic, formal written register. So, we may assume that these low-proficiency learners had not been exposed to these VACs yet, and this corpus may be more representative for advanced learners than young, low-proficiency learners. Each of these factors deserves further investigation that is beyond the scope of this article.

Second, unlike the previous cross-sectional study, there was a stronger relationship between Time and main verb frequency (R^2m = .457, R^2c = .467) than between the Time and verb-VAC frequency (R^2_m = .241). Further, there was no relationship between VAC frequency and Time, which preliminarily suggests that the variance accounted for by verb-VAC frequency may be attributable to main verb frequency. To investigate this hypothesis, post-hoc partial correlations were conducted using the ppcor package (Kim & Kim, 2015) in R. The results, which are summarized in Table 21, indicate a moderate relationship ($r = -.495$) between Time and verb-VAC frequency. However, when we control for the influence of main verb frequency on the relationship between Time and verb-VAC frequency, the relationship between Time and verb-VAC frequency is negligible ($r = -.029$). Conversely, when we control for the influence of verb-VAC frequency on the relationship between Time and main verb frequency, the relationship between Time and main verb frequency is still meaningful ($r = -.537$). These results suggest that the learners in this study exhibited more sophisticated verb use, but otherwise did not exhibit changes in sophisticated verb-VAC use (at least as measured by mean frequency and strength of association scores).

TABLE 21. Correlations between VAC frequency, main verb frequency, and Time

	R	*p*
Time and main verb frequency	−0.678	<0.001
Time and verb – VAC frequency	−0.495	<0.001
main verb frequency and verb – VAC frequency	0.706	<0.001
Time and main verb frequency while controlling for verb – VAC frequency	−0.537	<0.001
Time and verb – VAC frequency while controlling for main verb frequency	−0.029	0.834

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

RQ3 Results: Syntactic Complexity Versus VAC Sophistication Indices

To explore the degree to which syntactic complexity and VAC sophistication indices were complementary, we used both types of indices to predict the time point at which a text was produced (Time). However, we only selected those indices that showed unique meaningful relationships with Time in the previous analyses (DC/C and main-verb frequency).

The linear model with Time as the dependent variable reported significant main effects for both DC/C and main verb frequency with a large effect ($p < .001$, $R^2_{adjusted} = .583$). A summary of the model can be found in Table 22.

RQ3 Discussion: Syntactic Complexity Versus VAC Sophistication Indices

The results indicated that main verb frequency and DC/C contributed to a model of development in a complementary manner. A linear model using dependent clauses per clause and main verb frequency to predict the time point at which an essay was produced (Time) accounted for 58.3% of the variance. A follow-up analysis for the relative importance of each index to the model indicated that main verb frequency and DC/C contributed to the model to a similar degree. As the participants in the study spent time studying English, they tended to use less frequent main verbs and more dependent clauses in their writing. This was in line with Verspoor et al.'s cross-sectional study (2012) on similar learners as most frequent word types decreased across levels as determined by text scores.

CONCLUSION

This study investigates the degree to which indices of VAC sophistication, which previously were shown to be predictive of writing proficiency, can also be used to model language development over time. Most previous studies in the fields of SLA and SLW have measured syntactic development using indices of syntactic complexity such as MLT, MLC, and DC/C. Recently, the limitations of such indices have been pointed out (Biber et al., 2011; Kyle & Crossley, 2017; Norris & Ortega, 2009) including the difficulty of explaining findings in light of usage-based theories of SLA. Recently, Kyle and Crossley (Kyle, 2016; Kyle & Crossley, 2017) developed VAC-based indices of

TABLE 22. Linear model predicting Time using DC/C and main verb frequency

	Relative importance[a]	Estimate	SE	*t*	*p*
(Intercept)		18.978	3.566	5.321	<0.001
DC/C	0.271	5.369	1.288	4.168	<0.001
main verb frequency	0.327	−0.477	1.345	−4.935	<0.001

[a] The relative importance of the indices in each model was calculated using the calc.relimp() function in the relaimpo package (Grömping, 2006). Specifically, the metric lmg (Lindeman et al., 1980), which takes into account both the direct relationship between the independent and dependent variable (i.e., the bivariate correlation) and the indirect relationship between the independent and dependent variable (i.e., the amount of variance explained when included in a multivariate model) was used.

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

sophistication that operate at a similar level of abstraction (e.g., the clause) as many measures of syntactic complexity but are in line with usage-based theories of language learning (Ellis, 2002; Ellis & Ferreira-Junior, 2009). In a cross-sectional study (2017), Kyle and Crossley found that the VAC indices were much stronger predictors of writing proficiency (as measured by holistic quality scores) than indices of syntactic complexity such as MLT, MLC, and DC/C. While Kyle and Crossley (2017) provided support for the use of VAC indices to predict holistic TOEFL scores, a large body of research has indicated that cross-sectional findings do not always align with developmental trajectories (Bestgen & Granger, 2014; Bulté & Housen, 2014; Crossley & McNamara, 2014; Lowie & Verspoor, 2019).

This study builds on Kyle and Crossley (2017) by investigating the degree to which indices of syntactic complexity and VAC sophistication can model changes in writing for adolescent EFL learners studying English over a 2-year period. These analyses indicated that holistic writing scores increased over the course of the study, suggesting that writing development occurred. Meaningful developmental trends were observed for two indices of syntactic complexity (MLT and DC/C) and for two VAC sophistication indices (main verb frequency and verb-VAC frequency). Follow-up analyses indicated that a single index of syntactic complexity (DC/C) and a single index related to VAC sophistication (main verb frequency) significantly contributed to a model of longitudinal development. As participants spent time studying English, they tended to use more dependent clauses (RQ1) and less frequent main verbs (RQ2). The former results support commonly observed trends in studies of L2 writing development (e.g., Byrnes, 2009; Vyatkina, 2012) and some cross-sectional studies (Ortega, 2003; Verspoor et al., 2012). They differ, however, from the findings of Kyle and Crossley (2017), which found no meaningful relationship between DC/C and TOEFL holistic writing scores. The latter results are in line with previous work on lexical sophistication (e.g., Kyle & Crossley, 2015) and generally support usage-based theories of language learning. However, no meaningful trend was observed between VAC frequency or verb-VAC strength of association or frequency measures and Time. This result counters expected usage-based accounts of language learning and previous studies that have indicated such relationships both longitudinally (e.g., Ellis & Ferreira-Junior, 2009) and cross-sectionally (Kyle & Crossley, 2017), though this may have been due to the potential discrepancy between the input of the target learners in the present study and the written section of COCA, which was used to calculate VAC indices. Overall, the results of this study support the use of linguistic indices related to both clausal subordination (i.e., DC/C) and VAC use (i.e., main verb frequency) to model longitudinal development in lower proficiency L2 learners. Further research is needed to determine the degree to which age, educational context, instructional content, and register (among other factors) contributed to these findings (see limitations in the following text).

LIMITATIONS AND FUTURE DIRECTIONS

The current study had a number of limitations that future studies should address. First, the written sections (academic, fiction, magazine, and newspaper) of COCA may not have accurately represented the types of input that the participants were exposed to inside and outside of the classroom, which may affect the inferences drawn from the results.

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

Relatedly, the learner corpus does not include information regarding what the participants were learning in their language classes (L1 or L2). If possible, future research should attempt to control for these issues. Second, although topic was controlled for at each collection point (i.e., all students wrote on the same topic during a particular collection point), topics changed at each successive collection point. While this choice was made to maintain ecological validity, it may have affected the results, though no obvious systematic differences were identified for particular prompts in the plots. Nonetheless, future studies should consider using a counterbalanced design as is possible. Third, in this study we attempted to fit linear models, which has been the norm in L2 studies for some time. However, this is not the only way to measure longitudinal development (e.g., de Bot et al., 2013; Larsen-Freeman & Cameron, 2008), and indeed the plots indicate that the data investigated in this study were not strictly linear. In future studies it would be fruitful to additionally investigate nonlinear and complementary developmental trends. Fourth, this study examined a small set of large-grained syntactic complexity indices and a small set of VAC features, but ignored other important features of development such as accuracy and fluency and other linguistic features such as lexical diversity, all of which may explain important variance in growth beyond syntactic complexity or sophistication features. This study examined a small number of students in a particular setting and a particular style of writing. To confirm the generalizability of the observed trends, replication studies are needed that represent a variety of learning contexts and language production types. Further, our study included relatively short texts (the mean essay length was approximately 182 words). To our knowledge, no study has attempted to identify the length of text needed to obtain a reliable score for indices of syntactic complexity or VAC sophistication. This is an important area for future research. A final potential limitation is the use of automated tools for the identification of linguistic features. The use of such tools enable large amounts of data to be analyzed (e.g., the 360-million-word section of COCA analyzed in this study) and identification accuracy has improved markedly over the past 10 years. However, identification accuracy is still less than perfect (ranging from 81.25% for L2 VACs to 97.0% for L2 main verbs), which introduces error into any downstream analysis. Future research should further investigate methods of increasing parsing and tagging accuracy on L2 texts (e.g., Meurers & Dickinson, 2017).

NOTES

[1] Kyle and Crossley (2017) referred to these as indices of *syntactic sophistication*. However, as an anonymous reviewer pointed out, this term is not sufficiently precise (and indeed may be misleading). The indices introduced by Kyle and Crossly (2017) comprise indices of VAC (reference corpus) frequency, main verb (reference corpus) frequency, and the strength of association between VACs and the verbs that are used with them. Therefore, although some of these indices focus on the (reference corpus) frequency of syntactic forms (without lexical considerations), most of the indices are lexicogrammatical in nature (and not strictly syntactic).

[2] Note that various terms have been used to refer to this construct, including *phrasal sophistication*, *phrasal complexity*, and simply *collocation* (among others).

[3] For example, in purely syntactic terms, a VAC is a clause. However, usage-based perspectives are concerned with the frequency of the VAC, the frequency of the main verb, the strength of association between the VAC, and the particular verbs it co-occurs with, and not (necessarily) with the structural complexity of the clause.

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

[4] Lu (2010) found that identification of all structures achieved an F-score of above .900 with the exception of complex nominals, which achieved an F-score of .830. Automated complexity scores were strongly correlated with manually calculated complexity scores (MLC, $r = .941$; MLT, $r = .989$; DC/C, $r = .851$; CP/C, $r = .834$; CN/C, $r = .867$).

[5] Chen and Manning (2014) report that the Stanford Neural Network Dependency Parser achieves 89.7% labeled dependency accuracy on the Penn Treebank with Stanford Dependencies. We manually checked a random sample consisting of 10% of the VACs identified in the learner corpus. This analysis indicated that 97.9% of the main verbs and 81.25% of the verb-VAC combinations were correctly identified. In most cases of misidentification, the core VAC was correctly identified, but was missing a dependent (e.g., an adverb modifier). While the accuracy with L2 data is slightly lower than expected with L1 data, the VAC identification errors between the L1 and L2 corpora should be similar in nature because the same parser was used for each.

[6] The Fisher–Yates exact test is calculated as: $p_{\text{observed distribution}} = \frac{\left(\frac{a+c}{a}\right)*\left(\frac{b+d}{b}\right)}{\frac{N}{a+b}} + \Sigma\, p_{all\ more\ \text{extreme distributions}}$. The transformation introduced by Gries et al. (2005) uses the negative base of ten logarithm of this p value to index association strength.

[7] In these cases, the LME model reports a "singular fit" indicating that the random effects could not be calculated. In complex mixed-effects models, this can be due to model overfitting (i.e., using too many variables and/or using collinear variables). In simple models (such as the ones reported in the present study), this indicates that the observed differences between participants were not large enough to be fit by the model. Increasing sample size (and consequently, power) will increase the chances of fitting between participant differences. However, power is likely not a large issue in the results of this study given that particularly small between-participant differences were found in some of the analyses (see, e.g., Table 10). If larger samples were obtained (and similar patterns were observed), then fitted differences would likely still be very small.

REFERENCES

Barton, K. (2013). *MuMln: Multi-model inference*. *R package, version*, 1.9.

Bates, D. M. (2010). *lem4: Mixed-effects Modeling with R*. Unpublished Manuscript.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28–41. https://doi.org/10.1016/j.jslw.2014.09.004.

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

Biber, D., Gray, B., & Staples, S. (2014). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, *37*, 639–668. https://doi.org/10.1093/applin/amu059.

BNC Consortium. (2007). *The British National Corpus, version 3*. BNC Consortium. http://www.natcorp.ox.ac.uk/.

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

Bulté, B., & Housen, A. (2018). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics*, *28*, 147–164. https://doi.org/10.1111/ijal.12196.

Byrnes, H. (2009). Emergent L2 German writing ability in a curricular context: A longitudinal study of grammatical metaphor. *Linguistics and Education*, *20*, 50–66.

Chen, D., & Manning, C. D. (2014). *A fast and accurate dependency parser using. neural networks.*, 740–750. https://cs.stanford.edu/~danqi/papers/emnlp2014.pdf.

Crossley, S. A., Kyle, K., & Salsbury, T. (2016). A usage-based investigation of L2 lexical acquisition: The role of input and output. *The Modern Language Journal*, *100*, 702–715.

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

Crossley, S. A., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, *60*, 573–605. https://doi.org/10.1111/j.1467-9922.2010.00568.x.

Crymes, R. (1971). The relation of study about language to language performance: With special reference to nominalization. *TESOL Quarterly*, *5*, 217–230.

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

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

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*, 159–190. https://doi.org/10.1075/ijcl.14.2.02dav.

de Bot, K., Lowie, V., Thorne, S., & Verspoor, M. (2013). Dynamic systems theory as a comprehensive theory of second language development. In M. P. G. Mayo, M. J. G. Mangado, & M. M. Adrián (Eds.), *Contemporary approaches to second language acquisition* (pp. 199–220). John Benjamins Publishing.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL-International Review of Applied Linguistics in Language Teaching*, *47*, 157–177.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*, 143–188.

Ellis, N. C., & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, *7*, 188–221. https://doi.org/10.1075/arcl.7.08ell.

Eskildsen, S. W. (2009). Constructing another language—Usage-based linguistics in second language acquisition. *Applied Linguistics*, *30*, 335–357. https://doi.org/10.1093/applin/amn037.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *144*, 285–307.

Garner, J., Crossley, S., & Kyle, K. (2019). N-gram measures and L2 writing proficiency. *System*, *80*, 176–187. https://doi.org/10.1016/j.system.2018.12.001.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.

Gries, S. T. (2015). The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*, *10*, 95–125.

Gries, S. T., Hampe, B., & Schönefeld, D. (2005). Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, *16*, 635–676. https://doi.org/10.1515/cogl.2005.16.4.635.

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software*, *17*, 1–27.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels. NCTE Research Report No.*, *3*. http://eric.ed.gov/?id=ED113735.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983–997.

Kim, S., & Kim, M. S. (2015). Package “ppcor.” *Communications for Statistical Applications and Methods*, *22*, 665–674.

Knoch, U., Rouhshad, A., & Storch, N. (2014). Does the writing of undergraduate ESL students develop after one year of study in an English-medium university? *Assessing Writing*, *21*, 1–17. https://doi.org/10.1016/j.asw.2014.01.001.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). lmerTest: Tests in linear mixed effects models. *R package version 2.0–20*.https://Cran.Rproject.Org/Web/Packages/LmerTest

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication*. Georgia State University. http://scholarworks.gsu.edu/alesl_diss/35/.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*, 757–786. https://doi.org/10.1002/tesq.194.

Kyle, K., & Crossley, S. A. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, *34*, 513–535.

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, *102*, 333–349. https://doi.org/10.1111/modl.12468.

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

Larsen-Freeman, D., & Cameron, L. (2008). Research methodology on language development from a complex systems perspective. *The Modern Language Journal*, *92*, 200–213.

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

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

Laufer, B., & Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307–322. https://doi.org/10.1093/applin/16.3.307.

Lindeman, R. H., Merenda, P., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis* (p. 119). Foresman and Company.

Lowie, W. M., & Verspoor, M. H. (2019). Individual differences and the ergodicity problem. *Language Learning*, *69*, 184–206.

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

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

Meurers, D., & Dickinson, M. (2017). Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning*, *67*, 66–95.

Murakami, A., & Alexopoulou, T. (2016). Longitudinal L2 development of the English article in individual learners. In D. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 1050–1055). Cognitive Science Society.

Ninio, A. (1999). Pathbreaking verbs in syntactic development and the question of prototypical transitivity. *Journal of Child Language*, *26*, 619–653.

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

O'Donnell, M. B., & Ellis, N. (2010). *Towards an inventory of English verb argument constructions* (pp. 9–16). http://dl.acm.org/citation.cfm?id=1866732.1866734.

O'Grady, W. (2008). Innateness, universal grammar, and emergentism. *Lingua*, *118*, 620–631.

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.

Penris, W., & Verspoor, M. (2017). Academic writing development: A complex, dynamic process. In S. Pfenniger & Navracsics (Eds.), *Future Research Directions for Applied Linguistics* (Vol. *109*, pp. 215–242). Multilingual Matters Ltd.

R Core Team. (2016). *A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Römer, U., O'Donnell, M. B., & Ellis, N. C. (2015). Using COBUILD grammar patterns for a large-scale analysis of verb-argument constructions. N. Groom, M. Charles, & J. Suganthi (Eds.), (Vol. *73*, p. 43). John Benjamins Publishing Company.

Römer, U., Roberson, A., O'Donnell, M. B., & Ellis, N. C. (2014). Linking learner corpus and experimental data in studying second language learners' knowledge of verb-argument constructions. *ICAME Journal*, *38*, 115–135.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, *6*, 309–316.

Thornhill, D. E. (1969). *A quantitative analysis of the development of syntactical fluency of four young adult Spanish speakers learning English* [Unpublished doctoral dissertation]. The Florida State University.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Verspoor, M. (2017). Complex Dynamic Systems Theory and L2 pedagogy. In L. Ortega & Z. Han (Eds.), *Complexity Theory and Language Development: In celebration of Diane Larsen-Freeman* (Vol. *48*, pp. 143–162). John Benjamins.

Verspoor, M., & Behrens, H. (2011). Dynamic systems theory and a usage-based approach to second language development. *A dynamic approach to second language development: Methods and techniques* (pp. 25–38).

Verspoor, M., de Bot, K., & Xu, X. (2015). The effects of English bilingual education in the Netherlands. *Journal of Immersion and Content-Based Language Education*, *3*, 4–27.

Verspoor, M., Schmid, M. S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, *21*, 239–263. https://doi.org/10.1016/j.jslw.2012.03.007.

Verspoor, M., & Smiskova, H. (2012). Foreign language writing development from a dynamic usage-based perspective. In R. M. Manchón (Ed.), *L2 writing development: Multiple perspectives* (pp. 17–46). De Gruyter.

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal*, *96*, 576–598.

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press

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

Yates, F. (1934). Contingency tables involving small numbers and the $\chi 2$ test. *Supplement to the Journal of the Royal Statistical Society*, *1*, 217–235. https://doi.org/10.2307/2983604.

APPENDIX

HOLISTIC SCORING RUBRIC

The 0 represents the very beginning level where English is only barely emerging. There is usually very little text, and if there is text, it is mainly Dutch. Very simple sentence structure with many Dutch words and some English words thrown in, often misspelled.

The 1 represents English that has emerged to some degree. The language used is almost all English, with only a few Dutch words, but the language is simple, with mainly simple sentences, present tenses, often Dutch word order and Dutch expressions literally translated. Full of little errors.

The 2 represents English that has emerged. The English is still quite simple, simple sentence structure, simple tenses, an attempt at some creativity in vocabulary and syntax; the English may contain a Dutchism here and there, but it is mainly English. There are still many errors.

The 3 represents English that has emerged. The English is still quite simple with simple and compound sentences, but one or two dependent clauses may appear. There are mainly simple present and past tenses, but an occasional progressives or passive may appear. There is an attempt at some creativity in the vocabulary and syntax; even though the English still contains a few standard Dutchisms, there are also some authentic English collocations and expressions. There are still some errors.

The 4 represents English that has more variety in sentence structures (a few dependent clauses), some variety in tenses (past, future, progressive, passive, and use of modals). There are some authentic English collocations and expressions even though the English still contains a few standard Dutchisms. Some longer sentences, less choppy. There are still some errors, but mainly in mechanics and spelling.

The 5 represents English that has more variety in sentence structure (dependent clauses and nonfinite structures), variety in tenses (past, future, progressive, passive, and use of modals) where needed. There are several authentic English collocations and expressions, but there may also be a few standard Dutchisms. There are still some errors, but mainly in mechanics and spelling. The language flows.

The 6 represents English that has nativelike variety in sentence structure with dependent clauses and nonfinite structures, shows nativelike flexibility in time/tense/mood/voice. It contains many authentic English collocations and expressions, but there are still one or two Dutchisms. There are still some errors, but mainly in mechanics and spelling.

The 7 represents English that has a nativelike variety in sentence structure with dependent clauses and nonfinite structures, shows nativelike flexibility in time/tense/mood/voice. It contains mostly authentic English collocations and expressions, but there still be a Dutchism here and there. There are still some errors, but mainly in mechanics and spelling.

https://doi.org/10.1017/S0272263120000546 Published online by Cambridge University Press