Verb Argument Construction Complexity Indices and Second Language Writing Quality: A Task-Based Writing Analysis

Dr. Ayşe Demir[1]*

[1]Boğaziçi University, Department of Foreign Language Education, Istanbul, Turkey

Abstract

The purpose of the study is to investigate whether verb argument construction (VAC) based indices of syntactic complexity significantly vary in different second language (L2) writing tasks and whether such complexity measures vary in their relationship to writing quality. The present study used data from three different L2 writing tasks: descriptive, independent, and integrated. Descriptive essays (N=70) were taken from the Michigan State University corpus (Connor-Linton & Polio, 2014). Independent (N=70) and integrated (N=70) essays were selected from TOEFL public use dataset. Each corpus contained an equal number of essays written on two prompts (for a total of six prompts). Each essay was rated by two trained raters using TOEFL independent or integrated writing rubrics. The selected essays were stratified by writing scores, and it was made sure that the scores were normally distributed. The Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC et al., 2017) was used to calculate the VAC-based indices. Multinomial logistic regression and linear models were used to analyze the data. The findings indicate that VAC-based complexity measures vary by L2 writing tasks and that the relationship between VAC measures and L2 writing quality is also task-dependent with few prompt-based effects.

Keywords: verb argument construction, syntactic sophistication, L2 writing tasks, L2 writing prompts, L2 writing quality

Introduction

Syntactic complexity has been traditionally defined as the number of elements in a syntactic constituent and the number of connections between those elements, i.e., absolute complexity measures (Bulté & Housen, 2012; Kyle & Crossly, 2017). Previous studies found that more proficient L2 learners produce more complex syntactic structures (Lu, 2011; Ortega, 2003; Wolfe-Quintero et al., 1998). However, absolute complexity measures do not strongly

overlap with some aspects of second language (L2) acquisition theories, including usage-based approaches which suggests that L2 learning is closely related to the frequency of L2 forms in the input (Ellis, 2002; Ellis et al., 2013). From a usage-based perspective, more frequent constructions (or form-meaning pairings including syntactic constructions; Goldberg, 1995) are learned more easily than less frequent constructions (Ellis et al., 2013). In terms of production, usage-based studies have demonstrated that verb argument constructions (VAC: a verb slot and the related arguments in a syntactic construction; Kyle & Crossley, 2017) explain more variance in holistic rating of L2 argumentative writing than absolute complexity measures (Kyle & Crossley, 2017). However, beyond argumentative essays, the relationships between VAC-based indices and L2 writing tasks remain under-explored. Additionally, little is known about the mediating effects of writing tasks and prompts on the relationships between VAC-based indices and L2 writing quality, whereas in previous studies, both the variables (tasks and prompts) have been found to influence absolute complexity indices (Beers & Nagy, 2009; Way et al., 2000; Yang et al., 2015). Hence, the present study investigates whether VAC-based indices of syntactic sophistication significantly vary among different L2 writing tasks and whether the relationship between VAC-based complexity measures and L2 writing quality is mediated by the tasks and prompts of writing.

Syntactic Complexity

The construct of complexity has often been equated with a variety of terms such as late acquired, more advanced, more developed, more proficient, or more sophisticated (Housen et al., 2019; Pallotti, 2015). However, such diverse terms may indicate conceptually distinct constructs and may not be empirically correlated (Housen et al., 2019; Pallotti, 2015). Language complexity can be defined as relative complexity (difficulty or user-related complexity) and absolute

complexity (inherent or structural complexity), and these two types of complexity should be treated as distinct constructs in SLA research (Housen et al., 2019). Absolute complexity refers to the basic formal structures of linguistic units (e.g., phrases, clauses, or sentences) such as the number and variety of their constituent components (Housen et al., 2019). On the contrary, relative complexity is "generally understood as a property of language phenomena which are acquired late or which are cognitively taxing, with simple language phenomena being acquired early in development and/or taking up few cognitive resources in language processing and use" (Housen et al., 2019, p. 4). Relative complexity is related to the issues of learnability and difficulty. Objective determinants of L2 difficulty include the saliency of an L2 form, which might be a function of that form's frequency in the input. In SLA-based studies, very often the focus is on structural complexity although a form that is structurally more complex might not be more difficult to learn or process (Housen et al., 2019).

Additionally, it has been suggested based on empirical findings that syntactic complexity in L2 writing might be influenced by interactions among variables such as development, proficiency, and first language (L1) background (Ortega, 2003, 2015). For example, Lu and Ai (2015) found significant differences in structural or absolute complexity measures between groups of English as L1 and L2 writers when the L2 writers were grouped based on their L1 backgrounds. However, all areas of language might not be equally susceptible to be influenced by L1 background (Ortega, 2015). In addition to L1 background and proficiency, task-types can also influence syntactic complexity in L2 writing.

Writing Task Effects on Syntactic Complexity

Previous studies have investigated the effects of writing tasks on syntactic complexity features (e.g., Shaw & Weir, 2007; Yang et al., 2015). For example, Lu (2011) found that

argumentative essays exhibited higher syntactic complexity than narratives. Additionally, Beers and Nagy (2011) found that more subordinate clauses were used in persuasive essays than in narrative, descriptive, and compare/contrast essays, and that descriptive essays had more words per clause than persuasive essays.

Integrated writing and syntactic complexity. In addition to traditional independent writing tasks (that do not require any integration of information from a source text), integrated writing tasks (that involve integrating information from print and audio source texts) are increasingly used in L2 writing contexts. Cumming et al. (2005) found that the texts written for integrated tasks significantly differed from those written for independent tasks in terms of absolute syntactic complexity measures, e.g., number of words per T-unit and number of clauses per T-unit. For example, in the integrated essays, the writers used wider range of vocabulary and longer clauses than in the independent essays (Cumming et al., 2005). These differences likely result from integrated tasks demanding more attention from L2 learners (Plakans, 2010). Plakans (2008) found that integrated tasks involved a more interactive writing process, whereas independent writing tasks involved less online (but more initial) planning. Furthermore, Plakans (2009) showed that in integrated writing, L2 writers engaged in discourse synthesis processes that involve selecting content from source texts, organizing the content, and connecting the content by linking related ideas. The use of such discourse synthesis strategies might help L2 writers improve the quality of their writing and reduce their copying phrases from source texts (Yang & Plakans, 2012). However, in Plakans' (2009) study, L2 writers faced issues related to writing style and vocabulary during the discourse synthesis processes, and Gebril and Plakans (2009) also argued that limited L2 proficiency (e.g., in fluency and grammatical accuracy) of lower level writers may impact their use of source texts in integrated writing. In contrast, higher

level writers might have more facility in integrated writing tasks with higher comprehension of source materials (Gebril & Plakans, 2009).

Task-effects on the relationship between complexity and writing quality. The relationship between complexity measures and quality in L2 writing may also vary depending on writing tasks. For example, in an investigation into the relationship between syntactic complexity measures and writing quality of narrative and persuasive essays, Beers and Nagy (2009) found that words per clause was positively correlated with persuasive essay quality, but not with the narrative essay quality. On the contrary, clauses per T-unit was significantly and positively correlated with narrative essay quality but negatively correlated with persuasive essay quality (Beers & Nagy, 2009). Furthermore, clausal complexity features (e.g., use of subordinate clauses) have been found to be predictive of writing quality in L2 descriptive essays (Bulté & Housen, 2014; Crossley & McNamara, 2014) while fine-grained indices of phrasal complexity (e.g., adjective modifiers per object of the preposition and dependents per direct object) have been found to be predictive of writing quality in L2 argumentative essays (Kyle & Crossley, 2018).

Writing Prompt Effects on Syntactic Complexity

Several studies have shown that linguistic features in written texts are affected by writing prompts (Crossley et al., 2011; Hinkel, 2002; Huot, 1990; Tedick, 1990). Hinkel (2002) argued that the grammar and vocabulary as well as the thematic and contextual content of the prompts might affect the writing produced by L2 writers. For example, learners use simpler syntactic and lexical constructions when writing essays on prompts that are closer to their interests than in essays on prompts, unrelated to their interests (Hinkel, 2002). Additionally, Way et al. (2000) reported that writing quality as well as syntactic complexity (measured by mean length of T-unit)

varied depending on writing prompts for novice learners of French as L2. Similarly, Yang et al. (2015) found that L2 writers used significantly more finite subordination when writing essays on prompts that required causal reasoning than in essays on prompts that did not require such reasoning. Interestingly, finite subordination was not predictive of writing scores in the tasks that required causal reasoning, whereas finite subordination had a significant and positive relationship with writing quality scores in the tasks that did not require causal reasoning (Yang et al., 2015). These findings indicate potentially complex interactions between writing tasks, prompts, and writing quality scores.

Usage-based Approaches to Syntactic Complexity

Most of the studies discussed above examine large-grained indices that measure syntactic complexity at the clausal, phrasal, or sentence level (e.g., the length of clauses/T-units/sentences). Such absolute complexity indices have been subject to criticism in recent research (Biber et al., 2011; Bulté & Housen, 2012; Norris & Ortega, 2009). This criticism rests on the notion that T-unit and clausal subordination are more common in spoken conversations than in academic writing and that T-unit and subordination-based measures are not adequate discriminators of language proficiency differences (Biber et al., 2011). Additionally, large-grained indices of syntactic complexity such as mean length of T-unit cannot indicate what specific structures emerge in learners' writing as they develop their proficiency. This is because such large-grained measures do not indicate the type of structures (e.g., phrasal dependents and dependent clauses) that can increase the length of a T-unit (Kyle & Crossley, 2018). Therefore, more recent studies have begun to focus on relative as compared to absolute syntactic complexity features, also known as syntactic sophistication (Housen et al., 2019; Kyle & Crossley, 2017). Relative complexity involves "the relative difficulty of learning, using, and/or comprehending a

particular structure" (Kyle & Crossley, 2017, p. 514) and follows usage-based approaches to language learning (Ellis et al., 2013; Tomasello, 2003).

In usage-based view of language acquisition, units of language are constructions that are combinations of meanings and their related forms and are not dependent on any lexical items (Behrens, 2009; Goldberg, 1995; Goldberg et al., 2007). One of the main determinants of constructionist learning is the frequency of constructions in the input (Ellis et al., 2013; Tomasello, 2003) under the assumption that the more learners are exposed to a particular construction, the stronger it is entrenched in their memory and the more likely it is to be learned (Ellis, O'Donnell, et al., 2013; Ellis, Römer, et al., 2016). Many usage-based L2 acquisition studies focus on VACs, which consist of a verb slot and the related arguments (Kyle & Crossley, 2017). For example, the ditransitive construction, "$\text{She}_{\text{subject}}$ $\text{gave}_{\text{verb}}$ $\text{him}_{\text{indirect object}}$ a $\text{book}_{\text{direct object}}$," is a VAC related to transfer (Goldberg, 2013, p. 21). Previous research showed that VACs are important components of L2 acquisition (Bencini & Goldberg, 2000; Gries & Wulff, 2005). It has been argued that language acquisition moves from formulas or memorized chunks ("I gave him a pen"), to low-score patterns (a construction with a fixed part and an open slot, "I gave him ____"), to entirely abstract schematic knowledge of constructions, e.g., "$\text{NP}_{\text{subject}}$-Verb-$\text{NP}_{\text{indirect object}}$-$\text{NP}_{\text{direct object}}$ (Eskildsen, 2008). Such schematic knowledge derives from the speakers' experience of language input, and frequency is of primary importance in noticing "structural regularities that emerge from learners' implicit analysis of the distributional characteristics of the language input" (Ellis et al., 2016, p. 57).

VAC patterns can be measured using reference corpora as proxies for learners' experiences (Kyle & Crossley, 2017; Römer, O'Donnell, et al., 2015; Römer, Roberson, et al., 2014). For instance, Kyle (2016) and Kyle and Crossley (2017) used the Corpus of

Contemporary American English (COCA, Davis, 2010) to develop automated VAC features (e.g., frequency of main verb lemmas, VACs, and strength of association between main verb lemma and VACs). In Kyle and Crossley (2017), VAC-based indices were found to be strong predictors of L2 writing quality for argumentative essays, even when compared to absolute complexity measures (e.g., mean length of T-unit and mean length of clauses). These findings indicate that higher-scoring L2 argumentative essays contained less frequent and more strongly associated verb-VAC combinations (Kyle & Crossley, 2017).

The Present Study

Research has reported that absolute complexity indices vary among different L2 writing tasks (Beers & Naggy, 2009; Cumming et al., 2005) and prompts (Hinkel, 2002; Way et al., 2000). It is an empirical question whether different types of L2 writing feature different syntactic sophistication indices based on usage-based approaches. Examining such syntactic sophistication indices in varied writing tasks can indicate the extent to which learners use sophisticated forms (indicated by their frequency in the input) or strongly associated constructions that are schematized in their mental repertoire while writing different types of essays. Such syntactic sophistication indices related to the frequency of forms in the input might also be informative of the difficulty levels of different writing tasks because more frequent constructions are usually easier to learn than less frequent constructions (Ellis & Ferreira-Junior, 2009a). Kyle and Crossley (2017) examined the relative performance of traditional syntactic complexity indices (tapping into absolute complexity) and the VAC-based indices of syntactic sophistication in explaining variance in holistic scores of writing quality by focusing on only one type of task (TOEFL independent essays, also partially used in the present study as the independent task). However, there has been little research that investigated whether VAC-based syntactic

sophistication measures are predictive of different L2 writing tasks or what kind of effects tasks and prompts have on the relationships between syntactic sophistication measures and holistic ratings of L2 writing.

The present study addresses these gaps by examining VAC-based syntactic sophistication indices in three different L2 writing tasks: descriptive, independent, and integrated. In addition to integrated writing tasks that replicate writing in real life academic contexts (Cumming, 2013), the present study includes descriptive and independent tasks both of which are commonly used in testing and academic contexts (Connor-Linton & Polio, 2014). Whereas in the descriptive task, the topics were simple that may not elicit sophisticated academic writing skills (Connor-Linton & Polio, 2014), the independent task involved argumentation, and previous research found higher structural complexity in L2 argumentative essays (Beers & Nagy, 2011; Ravid & Berman, 2010). Due to such differences between descriptive and independent writing, both task types were included in the present study. Additionally, each task contains essays written on two different prompts, allowing for prompt-based effects to be considered. The study operationalizes writing quality as expert essay ratings based on holistic writing rubrics. The purpose of the present study is to investigate whether VAC-based syntactic sophistication indices are predictive of the three L2 writing tasks. The study also examines whether VAC-based indices predict L2 writing quality while co-varying with writing tasks and prompts. The specific research questions are the following:

1. Are VAC-based syntactic sophistication indices predictive of different L2 writing tasks?
2. Does the relationship between VAC-based indices and L2 writing quality vary depending on the writing tasks?

3. Does the relationship between VAC-based indices and L2 writing quality vary depending on the prompts of writing in each task?

Methods

Corpora

Descriptive essays. The descriptive essays were collected from 70 university-aged English as second language (ESL) learners at Michigan State University (Connor-Linton & Polio, 2014). The corpus included descriptive essays from both non-matriculated ESL learners in advanced level Intensive English Program (IEP) classes and matriculated ESL learners in English for academic purposes (EAP) classes. Each essay was written independently in 30 minutes without any opportunity to consult any source materials or to revise. The students chose from eight essay topics (see Table 1) divided into four sets that elicited descriptive essays (Connor-Linton & Polio, 2014). Those four sets of topics were counterbalanced among the participants, and at each time, the participants chose one of the two topics. The original corpus was longitudinal in design (the participants wrote three essays over a semester); however, for the current study, we only used essays collected from one time point to create a cross-sectional sub-corpus.

Table 1

Topics in the Descriptive Sub-Corpus (Connor-Linton & Polio, 2014)

Set	Topic
A	Describe your current home. Describe a place you have visited recently
B	Describe the campus of Michigan State University (MSU). Describe a good or bad teacher that you have had.
C	Describe your family. Describe a good friend of yours.
D	Describe a school that you have attended.

Describe a problem that the United States or some other country is facing.

No more than 10 participants wrote essays on a topic at each time. Therefore, to maximize the sample size in the corpus, the topics were divided into two broad categories: describing a place (including essays on the following four prompts: *Describe your current home; Describe a place you've visited recently; Describe the campus of MSU; Describe a school that you've attended*) and describing a person/persons (including essays on the following three prompts: *Describe your family; Describe a good friend; Describe a good or bad teacher that you've had*). No essays were selected from the following prompt: *Describe a problem that the USA or some other country is facing*.

Table 2 presents an overview of the descriptive sub-corpus used in this study.

Table 2

Overview of the Descriptive Sub-Corpus (word count)

	Prompt: Person (n= 34)	Prompt: Place (n= 36)	Total (n= 70)
Total word count	12,256	11,908	24,164
Minimum	196	207	196
Maximum	657	605	657
Mean	360.47	330.78	345.2
Std. Dev	101.91	93.85	98.27

Independent essays. A sub-corpus of independent argumentative essays (henceforth, independent essays) was formed from the Educational Testing Service (ETS) TOEFL public use data set. In that data set, 480 participants wrote one independent essay. There were two prompts, and half of the participants (n=240) wrote essays on one prompt and the other half (n=240) wrote essays on another. In each independent essay prompt, the test-takers were given two opposing viewpoints, and they had to support their positions with arguments in a timed 30 minutes writing task. In the prompt 1, the participants had to choose between selecting a study subject based on

their interests or selecting a study subject based on job prospects (henceforth, “choosing subject”). In the prompt 2, they had to choose between whether it is more important to cooperate with others now than it was in the past (henceforth, “cooperation”). The participants came from a variety of native language backgrounds. For the present study, 70 independent essays were selected to form the independent sub-corpus, and equal number of essays were selected from each prompt. As the essays were scored on a rubric that ranged from 1 to 5, we selected essays from each score level so that a range of scores were represented, and we also made sure that the scores (of the selected essays) were normally distributed.

Integrated essays. A sub-corpus of integrated essays was formed from the ETS TOEFL public use data set. Each of the same 480 participants who wrote one independent essay also wrote one integrated essay. For the integrated writing task, participants first read a reading passage and then listened to a lecture on the same topic. After that, they integrated the information from the listening lecture and explained in an essay how that information was related to that in the reading passage. The integrated writing task did not ask them to provide an opinion. One integrated writing prompt was related to different scientific theories on the navigation ability of migrating birds (henceforth, “migrating birds”), and the other prompt was on the environmental dangers of fish-farming (henceforth, “fish-farming”). Similar to the independent essay sub-corpus, the participants came from a variety of native language backgrounds. To form the integrated sub-corpus, 70 integrated essays, written by writers who were not included in the independent sub-corpus, were selected, and equal number of essays were selected from the two prompts. Similar to the independent essays, the integrated essays were also scored on a rubric that ranged from 1 to 5. Hence, we selected integrated essays from each score level (from 1 to 5),

and we made sure that the scores (of the selected essays) were normally distributed. Table 3 presents an overview of the integrated and independent sub-corpora used in the present study:

Table 3

Overview of the Integrated and Independent Sub-Corpora in the Present Study

	Integrated Sub-Corpus			Independent Sub-Corpus		
	Topic: Migrating birds (n=35)	Topic: Fish-farming (n=35)	Total (n= 70)	Topic: Choosing subject (n=35)	Topic: Cooperation (n= 35)	Total (n= 70)
Total word	6707	6664	13,371	10,798	9957	20,755
Min	92	103	92	61	85	61
Max	351	283	351	503	485	503
Mean	191.62	190.41	191.01	308.51.6	284.48.82	296.5
Std. deviation	58.71	46.16	52.43	92.17	88.23	90.38

Human Ratings

Each essay in the TOEFL independent and integrated sub-corpora was rated by two trained raters employed by ETS by using the TOEFL independent and integrated writing rubrics, respectively[1]. Each rubric was scaled from 1 to 5. In the independent writing rubrics, the low score of 1 was characterized by the following: serious lack of organization or development of ideas and presence of frequent language errors, and the high score of 5 was characterized by well-organized and well-developed ideas and effective language-use. For the integrated writing rubrics, a low score of 1 was characterized by the lack of any coherent content from the listening text and language use that obscured meaning; the high score of 5 was characterized by effective synthesis of information from the listening and reading texts with only occasional language errors that did not affect the meaning of content. In the rating procedure, if two scores for a text

[1] available at https://www.ets.org/Media/Tests/TOEFL/pdf/Writing_Rubrics.pdf

differed by a single point or less, the two scores were averaged. If the two scores differed by more than one point, a third rater scored the text.

Additionally, two trained raters rated the essays in the descriptive sub-corpus by using the TOEFL independent writing rubric. The authors used the independent writing rubric to score the descriptive essays because those essays were written independently with no opportunity to integrate information from a source material. Similar to the rating procedure of the TOEFL independent and integrated essays, if the two raters differed by one point or less, the scores were averaged, and if the difference was more than one point, a third rater rated the texts.

Table 4 shows Mean and Standard Deviations of the scores in the three sub-corpora used in the present study.

Table 4

Mean (Standard Deviation) of the Scores in the Three Sub-Corpora

	Descriptive Sub-Corpus			Independent Sub-Corpus			Integrated Sub-Corpus		
	Person (n=34)	Place (n=36)	Total (n=70)	Choosing subject (n=35)	Cooperation (n=35)	Total (n=70)	Fish-farming (n=35)	Migrating birds (n=35)	Total (n=70)
Mean (std.Dev)	3 (0.76)	3.45 (0.80)	3.22 (0.81)	3.22 (1.1)	3.14 (1.1)	3.185 (1.09)	3.21 (1.174)	3.18 (1.174)	3.20 (1.164)

VAC-based Syntactic Sophistication Indices

The present study uses the computational program Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) version 1.3.8 (Kyle, 2016; Kyle & Crossley, 2017) to calculate the VAC-based indices of syntactic sophistication for the essays in the three sub-corpora. The present study includes the syntactic sophistication indices in TAASSC that are based on the frequency profiles of all the written registers combined in COCA: fiction,

magazines, newspapers, and academic texts (Kyle, 2016; Kyle & Crossley, 2017). TAASSC calculates frequency-based indices for main verb lemmas (e.g., "to give"), VACs (e.g., subject-verb-indirect object-direct object), and verb-VAC combinations (e.g., subject-to give-indirect object-direct object) in a target text that also occur in the reference corpus, COCA (Davies, 2010). TAASSC also calculates the strength of association indices between VACs and the verbs that fill them (indicating the probability that a VAC and a main verb lemma will occur together). The VAC-based indices used in the present study are described briefly below.

Frequency-based indices. In TAASSC, average frequency scores for main verb lemmas, VACs, and verb-VAC combinations in a target text are calculated based on frequency counts derived from COCA. In the present study, in total 9 frequency indices based on the frequency counts from all the written sections of COCA are included as are shown in the Table 5.

Table 5

Summary of VAC-based Frequency Indices Included in the Present Study (Kyle & Crossley, 2017, p. 524)

	Main verb lemma frequency	VAC Frequency	Verb-VAC Combination Frequency
Mean token score	√	√	√
Mean token score (log transformed)	√	√	√
Mean type score	√	√	√
Total	3	3	3

Indices based on association strength. In TAASSC, the indices based on the strengths of association measure the probability that a main verb lemma and a VAC will co-occur in the COCA corpus. The present study includes three types of association strength measures calculated by TAASSC: faith (Gries et al., 2005), delta P (Ellis & Ferreira-Junior, 2009b), and a variant of collostructional strength (Stefanowitsch & Gries, 2003). Faith measures the conditional probability that a verb will occur with a particular VAC and vice versa (Kyle &

Crossley, 2017). Delta P is a variant of faith, and it calculates the probability of an outcome (e.g., a VAC) given a cue (e.g., a verb) minus the probability of the outcome without the cue (e.g., with any other verb). In contrast to faith and delta P, collostructional analysis measures the joint probability that two corpus items (for example, a verb and a construction) will occur together (Gries et al., 2005; Stefanowitsch & Gries, 2003). Thus, unlike delta P, collostructional analysis is not directional (Kyle & Crossley, 2017). For each of these measures, the present study considers the mean association strength scores for types and tokens (see Table 6).

Table 6

The Strength of Association Indices Included in the Present Study from TAASSC (Kyle & Crossley, 2017)

Type of the Association Strength Measure	Description of the Measure
Faith	*Average faith score verb (cue) - construction (outcome) – all* *Average faith score construction (cue) - verb (outcome) – all* *Average faith score verb (cue) - construction (outcome) (types only) – all* *Average faith score construction (cue) - verb (outcome) (types only) – all*
Delta P	*Average delta P score verb (cue) - construction (outcome) – all* *Average delta P score construction (cue) - verb (outcome) – all* *Average delta P score verb (cue) - construction (outcome) (types only) – all* *Average delta P score construction (cue) - verb (outcome) (types only) – all*
Collostructional Analysis	*Average approximate collostructional strength – all* *Average approximate collostructional strength (types only) – all* *Collostruction ratio – all* *Collostruction ratio (types only) – all*

Note: "all" refers to the reference corpus, "all COCA written"

Statistical Analyses

To answer the research question 1 (whether VAC-based indices are predictive of different L2 writing tasks), a multinomial logistic regression analysis was performed because the dependent or response variable was the categorical variable, task (with three levels: descriptive, independent, and integrated). In this model, VAC-based indices were included as the independent variables (i.e., predictors). In the multinomial logistic regression model, "one vs. rest" approach was followed where the odds of each outcome are modelled against all other outcomes (Levshina, 2015, p. 277). To fit the model in this approach, the function "polytomous ()" from the R package "polytomous" was used (Arppe, 2013; Levshina, 2015). Based on Levshina (2015), McFadden's R^2 values (which is analogous to R^2 values in linear regression) from 0.2 to 0.4 indicate a very good fit (which corresponds to 0.7 to 0.9 in linear models, Louviere et al., 2000, p. 55). Hence, in the present study, a R^2 value close to 0.2 (e.g., values above 0.15) will be used as the threshold level for a good effect size. To select the appropriate VAC indices, one-way ANOVA analyses were conducted. In each of those analyses, task (with three levels: descriptive, independent, and integrated) was the independent variable, and one VAC-based measure was the dependent variable. The significance level of the *p*-value was set to .002 after applying a Bonferroni correction. All the VAC indices were transformed into z-scores to maintain the uniformity of scaling. The VAC indices that demonstrated significant differences between the three tasks and were not multicollinear with other indices ($r < .70$) were included as predictor variables in the logistic regression model.

To answer the research questions 2 (whether the relationship between VAC-based indices and writing quality varies depending on the L2 writing tasks) and 3 (whether the relationship between VAC-based indices and writing quality varies depending on the prompts of writing in each task), linear models were developed with the writing scores as the dependent or response

variable in each model. As the observations were independent and the dependent variable was interval-scaled, linear models were used for addressing the research questions 2 and 3. To select the appropriate VAC-based indices as the predictors for the linear models, we first checked whether the VAC indices were normally distributed. Then the correlation of the VAC indices with the writing scores were examined, and all indices with correlation at or below 0.10 were discarded for reporting below a small effect size. The VAC-indices were also checked for multicollinearity, and any variable with correlation with another variable at or above .70 was discarded.

Then to examine whether the relationship between the VAC-based indices and writing quality varies depending on the L2 writing tasks (research question 2), three linear models (models 1, 2, and 3) were developed for the three tasks, descriptive, independent, and integrated, respectively. In each model, selected VAC indices were the fixed factors or independent variables and writing quality scores was the dependent variable. Additionally, three more linear models (models 4, 5, and 6) were developed, one for each task (descriptive, integrated, and independent, respectively), to investigate whether the relationship between the VAC-based indices and writing quality varies depending on the prompts of writing in each task (research question 3). In each of these models (4, 5, and 6), selected VAC predictors, prompts, and interactions between the VAC predictors and prompts were the fixed factors, and writing quality scores was the dependent variable. The reference levels for the categorical variable prompt were "person" in model 4, "fish-farming" in model 5, and "choosing subject" in model 6. The statistical program R was used to run all the analyses (R Core Team, 2015). A predictor was considered significant if the *p*-value was below .05.

Results

Research Question 1: Whether VAC-based Indices are Predictive of Different L2 Writing Tasks

After variable pruning, three VAC-based indices were entered in the multinomial logistic regression model*: average lemma frequency logged-all, average construction frequency logged-all, and average delta p score construction (cue)-verb(outcome)-all*. The output of the model reported a McFadden's R^2 value of 0.176, indicating a good fit (Levshina, 2015). This model also reported measures of accuracy (see Table 7).

Table 7
Cross-table: Measures of Accuracy of the Multinomial Logistic Regression Model

	Descriptive	Independent	Integrated
Descriptive	51	12	7
Independent	31	11	28
Integrated	13	8	49

The accuracy of the model (the total number of correct predictions divided by the number of observations) is reported as 0.53 (chance is 0.33). Table 8 reports the recall (the proportion of instances of each task predicted by the algorithm) and precision (how many times the predictions of descriptive, independent, and integrated texts made by the model were accurate) estimates of the multinomial logistic regression model. The results indicate that about 73% of the descriptive texts and 70% of the integrated texts were accurately predicted. However, independent texts had a lower recall value, only 15%, compared to the other texts. Additionally, the precision value for the descriptive texts was around 53% and for the integrated texts, 58%. Similar to the low recall value, independent texts also had a lower precision estimate (35%) than the other texts.

Table 8
Recall and Precision Estimates

	Descriptive	Independent	Integrated
Recall	0.731	0.150	0.700
Precision	0.532	0.354	0.581

The log odds ratios[2] and *p* values for the multinomial logistic regression model are reported in Table 9.

Table 9

Coefficients (log-odds) and p-values of the Multinomial Logistic Regression Model

	Descriptive		Independent		Integrated	
	Log-odds	*p*-value	Log-odds	*p*-value	Log-odds	*p*-value
Intercept	-34.65	<.001*	-2.45	.476	29.21	<.001*
Average lemma frequency logged-all	4.778	<.001*	0.316	.574	-4.215	<.001*
Average construction frequency logged-all	1.437	.037*	0.066	.908	-1.422	.037*
Average delta p construction (cue)-verb(outcome)-types only-all	32.28	.003*	-8.731	.336	-20.19	.06

Table 9 indicates that the descriptive essays used significantly more frequent constructions and more high frequency lemmas (indicated by the positive log-odds). Additionally, *average delta p construction (cue)-verb(outcome)-types only-all* was significantly predictive of descriptive essays with positive log-odds indicating that verb-construction combinations were more strongly associated in the descriptive essays than in the other essay types. On the contrary, the integrated essays used significantly less frequent lemmas and less frequent constructions (indicated by the negative log-odds) than the other essays.

Research Question 2: Whether the Relationship between VAC-based Indices and Writing Quality Varies Depending on the L2 Writing Tasks

[2] Log odds are centered around 0. Negative log odds indicate that an outcome is less probable than the other. The values of log odds can range from -Infinity (the natural logarithm of 0) to Infinity (Levshina, 2015).

After variable pruning, four VAC-based indices (*average construction frequency logged-all, average construction frequency [types only]-all, , average lemma construction frequency [types only]-all*, and *collostruction ratio [types only] – all* were entered as predictors in models 1, 2, and 3 (for descriptive, independent, and integrated tasks, respectively) to predict writing quality scores. The output of the model 1 (for descriptive task) reported *collostruction ratio [types only] – all* as the single significant predictor of writing quality scores. The positive coefficient of this predictor indicates that for every increase in *collostruction ratio [types only] – all,* the writing quality scores in the descriptive essays increased by 0.04. However, the overall model was not significant, multiple r-squared = 0.078, F (4, 65) = 1.376, p = .252.

Table 10 reports the output of the model 1.

Table 10
Output of the Model 1 (for Descriptive Tasks)

Predictors	Coefficient	Std error	*t*-value	*p*-value
Intercept	5.69	2.294	2.48	.015*
Average construction frequency logged-all	-0.522	0.532	-0.98	.33
Average. construction frequency [types only]-all	<0.001	<0.001	0.114	.909
Average lemma construction frequency [types only]-all	<-0.001	<0.001	-0.761	.449
Collostruction ratio [types only]-all	0.04	0.018	2.128	.037*

Additionally, the output of model 2 (for independent task) reported *average lemma construction frequency (types only)-all* as the only significant predictor of writing quality scores. Table 11 reports the output of the model 2 that explains 24% variance in the writing scores, multiple R squared=0.248, F (4, 65) = 5.362, p= <.001.

Table 11
Output of the Model 2 (for Independent Task)

Predictors	coefficient	Std error	t-value	p-value
Intercept	5.718	2.765	2.068	.042*
Average construction frequency logged-all	-0.389	0.642	-0.607	.545
Average. construction frequency [types only]-all	<0.001	<0.001	-0.312	.755

Average lemma construction frequency [types only]-all	<-0.001	<0.001	-3.117	.002*
Collostruction ratio [types only]-all	0.053	0.034	1.540	.128

As can be seen in the Table 11, for each increase in *average lemma construction frequency [types only]-all*, the scores in the independent task decreased by <-0.001 (indicated by the negative coefficient).

Lastly, the output of model 3 (for integrated task) reported *Average construction frequency [types only]-all as the* significant predictor of writing quality scores, which indicate that for every increase in *Average construction frequency [types only]-all*, the writing scores in the integrated tasks decreased by <-0.001. However, this model was not statistically significant, multiple R-squared=0.071, F (4, 65) = 1.248, p =.299. Table 12 reports the output of the model 3.

Table 12
Output of the Model 3 (for Integrated Task)

Predictors	Coefficient	Std error	*t*-value	*p*-value
Intercept	1.978	2.552	0.775	.441
Average construction frequency logged-all	0.43	0.595	0.724	.472
Average. construction frequency [types only]-all	<-0.001	<0.001	-2.019	.047*
Average lemma construction frequency [types only]-all	<0.001	<0.001	1.023	.310
Collostruction ratio [types only]-all	-0.022	0.044	-0.504	.615

Research Question 3: Whether the Relationship between VAC-based Indices and Writing Quality Varies Depending on the L2 Writing Prompts in Each Task

Three more linear models (models 4, 5, and 6 for the descriptive, integrated, and independent tasks, respectively) were developed to examine whether, within each task, the relationships between the VAC-based indices and L2 writing quality varied depending on the prompts of writing. Each model fitted interactions between prompts and the selected VAC-based

predictors (the same VAC indices used in the models for answering the research question 2) of the writing quality scores. The output of the models 4 and 5 (reported in Tables 13 and 14, respectively) that fitted interactions between prompts and the VAC indices for the descriptive and integrated tasks, respectively, did not report any significant interactions. On the contrary, the output of the model 6 that fitted interactions between prompts and the VAC indices for the independent task reported significant interactions between *average lemma construction frequency (types only)-all* and prompts. This model explained 34% variance in the writing quality scores, multiple R-squared = 0.344, F (9, 60) = 3.501, p =.001. Table 15 reports the results of the linear model 6 with the coefficients, standard errors, t-values, and p-values. The positive coefficient of the significant interaction between the "cooperation" prompt and *average lemma construction frequency (types only)-all* indicates that for every increase in *average lemma construction frequency (types only)-all*, the writing quality scores in the essays on the "cooperation" prompt were significantly higher (<0.001) compared to those on the other prompt ("choosing subject"). Hence, the high scoring essays written on the "cooperation" prompt contained more high frequent lemma construction types than the essays written on another independent task prompt ("choosing subject").

Table 13
Output of the Linear Model 4 for the Descriptive Task

	Coefficient	Std. Error	*t*-value	*p*-value
Intercept	1.961	3.730	0.526	.601
Average construction frequency logged-all	0.417	0.913	0.457	.649
Prompt: Place	3.685	4.733	0.779	.439
Average construction frequency [types only]-all	<-0.001	<0,001	-0.132	.895
Average lemma construction frequency [types only]-all	<-0.001	<0.001	-2.085	.041*
Collostruction ratio [types only]-all	0.041	0.025	1.652	.103
Prompt: Place x Average construction frequency logged-all	-0.806	1.131	-0.713	.478

Prompt:Place x Average construction frequency [types only]-all	<-0.001	<0.001	-0.198	.843
Prompt: Place x Average lemma construction frequency [types only]-all	<0.001	<0.001	1.355	.180
Prompt: Place x Collostruction ratio [types only]-all	0.003	0.037	0.098	.922

Note. Prompt: Person = reference

Table 14

Output of the Linear Model 5 for the Integrated Task

	Coefficient	Std. Error	*t*-value	*p*-value
Intercept	5.859	3.302	1.774	.081
Average construction frequency logged-all	-0.233	0.748	-0.312	.756
Prompt: Migrating birds	-8.634	5.322	-1.622	.110
Average construction frequency [types only]-all	<-0.001	<0.001	-1.742	.086
Average lemma construction frequency [types only]-all	<0.001	<0.001	0.651	.517
Collostruction ratio [types only]-all	-0.156	0.097	-1.607	.113
Prompt: Migrating birds x Average construction frequency logged-all	1.517	1.237	1.226	.225
Prompt: Migrating birds x Average construction frequency [types only]-all	<0.001	<0.001	0.459	.647
Prompt: Migrating birds x Average lemma construction frequency [types only]-all	<0.001	<0.001	0.957	.342
Prompt: Migrating birds x Collostruction ratio [types only]-all	0.176	0.111	1.588	.117

Note. Prompt: Fish-farming= reference

Table 15

Output of the Linear Model 6 for the Independent Task

	Coefficient	Std. Error	*t*-value	*p*-value
Intercept	1.424	5.155	0.276	.783
Average construction frequency logged-all	0.738	1.225	0.603	.549
Prompt: Cooperation	9.667	6.257	1.545	0.127
Average construction frequency [types only]-all	<-0.001	<0.001	-1.001	.320
Average lemma construction frequency [types only]-all	<-0.001	<0.001	-3.172	.002*
Collostruction ratio [types only]-all	0.089	0.082	1.083	.282
Prompt: Cooperation x Average construction frequency logged	-2.532	1.481	-1.709	.092
Prompt:Cooperation x Average construction frequency [types only]-all	<0.001	<0.001	1.212	.230
Prompt: Cooperation x Average lemma construction frequency [types only]-all	<0.001	<0.001	2.161	.034*

Prompt: Cooperation x Collostruction ratio [types only]-all	-0.084	0.091	-0.92	.361

Note. Prompt: Choosing subject = reference

Discussion

VAC-based Indices in Different L2 Writing Tasks

The purpose of the first research question was to investigate whether VAC-based indices are predictive of different L2 writing tasks. The output of the multinomial logistic regression analysis showed that with each increase in *average lemma frequency logged-all* and *average construction frequency logged-all,* the chances of a text being descriptive were significantly higher by 4.778 and 1.437, respectively (indicated by the positive log odds). Hence, compared to the other texts, the descriptive texts contained significantly more high frequency lemmas and constructions (that were more frequent in the COCA corpus). This finding indicates that descriptive essays might be less difficult than the other types of essays because frequent words and constructions are likely to be learned earlier or more easily than less frequent constructions (Ellis & Ferreira-Junior, 2009a). This result also underscores the findings of Way et al. (2000) who found descriptive writing task to be easier (as they had higher accuracy and lower syntactic complexity scores) among L2 learners of French compared to narrative and expository writing tasks. Additionally, the learners in the present study needed to draw on their own personal experiences while writing the descriptive essays on short prompts (e.g., “describe your family” and “describe your friend”), and such simple topics may not have primed writers to use sophisticated academic writing skills (Connor-Linton & Polio, 2014). However, these simple topics may still elicit verb-construction combinations that are strongly associated in common usage. Indeed, the findings show that with each increase in *average delta p construction (cue)-verb(outcome)-types only-all*, the chances of an essay being descriptive were significantly higher

(by 32.28) compared to the other essays. Thus, the descriptive essays were distinct from the other essays in the use of verb-VAC combinations that were strongly associated (in the COCA corpus).

The results also showed that with each increase in *average lemma frequency logged-all* and *average construction frequency logged-all*, the chances of integrated essays were significantly lower (by -4.215 and -1.422, respectively) compared to the other essays. Thus, integrated texts used significantly less frequent lemmas and constructions than the other types of texts. Since integrated writing tasks involve extensive borrowing from source materials (Cumming, 2013; Plakans & Gebril, 2013) that may not involve high frequent vocabulary and syntactic structures (Leki & Carson, 1997), the use of less frequent (and hence, more complex or sophisticated) constructions in the integrated essays might result from textual borrowing. Additionally, such textual borrowings might be influenced by writers' proficiency levels, and previous studies have shown that high proficiency students use higher amount of source text in integrated writing tasks than learners of lower proficiency levels who have difficulty in comprehending source texts (Cumming et al., 2005; Gebril & Plakans, 2009). Hence, the use of less frequent constructions in the integrated essays may also be indicative of the higher complexity or difficulty of this task because writers need sufficient comprehension of source materials to write effectively about them (Cumming, 2013; Gebril & Plakans,2009). Thus, VAC-based indices may suggest a different way of assessing text difficulty in addition to the measures usually used in literature, for example, readability formulas based on lexical, semantic, and syntactic features of texts (Crossley et al., 2008, 2011) and variables related to psycholinguistic models of text comprehensibility (e.g., measures of text cohesion and meaning construction) (Crossley et al., 2008).

However, none of the VAC indices were strong predictors of independent essays wherein the writers argued on a given topic. Developing and defending a persuasive argument is a challenging feature of L2 writing (Lee & Deakin, 2016; Silva, 1993), and previous research found higher syntactic complexity, measured by absolute complexity indices, in L2 argumentative essays compared to other essay types (e.g., more clauses per T-unit in Beers & Nagy, 2011 and more complex noun phrases in Ravid & Berman, 2010). In the present study, the VAC indices were based on not only academic but also all other written sections of COCA (fictions, magazines, and newspapers), and hence, such indices might not reflect the kind of complexity characteristic of the independent argumentative essays but, as the present findings show, were predictive of the descriptive essays that might have been less complex compared to the argumentative essays.

Effects of the Writing Tasks on the Relationship between VAC-based Indices and L2 Writing Quality

The second research question examined whether the relationship between VAC-based indices and L2 writing quality varies depending on the writing tasks. The findings show that VAC-based predictors of L2 writing quality vary depending on the writing tasks. For the independent task, it was found that highly scored essays contained significantly less frequent verb-VAC combination types. This output provides support for Kyle and Crossley (2017) who also found that highly scored independent essays contained less frequent verb-VAC combinations. Additionally, in the present study, the VAC predictors explained a higher amount of variance (24%) in the holistic writing scores of independent essays than the variance (14%) explained in Kyle and Crossley (2017). Furthermore, the present study also examined the relationship between VAC indices and writing quality in descriptive and integrated tasks.

Although for these two tasks, the linear models were not significant, the present findings showed that highly scored descriptive essays contained verb-VAC combinations that were strongly associated in the COCA corpus, and highly scored integrated essays contained construction types that were less frequent in the COCA corpus.

Thus, the relationship between VAC-based syntactic sophistication indices and L2 writing quality may vary depending on the writing tasks (descriptive, independent, integrated). Therefore, writing task type seems to play an important role in determining whether high proficiency L2 writers use low frequency (hence, more sophisticated; Ellis & Ferreira-Junior, 2009a) lemma-constructions and strongly associated verb-VAC combinations in their essays. Similar to the present findings, Beers and Nagy (2009) also found that the relationship between structural or absolute complexity measures (e.g., words per clause and clauses per T-unit) and writing quality scores varied in different types of essays (narrative and persuasive).

Mediating Effects of the Writing Prompts on the Relationship between VAC-based Indices and L2 Writing Quality in Each Task

The third research question examined whether the relationship between VAC-based indices and L2 writing quality varies depending on the writing prompts in each task. The results showed no significant interactions between the VAC indices and prompt for either descriptive or integrated essays. This finding indicates that prompts likely played no significant role in the production of VAC features for these two tasks. There was a single significant interaction between prompt and *average lemma construction frequency (types only)-all* for independent essays. This finding suggests that the essays written on the “cooperation” prompt used more high frequent lemma-construction combinations (types only), which may have allowed for higher scores in those essays compared to the essays written on a different prompt (“choosing subject”).

For each prompt in the independent task, the writing scores were normally distributed. Thus, it might be argued that the writing prompt might have been responsible for the production of high frequent lemma-construction combinations types in highly scored independent essays.

Implications

Whereas previous studies on L2 writing often focused on absolute or structural complexity indices (e.g., Lu, 2011), the present study examines complexity in different L2 writing tasks from usage-based perspective, and the findings have implications for L2 writing instructors, raters, and other stakeholders. For example, whereas structural complexity features are often emphasized in instructional contexts, the present findings provide evidence that different L2 writing tasks might have varied levels of difficulty based on the inclusion of high or low frequency constructions, and such understanding may facilitate pedagogical practices for L2 writing tasks. Additionally, the current findings might inform test developers, administrators, and raters that there might be different indicators of writing proficiency related to the frequency of linguistic constructions and the closeness between particular verbs and constructions in different L2 writing tasks. The findings may also inform test-takers of the importance of including low frequency and strongly related verb-constructions for scoring higher in varied types of L2 writing tasks. Similarly, the results of the present study may lead L2 writing programs to emphasize tasks and prompts while considering whether L2 writing samples should include low frequency verbs, constructions, and strongly associated verb-VAC combinations.

Limitations and Future Directions

The present study has some limitations. Although the study used three different corpora, a small number of essays (n=70) was selected from each corpus (to ensure that the sample sizes are comparable across the three tasks), which might limit the generalizability of the conclusions

drawn from the analyses. Because of such small sample size, the relevant analyses may also be underpowered. However, while extracting the TOEFL independent and integrated sub-corpora, we made sure that the scores of the extracted samples follow similar distribution as that of the original TOEFL corpora, and thus, the two TOEFL sub-corpora are representative of the original TOEFL corpora. Replication studies are warranted to examine if generalizability of the present findings holds. Additionally, in the present study, the VAC-based complexity indices were calculated based on COCA, which may not be representative of the language experiences to which the L2 writers were exposed (Monteiro et al., 2018). Moreover, syntactic complexity in L2 writing might be influenced by variables such as L1 background (Lu & Ai, 2015; Ortega, 2003, 2015; Römer et al., 2014). However, it was beyond the scope of the present study to control for L1 background in the selection of the study participants. Future research should control for L1 background of participants to assess the relative effects of linguistic background on VAC-based syntactic sophistication indices. Additionally, in the present study, there was no control on textual borrowing in the integrated tasks that might affect the linguistic features in those tasks. Moreover, the descriptive essays were scored using the rubric for independent argumentative essays, whereas the use of a rubric specifically designed for descriptive writing would have been more desirable. Furthermore, in addition to language tasks, time might be another variable over which VAC-based complexity indices might vary since traditional absolute complexity indices have been found to develop longitudinally (e.g., Bulté & Housen, 2014). Although a few studies have focused on longitudinal analyses of VACs (e.g., Ellis & Ferreira-Junior, 2009a), there needs to be more investigations into longitudinal development in VAC-based complexity indices, especially those that control for prompt and task. Future studies should also investigate VAC-

based indices in a wider variety of writing tasks including narrative and compare/contrast essays that are also commonly used in L2 learning contexts.

Conclusion

In contrast to most studies that examined syntactic complexity in L2 writing using traditional or absolute complexity measures (Bulté & Housen, 2014; Lu, 2011; Wolfe-Quintero et al.,1998), the present study investigated complexity in L2 writing using VAC-based indices that align with usage-based approaches to language learning. The findings have implications about how different L2 writing tasks may lead to different levels of VAC-based feature production, which may be indicative of the varied difficulty levels of those tasks. Additionally, although high proficient L2 writers are theoretically expected to use verb-construction combinations that are strongly associated in the input (Kyle & Crossley, 2017), the findings of the present study imply that the relationships between the association strengths of verb-VAC combinations and writing quality may vary depending on the L2 writing tasks with few prompt-based effects. Hence, the findings suggest that the nature of writing tasks and prompts are important mediating variables that may influence how frequency and association-strength based complexity indices are related to L2 writing quality.

References

Arppe, A. (2013). Polytomous: Polytomous logistic regression for fixed and mixed effects. R package version 0.1.6. Retrieved from https://CRAN.R project.org/package=polytomous

Beers, S. F., & Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre? *Reading and Writing*, *22*(2), 185-200.

Beers, S. F., & Nagy, W. E. (2011). Writing development in four genres from grades three to seven: Syntactic complexity and genre differentiation. *Reading and Writing*, *24*(2), 183-202.

Behrens, H. (2009). Usage-based and emergentist approaches to language acquisition. *Linguistics*, *47*(2), 383-411.

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.

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.

Bulté, B., & Housen, A. (2012). Defining and operationalizing L2 complexity. In A. Housen, I. Vedder, & F. Kuiken (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA* (pp. 21-46). Amsterdam: John Benjamins Publishing Company.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42-65.

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

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

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a foreign language*, *23*(1), 84-101.

Crossley, S. A., Weston, J. L., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, *28*(3), 282-311.

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.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly*, *10*(1), 1-8.

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. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*(4), 447-464.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143-188.

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.

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.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2013). Usage-based language: Investigating the latent structures that underpin acquisition. *Language Learning*, *63*(s1), 25-51.

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Language usage, acquisition, and processing: Cognitive and corpus investigations of construction grammar*. Malden, MA: Wiley-Blackwell.

Eskildsen, S. W. (2008). Constructing another language—Usage-based linguistics in second language acquisition. *Applied Linguistics*, *30*(3), 335-357.

Gebril, A., & Plakans, L. (2009). Investigating source use, discourse features, and process in integrated writing tests. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *7*(1), 47-84.

Goldberg, A. (1995). *Construction: A construction grammar approach to argument structure.* Chicago, IL: The 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, UK: Oxford University Press.

Goldberg, A. E., Casenhiser, D., & White, T. R. (2007). Constructions as categories of language. *New Ideas in Psychology*, *25*(2), 70-86.

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.

Gries, S. T., & Wulff, S. (2005). Do foreign language learners also have constructions? *Annual Review of Cognitive Linguistics*, *3*(1), 182-200.

Hinkel, E. (2002). *Second languages writers' text: Linguistic and rhetorical features.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*(2), 237-263.

Housen, A., De Clercq, B., Kuiken, F., & Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second Language Research*, *35*(1), 3-21.

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* (Unpublished doctoral dissertation). Georgia State University.

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, *34*(4), 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*(2), 333-349.

Lee, J. J., & Deakin, L. (2016). Interactions in L1 and L2 undergraduate student writing: Interactional meta-discourse in successful and less-successful argumentative essays. *Journal of Second Language Writing*, *33*, 21-34.

Leki, I., & Carson, J. (1997). “Completely different worlds”: EAP and the writing experiences of ESL students in university courses. *TESOL Quarterly*, *31*(1), 39-69.

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). *Stated choice methods: Analysis and applications*. Cambridge: Cambridge university press.

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.

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, *29*, 16-27.

Monteiro, K. R., Crossley, S. A., & Kyle, K. (2018). In Search of New Benchmarks: Using L2 Lexical Frequency and Contextual Diversity Indices to Assess Second Language Writing. *Applied Linguistics*, 1-22.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555-578.

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.

Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing*, *29*, 82-94.

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*(1), 117-134.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing*, *13*(2), 111-129.

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, *26*(4), 561-587.

Plakans, L. (2010). Independent vs. integrated writing tasks: A comparison of task representation. *TESOL Quarterly*, *44*(1), 185-194.

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing*, *22*(3), 217-230.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Ravid, D., & Berman, R. A. (2010). Developing noun phrase complexity at school age: A text-embedded cross-linguistic analysis. *First Language, 30*, 3–26.

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.

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.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing.* Cambridge, England: UCLES/Cambridge University Press.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly*, *27*(4), 657-677.

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, *8*(2), 209-243.

Tedick, D. J. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes*, *9*(2), 123-143.

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

Way, D. P., Joiner, E. G., & Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of French. *The Modern Language Journal*, *84*(2), 171-184.

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.

Yang, H. C., & Plakans, L. (2012). Second language writers' strategy use and performance on an integrated reading-listening-writing task. *TESOL Quarterly*, *46*(1), 80-103.