Assessing syntactic sophistication in L2 writing: A usage-based approach

Kristopher Kyle
University of Hawaii at Manoa, USA

Scott Crossley
Georgia State University, USA

Abstract

Over the past 45 years, the construct of syntactic sophistication has been assessed in L2 writing using what Bulté and Housen (2012) refer to as absolute complexity (Lu, 2011; Ortega, 2003; Wolfe-Quintero, Inagaki, & Kim, 1998). However, it has been argued that making inferences about learners based on absolute complexity indices (e.g., mean length of t-unit and mean length of clause) may be difficult, both from practical and theoretical perspectives (Norris & Ortega, 2009). Furthermore, indices of absolute complexity may not align with some prominent theories of language learning such as usage-based theories (e.g., Ellis, 2002a,b). This study introduces a corpus-based approach for measuring syntactic sophistication in L2 writing using a usage-based, frequency-driven perspective. Specifically, novel computational indices related to the frequency of verb argument constructions (VACs) and the strength of association between VACs and the verbs that fill them (i.e., verb–VAC combinations) are developed. These indices are then compared against traditional indices of syntactic complexity (e.g., mean length of T-unit and mean length of clause) with regard to their ability to model one aspect of holistic scores of writing quality in Test of English as a Foreign Language (TOEFL) independent essays. Indices related to usage-based theories of syntactic development explained greater variance (R^2 = .142) in holistic scores of writing quality than traditional methods of assessing syntactic complexity (R^2 = .058). The results have important implications for modeling syntactic sophistication, L2 writing assessment, and AES systems.

Corresponding author:
Kristopher Kyle, University of Hawaii at Manoa, Moore Hall 587, 1890 East-West Road, Honolulu, HI 96822, USA.
Email: kristopherkyle1@gmail.com

Linguistic features are often used in language performance assessments to investigate relationships between performance and target language use (Biber & Gray, 2013; Biber, Gray, & Staples, 2014; LaFlair & Staples, 2017) and to determine whether different tasks elicit distinct linguistic features (Cumming et al., 2005; Kyle, Crossley, & McNamara, 2015; Yang, Lu, & Weigle, 2015). Linguistic features are also used to develop automatic scoring (AES) models that link features to rater judgments (Attali & Burstein, 2006; Burstein, Tetreault, & Madnani, 2013; Enright & Quinlan, 2010; Guo, Crossley, & McNamara, 2013). Arguably, these linguistic features should be related to theories of language development and/or use (Bachman & Cohen, 1999; Chapelle, Enright, & Jamieson, 2011; Chodorow & Burstein, 2004; Deane, 2013).

One component of language that has been of interest to both writing assessment and language acquisition researchers is syntactic sophistication (Bulté & Housen, 2012; Knoch, Rouhshad, & Storch, 2014; Lu, 2011; Norris & Ortega, 2009). Over the past 45 years syntactic sophistication, which is analogous to what Bulté and Housen (2012) refer to as *relative complexity* (i.e., the relative difficulty of learning, using, and/or comprehending a particular structure) has been primarily operationalized in terms of *absolute complexity* (i.e., text internal, formal syntactic features such as the number of words per T-unit). This practice has been supported by studies that have generally shown that higher proficiency learners produce more complex syntactic structures (Cumming et al., 2005; Larsen-Freeman, 1978; Lu, 2011; Ortega, 2003; Wolfe-Quintero et al., 1998). The use of measures related to absolute complexity, however, may not be relevant to some widely held theoretical perspectives in language acquisition, namely usage-based theories of language learning (e.g., Ellis, 2002a; Langacker, 1987; Tomasello, 2003).

Usage-based theories of language development suggest that frequency, not absolute complexity, is a key driver in language learning (Ellis, 2002a; 2002b). From a usage-based perspective, frequent form–meaning pairs called constructions (Goldberg, 1995), are learned earlier/more easily than less frequently encountered constructions (e.g., Ellis & Ferreira-Junior, 2009a). Less frequently encountered constructions are therefore likely to appear in the language produced by more proficient speakers. In language testing, usage-based approaches are commonly operationalized at the lexical level as evidenced by the use of corpus-based word frequency features in both the evaluation of test takers' performance on writing tasks (e.g., Cumming et al., 2005) and the development of automatic essay scoring (AES) models (Attali & Burstein, 2006; Guo et al., 2013). However, usage-based perspectives can extend beyond word-level constructions to phrasal constructions (Ellis, 2002a; Goldberg, 1995; Tomasello, 2003) as seen in more recent assessment research that has reported strong links between scores of writing quality and the degree of test takers' use of certain multi-word units identified through corpus-based research (Bestgen & Granger, 2014; Crossley, Cai, & McNamara, 2012; Kyle & Crossley, 2015). Beyond phrases, a growing body of research also supports links between the development of verb argument construction use (i.e., syntactic development) and the frequency of those structures in a learner's input (Ellis & Ferreira-Junior, 2009b; Lieven, Pine, & Baldwin, 1997; Ninio, 1999).

In this study, we introduce an approach to assessing syntactic sophistication in L2 writing using computational indices inspired by usage-based theories of language learning. Specifically, we develop and test indices related to the frequency of verb argument

constructions (VACs; e.g., *subject-verb-indirect object-direct object*), main verb lemmas (e.g., *to give*), and verb–VAC combinations (e.g., *subject-**to give**-indirect object-direct object*). Additionally, we develop indices related to the strength of association of verb–VAC combinations (i.e., the probability that a VAC and a main verb lemma will co-occur). The utility of the indices are tested by assessing their ability to predict holistic scores of writing quality in Test of English as a Foreign Language (TOEFL) independent essays when compared against traditional indices of syntactic complexity (e.g., mean length of T-unit and mean length of clause).

Traditional indices of syntactic complexity

Complexity has been an important construct in first language (L1) and second language (L2) development and assessment for the past 45 years. Larsen-Freeman (1978), drawing on previous work in L1 development (Hunt, 1965), cited complexity as one of three important constructs of language development (in addition to accuracy and fluency). At the syntactic level, complexity has generally been conceptualized as a text-internal formal characteristic (i.e., absolute complexity), and has been operationalized with regard to clausal subordination and/or sentence length. Many different indices have been used, but large-grained indices such as mean length of T-unit (MLTU) and mean length of clause (MLC), and clauses per T-unit (C/TU) have been used with the most consistency (Ortega, 2003; Wolfe-Quintero et al., 1998).

In L2 writing development and assessment research, a relatively consistent positive relationship has been found between indices such as MLTU, C/TU, and MLC and both general language proficiency (as measured by program level or external test criteria) and writing quality scores (Cumming et al., 2005; Lu, 2011; Ortega, 2003; Wolfe-Quintero et al., 1998). The results indicate that as L2 writers become more proficient language users and earn higher writing quality scores, they tend to write longer T-units and clauses and tend to use more clauses per T-unit (c.f. Knoch et al., 2014). In language assessment, for example, Cumming et al. (2005) used MTLU and C/TU to measure syntactic complexity in test takers' performance on prototype TOEFL writing tasks. They found that higher rated essays tended to include more words per T-unit than lower rated essays, and, for some essay types, higher rated essays also tended to include T-units with more clauses.

Although large-grained indices of syntactic complexity have been used prominently (and somewhat successfully) to measure sophistication in L2 writing development and assessment studies, a number of limitations have been raised (Biber, Gray, & Poonpon, 2011; Bulté & Housen, 2012; Larsen-Freeman, 2009; Norris & Ortega, 2009). Biber et al. (2011), for example, provide corpus-based evidence that suggests phrasal (not clausal) complexity is a feature of academic writing. Larsen-Freeman (2009) and Norris & Ortega (2009) suggest that large-grained clausal indices also obscure the linguistic variation that occurs in language development, and if used for assessing test performance, may contribute to misleading inferences about test takers. Additionally, and of critical importance, is the lack of a cohesive theoretical rationale for the use of indices of syntactic complexity (Bulté & Housen, 2012; Norris & Ortega, 2009). A lack of strong theoretical rationale for the use of particular measurement instruments weakens claims related to construct validity (Chapelle, 1999) and limits the inferences that can be made (Norris & Ortega, 2009).

Usage-based perspectives on syntactic development

A usage-based perspective may be useful for addressing many of the limitations of the use of traditional indices of syntactic complexity to measure syntactic sophistication in test takers performance. Usage-based perspectives to language acquisition posit that language learning is no different from other types of experiential human learning (Bybee, 2006; Langacker, 1987; Tomasello, 2003) in that repeated language experiences mixed with a combination of two human cognitive abilities (intention reading and pattern finding) allow children to acquire, over time, the language system of the adults with whom they interact (Tomasello, 2003). In lay terms, language that is encountered more frequently will be learned earlier and more easily. Starting in the 1990s, usage-based theories of language acquisition began to be empirically investigated in first language (L1) acquisition (e.g., Goldberg, Casenhiser, & Sethuraman, 2004; Lieven, Pine, & Baldwin, 1997; Tomasello & Brooks, 1999). By the early 2000s, usage-based perspectives began to gain traction in the field of second language (L2) acquisition as well (Ellis, 2002a, 2002b).

Usage-based perspectives posit that all linguistic forms (e.g., words, phrases, syntactic patterns, etc.), which are called *constructions* (Goldberg, 1995), are functional form–meaning pairings. The unit of investigation in most construction research has been verb argument constructions (VAC), which consist of a verb slot and the arguments it takes. A ditransitive construction, for example, includes a subject, a main verb, an indirect object, and a direct object, such as in the sentence $She_{subject}$ $spugged_{verb}$ $him_{indirect\_object}$ $something\text{-}direct\_object$. Research has suggested that VACs are not merely bare forms on which to place words, but carry meaning for both L1 (Bencini & Goldberg, 2000; Chang, Bock, & Goldberg, 2003; Hare & Goldberg, 1999) and L2 (Gries & Wulff, 2005) users. The ditransitive construction above, for example, carries a meaning of literal or metaphorical transfer even in the absence of a transfer-related verb (such as *give*) (Goldberg, 2013). This allows proficient users of English to interpret the nonsense verb *spugged* in the example above as something related to transference.

L1 studies suggest that VACs (and other constructions) are learned as a function of frequency (Goldberg et al., 2004; Lieven et al., 1997; Ninio, 1999). VACs are generally first used by language learners with a single prototypical "pathbreaking" verb before they learn that other verbs can also be used (Ninio, 1999). Low-variance, high frequency experiences with constructions allow language learners to generalize syntactic frames (and therefore overcome the "poverty of stimulus" that some linguists theorize challenges language learners). L2 research has also suggested that input frequency is a key component of language learning (Ellis & Ferreira-Junior, 2009a, 2009b) and that construction knowledge progresses from fixed (e.g., *She kicked him the ball*) to schematic (i.e., $NP_{\_subject}\text{-}Verb\text{-}NP_{indirect\ object}\text{-}NP_{direct\_object}$) constructions (Eskildsen, 2009; Eskildsen & Cadierno, 2007).

An important limitation of many usage-based studies is the small amount of input/output data analyzed. It has been common practice to analyze small datasets because it is extremely time consuming (and therefore cost prohibitive) to analyze syntactic structures in learner data (let alone in large reference corpora). Ideally, usage-based research would involve recording all of the language experiences to which a child or L2 learner is exposed. However, for the practical reasons discussed above, this does not happen, and,

as a result, most usage-based studies involve the analysis of only a few hours of interactions. Even the most data-rich studies such as that of Ellis and Ferriera Junior (2009a) still only include a very small percentage of the language experiences a learner has, and are usually restricted to a particular context/register (e.g., an informal interview). One alternative to the analysis of snippets of a learner's linguistic experience is to use a reference corpus as a proxy for the experiences of a learner. Recent research has indicated strong links between reference corpus frequencies and VAC knowledge of L1 users and advanced L2 users (Römer, O'Donnell, & Ellis, 2015; Römer, Roberson, O'Donnell, & Ellis, 2014). For instance, Römer et al. (2014) found similarities between the frequency of verbs used in a variety of subject-verb-prepositional phrase (SVPP; e.g., *subject-verb-across*) VACs in learner corpus data and in survey data collected from advanced L2 English users. In a similar study, Römer et al. (2015) compared the verbs produced in SVPP constructions by advanced L1 German learners of L2 English, L1 English speakers (via a gap-filling exercise), and their relative frequencies in the British National Corpus (BNC; BNC Consortium, 2007). Although the results varied by construction type, the verbs produced for each VAC by the L1 and L2 users of English were generally highly correlated both with each other and with corpus frequencies.

A potential solution to overcome the small sample sizes found in most VAC studies is to automate VAC identification. While computers have been able to automatically and accurately extract word and phrase frequencies for some time, until recently this has not been the case with syntactic features. Some advances in the automatic analysis of language data used to investigate learners' syntactic development have been made (Biber et al., 2004; Lu, 2011; O'Donnell & Ellis, 2010), but no approaches developed so far have been able to identify and account for all VACs in a corpus of learner language. O'Donnell, Ellis, and Römer (O'Donnell & Ellis, 2010; Römer et al., 2015) have conducted pioneering work in this area using the Robust Accurate Statistical Parsing (RASP; Briscoe, Carroll, & Watson, 2006) system. However, because of the relatively low parsing accuracy (around 75%) of RASP, only a limited number of VAC frequency profiles were extracted. Recent advances in syntactic parsing accuracy (i.e., Chen & Manning, 2014), however, hold promise for automating VAC analyses. Specifically, state of the art parsers now achieve labeling accuracies of around 90% and allow for the identification of the types of syntactic dependencies represented by VACs.

Current study

This study introduces a computational approach for the analysis of syntactic sophistication in L2 writing that aligns with usage-based theories of language learning. This approach includes the automatic calculation of a number of lexicogrammatical features related to a reference corpus' frequency of main verb lemmas and VACs and the strength of association between VACs and the verbs that fill them. Such an approach contrasts with traditional approaches to the analysis of syntactic sophistication that consider the text internal and formal features of L2 texts (i.e., absolute complexity) such as the mean length of T-unit (MLTU) and the mean length of clause (MLC). In developing new indices of syntactic sophistication, we hope to address some of the criticisms that have been raised regarding the use of syntactic complexity indices while providing methods for

analyzing target language sophistication in test takers' and learners' performance in various domains and tasks that align with usage-based theories of language learning. In this study, we compare the performance of newly developed VAC indices with traditional syntactic complexity indices for predicting holistic scores of writing quality on an essay writing task.

This study is guided by the following research questions:

1. What is the relationship between syntactic complexity indices and holistic scores of writing quality?
2. What is the relationship between usage-based indices of syntactic sophistication and holistic scores of writing quality?
3. Are there differences between the syntactic complexity indices and usage-based indices of syntactic complexity in terms of explaining holistic scores of writing quality?

Method

To investigate the three research questions, we examined whether linear relationships existed between human ratings of L2 essay quality and traditional indices of syntactic sophistication (i.e., indices of syntactic complexity; Research Question 1) and usage-based indices of syntactic sophistication (i.e., the VAC indices; Research Question 2), and then compared the strength of those relationships. Stepwise multiple regression statistics were used to model the relationships between indices related to each research question and holistic TOEFL independent writing scores and comparisons were made between models to assess differences in accuracy. The data and indices are described below.

Learner corpus

The writing quality corpus we selected comprises 480 argumentative essays included in the TOEFL Public Use Dataset. Essays were collected by the Educational Testing Service (ETS) during operational administrations of the TOEFL. The essays comprise responses to two independent prompts (240 texts for each prompt) that ask test takers to compose an essay that asserts and defends an opinion on a particular topic based on life experience (see Table 1). Test takers are given 30 minutes to complete the writing task and are expected to produce at least 300 words. See Table 2 for the summary statistics of scores test takers obtained on the essays included in this corpus.

Each essay was given a score on a 5-point scale by at least two raters trained by the Educational Testing Service (ETS). If the scores given by the raters differed by 1 point or less, scores were averaged. If any two scores given by raters differed by more than 1 point, a third rater was used to adjudicate the score. Scores range from 1.0 to 5.0 in .5-point intervals. The holistic rating score used included descriptors related to the completion of the task, organization, development of ideas, coherence, word use, and syntax. See Table 3 for the score descriptors for low- and high-quality essays.[1]

Table 1. Writing prompts for independent essays in TOEFL public use dataset.

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

Table 2. Overview of writing quality corpus.

Prompt	N	Number of words	Min score	Max score	Mean score	Standard deviation	Skewness	Kurtosis
1	240	77,238	1	5	3.383	0.864	0.055	2.353
2	240	74,252	1	5	3.471	0.910	−0.068	2.419
Combined	480	151,490	1	5	3.427	0.887	−0.003	2.387

Table 3. Abbreviated TOEFL rubric for independent writing tasks.

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

Indices of syntactic sophistication

The indices employed in this study are calculated by the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) version 1.0 (Kyle, 2016). TAASSC is an easy to use text analysis tool that is freely available[2] and works on most operating systems (Windows and Mac). TAASSC features a simple user interface and requires no programming knowledge to operate. See Figure 1 for a screenshot of TAASSC.
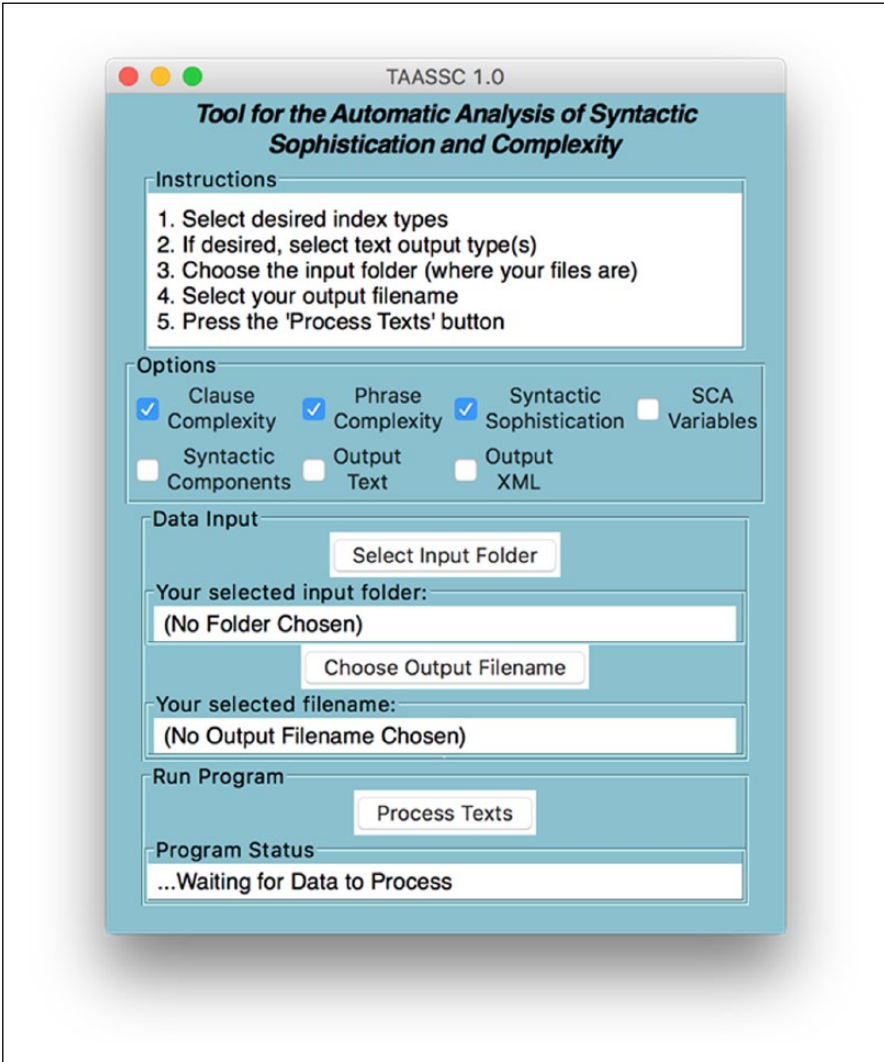
Figure 1. The TAASSC GUI.

Syntactic complexity indices. To address the first research question, we compute traditional syntactic complexity measures using the indices reported by Lu's (2010) L2 syntactic complexity analyzer (SCA). These indices are also reported by TAASSC. SCA includes 14 indices of syntactic complexity drawn from those identified by Wolfe-Quintero et al. (1998) and Ortega (2003) inter alia. Table 4 includes a description of each of the structures counted by SCA, and Table 5 comprises a list of the 14 SCA indices including a short description of each. For further information, refer to Lu (2010, 2011).

Table 4. A description of syntactic structures counted by SCA.

Structure	Description	Examples
Word	A sequence of letters that are bounded by white space	*I* *ate*
Verb phrase	A finite or non-finite verb phrase that is dominated by a clause marker	*ate pizza* *was hungry*
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*
Complex T-unit	A T-unit that includes a dependent clause	*I ate pizza because I was hungry*
Sentence	A group of words bounded by sentence-ending punctuation (., ?, !, ", ...)	*I went running today.*

Note: Adapted from Lu (2010, pp. 7–13).

Table 5. A description of SCA variables.

Index abbreviation	Index name	Index description
MLS	Mean length of sentence	Number of words per sentence
MLT	Mean length of T-unit	Number of words per T-unit
MLC	Mean length of clause	Number of words per clause
C/S	Clauses per sentence	Number of clauses per sentence
VP/T	Verb phrases per T-unit	Number of verb phrases per sentence
C/T	Clauses per T-unit	Number of clauses per T-unit
DC/C	Dependent clauses per clause	Number of dependent clauses per clause
DC/T	Dependent clauses per T-unit	Number of dependent clauses per T-unit
T/S	T-units per sentence	Number of T-units per sentence
CT/T	Complex T-unit ratio	Number of complex T-units divided by T-units
CP/T	Coordinate phrases per T-unit	Number of coordinate phrases per T-unit
CP/C	Coordinate phrases per clause	Number of coordinate phrases per clause
CN/T	Complex nominals per T-unit	Number of complex nominals per T-unit
CN/C	Complex nominals per clause	Number of complex nominals per clause

Table 6. Main verb lemma frequencies in the written section of COCA.

Rank	Frequency (per million)	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

VAC indices. To address the second research question, we use VAC-centered indices of syntactic sophistication calculated by TAASSC that reflect usage-based perspectives of language acquisition (Ellis, 2002a; Goldberg, 1995; Langacker, 1987). For all VAC indices, the Corpus of Contemporary American English (COCA; Davies, 2010) was used as a reference corpus and as a proxy for language experience. COCA is a large (~450 million words), recently collected (includes texts from 1990 to present) corpus that is subdivided into five registers (academic, fiction, magazine, newspaper, and spoken). TAASSC includes indices based on frequency profiles from the combination of all written registers (i.e., academic, fiction, magazine, and newspaper) and each individual written register, respectively. Main verb lemma frequencies, VAC frequencies, and verb–VAC combination frequencies were extracted from COCA with a Python script developed by the authors. The script used the Stanford neural network dependency parser (Chen & Manning, 2014), which is a fast, highly accurate parser (90% labeling accuracy for L1 data) to initially process each text. The script performed a number of functions that identified and extracted each main verb and all direct dependents of that verb. Frequency profiles were then compiled, resulting in comprehensive frequency lists for main verb lemmas (see Table 6), VACs (see Table 7), and verb–VAC combinations (see Table 8). These frequency profiles are also used to create indices that measure the strength of association between VACs and main verb lemmas (Gries & Ellis, 2015). For the current study, indices related to the academic register were used. Each index type used in the current study is described below.

Frequency. TAASSC calculates average frequency scores for main verb lemmas, VACs, and verb–VAC combinations in a target text (in this case the TOEFL writing samples) based on counts derived from COCA. Index variants include mean frequency (both with and without log transformation) scores for tokens and types in a text and standard deviations for these scores. 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. Indices are also calculated that comprise the percentage of main verb lemmas, VACs and main verb lemma – VAC combinations – that occur in a target text occur in the reference

Table 7. Verb argument construction frequencies in COCA.

Rank	Frequency (per million)	Verb argument construction	Most frequent main verb lemma
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

Table 8. Most common verb argument construction–main verb lemma combinations in COCA.

Rank	Frequency (per million)	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	6843.83	be	subordinator - subject – verb – adjective complement	She hears **that he is arrogant**. (news)
4	6318.98	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	5335.93	have	subject – verb – direct object	**Iran has** obvious **interests** in Iraq. (magazine)
6	5124.34	be	verb – nominal complement	That's what's great about **being a teen**. (news)
7	4986.51	be	subordinator - subject – verb – nominal complement	Even before the man reached the car, she knew **that it was Frank**. (fiction)
8	4258.04	be	verb – adjective complement	This is the reason I have found life **to be harder than fiction**… (fiction)
9	3865.16	say	subject – verb – clausal complement	**He said [that health decisions should be made by patients and doctors]** (magazine)
10	3516.17	say	clausal complement – verb – subject	[**"We have an all-new situation now"], says Europol's Storbeck** (magazine)

Table 9. Summary of usage-based frequency indices.

	Main verb lemma frequency	VAC frequency	Verb–VAC combination frequency
Mean token score	✓	✓	✓
Mean token score (log transformed)	✓	✓	✓
Standard deviation token score	✓	✓	✓
Standard deviation token score (log transformed)	✓	✓	✓
Mean type score	✓	✓	✓
Proportion of items attested in corpus	✓	✓	✓
Total	**6**	**6**	**6**

Table 10. Example contingency table used to calculate indices of association strength for the verb "have".

	Construction C (nsubj-v-dobj)	Not construction C (not nsubj-v-dobj)	Totals
Verb V (have)	**a** (212,970)	**b** (991,685)	**a + b** = frequency of V (1,204,655)
Not Verb V (not have)	**c** (1,733,964)	**d** (30,909,494)	**c + d =** combinations that are not V + C (32,643,458)
Totals	**a + c** (1,946,934) frequency of C	**b + d** (37,965,533)	**(a+ b) + (c + d)** = N (total number of VAC tokens in the corpus) = (33,635,143)

Note: Adapted from Gries et al. (2005).

corpus. These indices comprise a rough measure of frequency. In the current analysis, a total of 18 frequency indices based on the frequency norms derived from the academic section of COCA are used (see Table 9).

Association strength. Strength of association indices measure the conditional probability that two items (in this case a main verb lemma and a VAC) will occur together within the COCA corpus. Strength of association has been suggested to supplement frequency in explaining language learning (Ellis & Ferreira-Junior, 2009a, 2009b) because it accounts for respective relative frequency of verbs and constructions. TAASSC calculates three types of association strength measures[3] suggested in the literature including faith (Gries, Hampe, & Schönefeld, 2005), delta P (Ellis & Ferreira-Junior, 2009b), and a variant of collostructional strength (Stefanowitsch & Gries, 2003). In each case, TAASSC calculates mean association strength scores for types and tokens. TAASSC also calculates intra-text standard deviation scores for tokens. Table 10 provides an example of a contingency table used to calculate these three types of association strength measures.

Faith. Faith calculates the conditional probability that a particular verb will occur with a particular VAC (and vice versa). We calculate the probability that a particular VAC X will occur given verb Y (i.e., P(construction|verb) as $\left(\dfrac{a}{a+b}\right)$ (Gries, Hampe, & Schönefeld, 2005). With reference to Table 10, the conditional probability that the transitive (nsubj-v-dobj) construction will be the outcome given the main verb *have* is $\dfrac{212,970}{212,970+991,685} = .177$, indicating that there is a 17.7% chance that the SVO will occur given the main verb *have*. In the current study, we use indices based on the mean faith score for types and tokens in the target text and for the standard deviations of those scores within a target text. These indices are calculated with both the verb and the construction as the cue (six total indices).

Delta P. Delta P is a variant of faith that 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: $\left(\dfrac{a}{a+b}\right)-\left(\dfrac{c}{c+d}\right)$. The delta P value for the outcome of the SVO given the cue *have* is calculated $\left(\left(\dfrac{212,970}{212,970+991,685}\right)=.177\right)-\left(\left(\dfrac{1,733,964}{1,733,964+30,909,494}\right)=.053\right)=.124$. The probability 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 indices based on the mean delta P score for types and tokens in the target text and for the standard deviations of those scores within a target text. These indices are calculated with both the verb and the construction as the cue (six total indices).

Collostructional analysis. Unlike faith and delta P, collostructional analysis (Gries et al., 2005; Stefanowitsch & Gries, 2003) calculates the joint probability (i.e., it is not directional, unlike delta P) that two corpus items will co-occur. Collostructional strength is calculated using the Fisher-Yates exact test (Fisher, 1934; Yates, 1934), which is calculated as:

$$P_{observed\ distribution} = \dfrac{\left(\dfrac{a+c}{a}\right)*\left(\dfrac{b+d}{b}\right)}{\dfrac{N}{a+b}} + \sum P_{all\ more\ extreme\ distributions}.$$ Gries et al. (2005) used the negative base ten logarithm of the p value to rank-order the strength of association between verbs and constructions. We use a variant calculation of collostructional strength, which correlates almost perfectly with the original method (Gries, pers. comm.), but is much easier to compute with large frequency values (such as those found in COCA). This method is calculated as follows: *approximate collexeme strength* $=\left(\left(\dfrac{a}{a+b}\right)-\left(\dfrac{c}{c+d}\right)\right)*(a+b).$

Table 11. Correlations between holistic essay score and SCA variables entered into regression model.

Variable	Correlation with holistic score
Mean length of clause	0.240
Coordinate phrases per clause	0.190

We include mean indices for types and tokens, along with standard deviations for tokens. We also calculate the ratio of attracted verb–VAC combinations to repelled verb–VAC combinations (for a total of five indices).

Results and discussion

Research Question 1: Syntactic complexity

A number of preliminary analyses were conducted to ensure the data was appropriate for stepwise multiple regression analysis. Each of the 14 indices of syntactic complexity measured by SCA (Lu, 2010, 2011) demonstrated normal distributions. Scatterplots suggested that each of the 14 indices demonstrated a linear relationship with holistic scores. Eleven of the 14 indices did not reach both Cohen's (1988) threshold for a meaningful (small) effect of $r \geqslant 0.100$ and a conservative threshold of statistical significance of $p <$ .001 with TOEFL essay scores and were removed from the analysis. Of the remaining three variables, one (complex nominals per clause) was removed due to multicollinearity (Tabachnick & Fidell, 2014) with mean length of clause. The remaining two variables (mean length of clause and coordinate phrases per clause) were entered into a stepwise regression (see Table 11 for correlations between these variables and the TOEFL essay scores). The resulting model, which included one variable (mean length of clause), explained 5.8% ($r =.240$, $R^2 = .058$) of the variance in holistic essay scores (see Table 12 for the model). A tenfold cross-validated model explained 8.2% of the variance in holistic essay scores. The model explained 2.7% ($r = .163$, $R^2 = .027$) of the variance in prompt 1 scores and 8.9% ($r = .298$, $R^2 = .089$) of the variance in prompt 2 scores. A Fisher's r to z transformation indicated that the amount of variance explained by the model across the two prompts did not differ significantly ($z = -1.56$, $p = .119$).

The results indicate that the relationship between indices of syntactic complexity calculated by the syntactic complexity analyzer (SCA) and TOEFL independent essay scores was significant, but small. Two indices met the index selection criteria (MLC and CP/C), and a single index (MLC) was included in the predictor model. The results indicate that higher rated essays tend to include longer clauses. These findings support previous studies, such as Lu (2010, 2011), who found similar results across university levels (i.e., as university level increased, writers used longer clauses and more coordinate phrases per clause). These results also align with the findings from Ortega's (2003) synthesis of L2 writing studies, which found either neutral or positive relationships between MLC and holistic scores of writing quality. Overall, however, these differences demonstrated small effects and explained only a small portion of the variance in holistic scores

Table 12. Summary of SCA multiple regression model.

Entry	Predictors included	r	R^2	R^2 change	β	SE	B
1	Mean length of clause	.240	.058	.058	.110	.201	.240

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

Table 13. Examples from TOEFL Essays: Mean length of clause.

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

of writing quality. In Table 13, we present examples of shorter and longer clauses found in in the TOEFL writing data.

From a theoretical interpretation standpoint, the results are arguably opaque. They suggest that higher proficiency writers tend to employ longer clauses, but leave gaps in our understanding regarding the specific structures that contribute to increased clause length (e.g., phrasal coordination, adverbs, adverbial phrases, noun modifiers, etc.). Furthermore, it is unclear whether the inclusion of particular clause-lengthening structures is uniform across participants and/or score levels (Larsen-Freeman, 2009; Norris & Ortega, 2009).

Research Question 2: Usage-based indices

A number of preliminary analyses were conducted to ensure the data was appropriate for stepwise multiple regression analysis. Thirty-five VAC indices were initially considered (18 related to frequency and 17 related to association strength). Seven VAC indices violated the assumption of normality and were removed from further consideration. Of the remaining 28 indices, 16 did not reach both Cohen's (1988) threshold for a meaningful (small) effect of $r \geqslant 0.100$ and a conservative threshold of statistical significance of $p < .001$ with TOEFL essay scores and were removed from the analysis. Of the remaining 12 variables, eight were removed due to multicollinearity (Tabachnick & Fidell, 2014). The remaining four variables (see Table 14) were entered into a stepwise regression. The resulting model, which included four variables, explained 14.2% ($r = .376$, $R^2 = .142$) of the variance in holistic essay scores (see Table 15 for the model). The tenfold cross-validated model explained 14.0% ($r = .374$, $R^2 = .140$) of the variance, suggesting that the model was consistent across the dataset. The model explained 17.6% ($r = .420$,

Table 14. Correlations between holistic essay score and syntactic sophistication variables entered into regression.

Variable	Correlation with holistic score
Average delta p score verb (cue) – construction (outcome) (types only) –academic	0.251
Average lemma construction frequency (types only) – academic	−0.234
Average faith score construction (cue) – verb (outcome) (types only) – academic	0.166
Collostruction ratio (types only) – academic	0.155

Table 15. Summary of usage-based multiple regression model.

Entry	Predictors included	r	R^2	R^2 change	β	SE	B
1	Average delta p score verb (cue) – construction (outcome) (types only) –academic	0.251	0.063	0.063	17.81	3.548	0.22
2	Average lemma construction frequency (types only) –academic	0.345	0.119	0.056	−3.06E–05	5.68E–06	−0.23
3	Average faith score construction (cue) – verb (outcome) (types only) - academic	0.366	0.134	0.015	4.863	1.876	0.112
4	Collostruction ratio (types only) – academic	0.376	0.142	0.008	0.025	0.012	0.09

Note: Estimated constant term = 3.280, β = unstandardized beta, *SE* = standard error; *B* = standardized beta.

R^2 = .176) of the variance in prompt 1 scores and 11.4% (r =.338, R^2 =.114) of the variance in prompt 2 scores. A Fisher's r to z transformation indicated that the amount of variance explained by the model across the two prompts did not differ significantly (z = .105, p = .294).

The results indicate that the relationship between usage-based indices and holistic scores of writing quality was significant and demonstrated a medium effect. Four variables were included in a model that explained 14.2% of the variance in essay score. Essays that included more strongly associated verb–VAC combinations and less frequent verb–VAC combinations tended to earn higher scores. In Table 16, we present examples of weakly and strongly associated verb–VAC combinations. In Table 17, we present examples of low- and high-frequency verb–VAC combinations.

The results are relatively straightforward to interpret from a theoretical standpoint. The results suggest an interplay between frequency and strength of association. Higher rated essays, which ostensibly were written by more proficient L2 users, tended to include less frequent verb–VAC combinations. These findings support usage-based perspectives on language learning (e.g., Behrens, 2009; Ellis, 2002a; Tomasello, 2003) which suggest that frequent constructions will be learned earlier/more easily. As learners have more language experiences (and become more proficient), they will learn (and

Table 16. Examples of weak and strong verb–VAC associations in TOEFL essays.
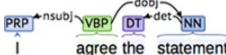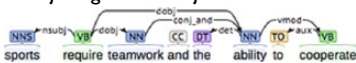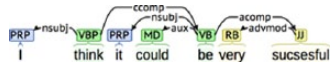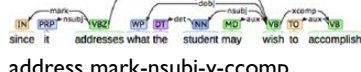
Essay score	Association	Verb–VAC Combination
2	Weak	I agree the statement nsubj – *agree* – dobj
5	Strong	sports require teamwork and the ability to cooperate subject – *require* – direct object – direct object

Table 17. Examples of high and low frequency verb–VAC combinations.

Essay score	Frequency	VAC
2	High	I think it could be very sucsesful think nsubj-v-ccomp
5	Low	since it addresses what the student may wish to accomplish address mark-nsubj-v-ccomp

use) constructions that are less frequent, as is evident in the findings from this study. Higher-rated essays also included verb–VAC combinations that were more strongly associated. Usage-based research has also suggested that the verb–VAC profiles of advanced proficiency L2 users also reflect corpus-based profiles (Römer et al., 2015, 2014). Ninio (1999) posits that verb–VAC combinations are first learned as fixed constructions which represent frequent and strongly associated verb–VAC combinations in the input. After adequate, varied input learners begin to learn the schematicity of the verb slot by using a new verb. Shortly after the first new verb is used in a construction many more follow, providing evidence of generalization. The data in the current study supports the notion that early stages of generalization are represented by overgeneralization (Verspoor & Behrens, 2011), wherein non-input like verbs are used in newly learned constructions. Through more exposure to input/use, however, verb–VAC combination are tuned and become more target like.

Research Question 3: Model comparison

The results of a Fisher r to z transformation indicated that the predictor model derived from usage-based indices of syntactic sophistication explained a significantly larger portion ($z = 2.33$; $p = .02$) of the variance in holistic writing scores than the predictor model based on syntactic complexity indices.

Conclusion

The sophistication of syntactic forms used by L2 users has been of interest in second language acquisition and language assessment over the past 45 years (Cumming et al.,

2005; Larsen-Freeman, 1978; Wolfe-Quintero et al., 1998). Most previous research has investigated syntactic sophistication using the text internal, formal construct of syntactic complexity, often through large-grained indices such as mean length of T-unit (MLTU) and mean length of clause (MLC). Results have been mixed, but a tentative trend suggests that as learners become more proficient writers and/or earn higher writing quality scores they tend to use more complex syntactic structures (i.e., longer clauses and T-units; Ortega, 2003; Cumming et al., 2005), though this is not always the case (Knoch et al., 2014; Ortega, 2003). One important limitation with the use of large-grained indices of syntactic complexity is that the results are difficult to interpret from a theoretical standpoint (Larsen-Freeman, 2009; Norris & Ortega, 2009). One alternative is to examine syntactic complexity through the lens of usage-based perspectives (e.g. Ellis, 2002b; Tomasello, 2003).

In light of this previous research, the current study compared two methods of measuring syntactic sophistication in L2 writing quality data. The first method assessed the use of traditional indices of syntactic complexity such as MLTU and MLC to predict L2 writing quality. The second method assessed usage-based indices derived from the average main verb lemma frequency, VAC frequency, verb–VAC frequency, and strength of association based on the academic section of COCA to predict L2 writing quality. The results indicated that both methods resulted in significant predictor models. The usage-based model, however, explained a significantly larger portion (14.2%) of the variance in holistic scores of writing quality than traditional methods (5.8%). In addition to being a stronger model, the usage-based indices provide a clearer path to a theory-based interpretation of the results.

The results suggest that human ratings of essay quality may be sensitive to both the relative frequency of constructions themselves and the strength of association between constructions and the verbs that fill them. This may be captured in the TOEFL independent writing rubric wherein verb–VAC combinations may be subsumed under the descriptor "appropriate word choice," and in some cases under "lexical or grammatical errors." Evidence for this is seen in essays that include weakly associated verb–VAC combinations earning lower quality scores and in essays that include strongly associated verb–VAC combinations earning higher quality scores. Future analytic scoring techniques may benefit by making this connection explicit both to raters (to help reduce rater variability) and to test takers (to explicitly outline rater expectations), though the efficacy of both of these suggestions should be empirically investigated. Beyond rubric development, the findings have implications for AES models, which could also benefit from the inclusion of usage-based indices of syntactic development. The inclusion of such indices could both increase model accuracy and construct coverage.

A number of limitations in this study should be considered in future research. First, although the results indicated that usage-based indices of syntactic sophistication explained a larger proportion of the variance than traditional syntactic complexity indices, neither model was particularly strong when compared to more comprehensive models (Burstein et al., 2013; Guo et al., 2013). This is not surprising given that most (if not all) essay scoring rubrics include descriptors from a range of different language proficiency areas (e.g., lexical proficiency and cohesion) in addition to syntax. Future research

should investigate the degree to which these indices add to both the construct coverage and the predictive power of models that include a wider array of language features.

A second potential limitation in this study is the use of COCA as a proxy for L2 language experience for the usage-based indices. COCA was designed to be representative of general English language use in America (Davies, 2009, 2010), but likely does not fully represent the types of language to which L2 learners are exposed. A corpus that included the types of language that language learners are commonly exposed to would likely serve as a better proxy for language experience, and may yield stronger (and more representative) results. Outlining the characteristics for such a corpus, collecting appropriate texts, and replicating this study may be a rich area of investigation. Corpora such as Touchstone Applied Science Associates (TASA) corpus (Landauer, Foltz, & Laham, 1998), Michigan Corpus of Academic Spoken English (MICASE) (Simpson-Vlach & Leicher, 2006), and the TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL) (Biber, Conrad, Reppen, Byrd, & Helt, 2002) may be appropriate starting points for such research.

Acknowledgements

Declaration of Conflicting Interest

Funding

Notes

1. The entire TOEFL independent writing rubric can be found at www.ets.org/Media/Tests/ TOEFL/pdf/ Writing_Rubrics.pdf
2. TAASSC is freely available at www.kristopherkyle.com/taassc.html
3. Verb–VAC association strength norms in TAASSC do not include copular constructions (which are very strongly associated with the verb to be) to avoid skewing mean association strength scores.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Bachman, L. F., & Cohen, A. D. (1999). Language testing – SLA interfaces: An update. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 1–31). Cambridge: Cambridge University Press. http://doi. org/10.1017/CBO9781139524711.003

Behrens, H. (2009). Usage-based and emergentist approaches to language acquisition. _Linguistics_, _47_(2), 383–411. http://doi.org/10.1515/LING.2009.014

Bencini, G. M., & Goldberg, A. E. (2000). The contribution of argument structure constructions to sentence meaning. _Journal of Memory and Language_, _43_(4), 640–651. http://doi.org/10.1006/jmla.2000.2757

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. http://doi.org/http://dx.doi.org/10.1016/j.jslw.2014.09.004

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., … Urzua, A. (2004). _Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus_. TOEFL Monograph Series. Retrieved from www.ets.org/Media/Research/pdf/RM-04–03.pdf

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. _TESOL Quarterly_, _36_(1), 9–48. http://doi.org/10.2307/3588359

Biber, D., & Gray, B. (2013). Discourse characteristics of writing and speaking task types on the TOEFL IBT® test: A lexico-grammatical analysis. _ETS Research Report Series_, _2013_(1).

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? _TESOL Quarterly_, _45_(1), 5–35. Retrieved from www.jstor.org/stable/41307614

Biber, D., Gray, B., & Staples, S. (2014). Predicting patterns of grammatical Complexity across language exam task types and proficiency levels. _Applied Linguistics_, amu059. http://doi.org/10.1093/applin/amu059

BNC Consortium (2007). _British national corpus_, version 3 (BNC XML ed.). Retrieved from http://www.natcorp.ox.ac.uk

Briscoe, T., Carroll, J., & Watson, R. (2006). The second release of the RASP system. In _Proceedings of the COLING/ACL on Interactive presentation sessions_. (pp. 77–80). http://doi.org/10.3115/1225403.1225423

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken & F. Vedder (Eds.), _Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA_. Amsterdam: John Benjamins.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), _Handbook of automated essay evaluation: Current applications and new directions_ (pp. 55–67). New York: Routledge.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. _Language_, _82_(4), 711–733. Retrieved from www.jstor.org/stable/4490266

Chang, F., Bock, K., & Goldberg, A. E. (2003). Can thematic roles leave traces of their places? _Cognition_, _90_(1), 29–49. http://doi.org/10.1016/S0010–0277(03)00123–9

Chapelle, C. A. (1999). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), _Interfaces between second language acquisition and language testing research_ (pp. 32–70). Cambridge: Cambridge University Press. http://doi.org/10.1017/CBO9781139524711.004

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2011). _Building a validity argument for the Test of English as a Foreign Language_. New York: Routledge.

Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In _Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)_ (pp. 740–750). Retrieved from https://cs.stanford.edu/~danqi/papers/emnlp2014.pdf

Chodorow, M., & Burstein, J. (2004). Beyond essay length: evaluating e-rater®'s performance on toefl® essays. *ETS Research Report Series*, *2004*(1).

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

Crossley, S. A., Cai, Z., & McNamara, D. S. (2012). Syntagmatic, paradigmatic, and automatic n-gram approaches to assessing essay quality. In *Twenty-Fifth International FLAIRS Conference*.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, *10*(1), 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*(2), 159–190. http://doi.org/10.1075/ijcl.14.2.02dav

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*(4), 447–464. http://doi.org/10.1093/llc/fqq018

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24. http://doi.org/10.1016/j.asw.2012.10.002

Ellis, N. C. (2002a). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*(2), 143–188. Retrieved from http://dx.doi.org/10.1017/S0272263102002024

Ellis, N. C. (2002b). Reflections on frequency effects in language processing. *Studies in Second Language Acquisition*, *24*(2), 297–339. Retrieved from http://dx.doi.org/10.1017/S0272263102002140

Ellis, N. C., & Ferreira-Junior, F. (2009a). Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal*, *93*(3), 370–385. http://doi.org/10.1111/j.1540–4781.2009.00896.x

Ellis, N. C., & Ferreira-Junior, F. (2009b). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, *7*(1), 188–221. http://doi.org/10.1075/arcl.7.08ell

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, *27*(3), 317–334.

Eskildsen, S. W. (2009). Constructing another language—usage-based linguistics in second language acquisition. *Applied Linguistics*, *30*(3), 335–357. http://doi.org/10.1093/applin/amn037

Eskildsen, S. W., & Cadierno, T. (2007). Are recurring multi-word expressions really syntactic freezes? Second language acquisition from the perspective of usage-based linguistics. In *Nordic Conference on Syntactic Freezes*. Retrieved from www.forskningsdatabasen.dk/en/catalog/2185991292

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*(852), 285–307. Retrieved from www.jstor.org/stable/2935559

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago, IL: University of Chicago Press.

Goldberg, A. E. (2013). Constructionist approaches. In T. Hoffman & G. Trousdale (Eds.), *The Oxford handbook of construction grammar* (pp. 15–31). Oxford: Oxford University Press.

Goldberg, A. E., Casenhiser, D., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics*, *15*(3), 289. http://doi.org/10.1515/cogl.2004.011

Gries, S. T., & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Learning*, *65*(S1), 228–255. http://doi.org/10.1111/lang.12119

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*(4), 635–676. http://doi.org/10.1515/cogl.2005.16.4.635

Gries, S. T., & Wulff, S. (2005). Do foreign language learners also have constructions? *Annual Review of Cognitive Linguistics*, *3*(1), 182–200. http://doi.org/10.1075/arcl.3.10gri

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*(3), 218–238.

Hare, M. L., & Goldberg, A. E. (1999). Structural priming: Purely syntactic. In *Proceedings of the 21st annual meeting of the Cognitive Science Society* (pp. 208–211). London: Lawrence Erlbaum.

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

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. http://doi.org/10.1016/j.asw.2014.01.001

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. Retrieved from 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*(4), 757–786. http://doi.org/10.1002/tesq.194

Kyle, K., Crossley, S. A., & McNamara, D. S. (2015). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing, OnlineFirs*. http://doi.org/10.1177/0265532215587391

LaFlair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing*, Vol. 34(4): 451–475

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Stanford, CA: Stanford University Press.

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly*, *12*(4), 439–448. http://doi.org/10.2307/3586142

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 579–589. http://doi.org/10.1093/applin/amp043

Lieven, E. V. M., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, *24*(1), 187–219.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. http://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*(1), 36–62. Retrieved from www.jstor.org/stable/41307615

Ninio, A. (1999). Pathbreaking verbs in syntactic development and the question of prototypical transitivity. *Journal of Child Language*, *26*(3), 619–653. Retrieved from http://dx.doi.org/10.1017/S0305000999003931

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4). http://doi.org/10.1093/applin/amp044

O'Donnell, M. B., & Ellis, N. (2010). Towards an inventory of English verb argument constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics* (pp. 9–16). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1866732.1866734

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492–518.

Römer, U., O'Donnell, M. B., & Ellis, N. C. (2015). Using COBUILD grammar patterns for a large-scale analysis of verb-argument constructions. In N. Groom, M. Charles & J. Suganthi (Eds.), *Corpora, grammar and discourse: In honour of Susan Hunston* (Vol. 73, p. 43). Amsterdam: John Benjamins.

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*(1), 115–135.

Simpson-Vlach, R. C., & Leicher, S. (2006). *The MICASE handbook: A resource for users of the Michigan corpus of academic spoken English*. Ann Arbor, MI: University of Michigan Press.

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, *8*(2), 209–243. http://doi.org/10.1075/ijcl.8.2.03ste

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics*. Harlow, Essex: Pearson Education. Retrieved from http://lib.myilibrary.com?id=526967

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

Tomasello, M., & Brooks, P. J. (1999). Early syntactic development: A Construction Grammar approach. In M. Barret (Ed.), *The development of language* (pp. 161–190). New York: Psychology Press.

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

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

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, *28*, 53–67. http://doi.org/10.1016/j.jslw.2015.02.002

Yates, F. (1934). Contingency tables involving small numbers and the χ2 test. *Supplement to the Journal of the Royal Statistical Society*, *1*(2), 217–235. http://doi.org/10.2307/2983604