Language Learning & Technology
ISSN 1094-3501

February 2021, Volume 25, Issue 1
pp. 53–72

ARTICLE

Lexical complexity, writing proficiency, and task effects in Spanish Dual Language Immersion

Erin Schnur, Cambly

Fernando Rubio, University of Utah

Abstract

Using the 2.4-million-word written Spanish subsection of the Corpus of Utah Dual Language Immersion, collected from a large-scale standardized writing assessment of K-12 Spanish second language (L2) students, this study focuses on lexical complexity as operationalized by three measures: lexical diversity, lexical density, and lexical sophistication. The study goes beyond most previous work on lexical complexity by investigating the effect of task type on all three measures of lexical complexity. Patterns in variation are identified across proficiency levels and between task types. Results show that all three measures increase at each proficiency score between Novice High and Advanced, except at Intermediate Mid where scores dip or flatten. Diversity and sophistication are both shown to increase rapidly after this mid-point, indicating that a broad and deep lexical repertoire is a key feature of more advanced proficiency levels. Results for the effect of task indicate that text genre impacts learners' lexical density, while tasks that are more complex elicit higher lexical sophistication.

Keywords: *Corpus, Second Language Acquisition, Vocabulary, Writing*

Language(s) Learned in This Study: *Spanish*

APA Citation: Schnur, E., & Rubio, F. (2021). Lexical complexity, writing proficiency, and task effects in Spanish Dual Language Immersion*. Language Learning & Technology*, *25*(1), 53–72. https://hdl.handle.net/10125/73425.

Introduction

The lexical characteristics of second language (L2) writers are of interest to both Second Language Acquisition (SLA) researchers and language teaching professionals. Measures of lexical complexity inform our understanding of development and proficiency, influence the design and rating of major international language tests, and have implications for the design of pedagogic materials and classroom practices. Patterns in the expression of learners' lexical repertoires as their proficiency advances not only provide benchmarks for measuring advancement but also illuminate key elements that distinguish one proficiency level from the next. In other words, understanding the lexical differences between language produced by learners at different levels can help educators make those decisions about lexical pedagogy that best help students advance.

The effect of task, however, has largely been ignored in discussions of the relationship between lexical complexity and proficiency. Polio and Park (2016), after providing an extensive review of recent studies on language development in writing, conclude, among other things, that we need to "improve quantitative research by controlling for writing task" (p. 300). There is ample research that supports the assumption that different levels of task complexity impose different requirements on the writer and therefore elicit more or less complex language. Similar conclusions have been reached regarding the impact that the genre or type of writing elicited by a task (e.g., descriptive, narrative, argumentative) can have on written production. Yet while measures of lexical complexity remain a staple of evaluation of learner development, little research addresses the effect of task type on these measures.

This study aims to investigate the relationship between lexical complexity and both proficiency score and task type. Using a corpus comprised of 2.4 million words of writing produced by L2 Spanish students in Utah's Dual Language Immersion (DLI) program during ACTFL proficiency testing, we seek to extend the discussion of lexical complexity and its implications for researchers and educators by first investigating how three observations of lexical complexity (diversity, density, and sophistication) vary by proficiency score. Because our data were collected from large-scale proficiency testing, they provide a unique opportunity to examine lexical complexity, proficiency, and task effect in a highly standardized setting with standardized prompts administered across learners at a variety of proficiency levels. In addition to investigating the patterns in lexis across proficiencies, we also consider differences in task type, comparing texts produced in response to more and less complex tasks across the same three measures. Our findings shed light on how lexical complexity develops and how task characteristics impact its expression and point to pedagogical implications for language educators.

Background

ACTFL Proficiency Testing in Utah

Utah's DLI program administers the ACTFL Assessment of Performance toward Proficiency in Languages (AAPPL) as an evaluation of students' language progress. The AAPPL assesses learners across three modes of communication and students receive a separate rating in Interpersonal Listening/Speaking, Presentational Writing, Interpretive Reading, and Interpretive Listening. This study focuses on learners' performance on the writing component of the AAPPL test, which is administered to students in the Utah DLI program in grades four, six, eight, and nine.

AAPPL scoring is conducted by trained raters and is determined for the entirety of the writing that a test-taker produces. In other words, human raters score each response and a global rating is assigned via an algorithm that takes all ratings into account.

Ratings are based on the ACTFL Performance Descriptors for Language Learners, which evaluate proficiency across functions, contexts, and text types, and consider the learner's language control, vocabulary, communication strategies, and cultural awareness. AAPPL scores run from Novice Low to Advanced, with test form A discriminating between proficiency levels from Novice Low to Intermediate Mid, and test form B discriminating between Novice High and Advanced (see Table 1). The AAPPL rating scale uses more sublevels than the ACTFL scale (four in the Novice range and five for the Intermediate range) so, for clarity purposes, in this paper we use abbreviations to refer to each level and the corresponding sublevels as indicated by the codes in Table 1.

Table 1. *AAPPL Rating Scale*

Code	AAPPL Score	Proficiency Level (CEFR equivalent)	Form A	Form B
NL	N1	Novice Low (0)	Form A Range	
NM-Low	N2	Novice Mid (0)	▨	
NM-High	N3	Novice Mid (0)	▨	
NH	N4	Novice High (A1)	▨	Form B Range
IL	I1	Intermediate Low (A2)	▨	▨
IM-Low	I2	Intermediate Mid (B1.1)	▨	▨
IM-Mid	I3	Intermediate Mid (B1.)	▨	▨
IM-High	I4	Intermediate Mid (B1.1)	▨	▨
IH	I5	Intermediate High (B1.2)		▨
A	A	Advanced (B2)		▨

In addition to the scores represented in Table 1, students whose writing does not meet the requirements sufficient to earn the lowest score for each form are assigned a rating of “below” that score (“below N1” for form A, “below N4” for form B).

ACTFL scoring can be correlated with CEFR ratings (ACTFL, n.d.), as indicated in Table 1.

Lexical Complexity

Researchers have established that measures of lexical complexity—that is, the size, variety, and quality of a learner’s vocabulary—are a good predictor of writing quality (e.g., Crossley et al., 2012; Yu, 2010). Previous work on lexical complexity has broadly defined the construct as consisting of two theoretical parts: systemic complexity (breadth) and structural complexity (depth) (Bulté & Housen, 2012; Skehan, 2003). These two theoretical subconstructs have been investigated using a variety of measures including those of fluency and compositionality, however three main types of measurements comprise most lexical complexity research: measures of lexical diversity, density, and sophistication (Read, 2000).

Lexical diversity is a measure of the number of different words in a writer’s lexical repertoire (Laufer & Nation, 1995), and informs our understanding of systemic complexity. Research into the relationship between lexical diversity and proficiency has broadly established a positive correlation between the two constructs (e.g., Jarvis, 2002; Yu, 2010) for both speaking and writing, although much discussion surrounds its measurement. As McCarthy and Jarvis note in a 2007 study comparing 13 lexical diversity indices, there is a great need to analyze lexical diversity; it is often measured, and yet “the formulation of a fully reliable and valid LD measure has proven to be elusive” (p. 460).

The fundamental difficulty of measuring lexical diversity is its instability over texts of varying lengths. The most well-known measure is type-token ratio (TTR), however a variety of other measures have been proposed and tested. One frequently used measure is *D* (Malvern et al., 2004). The use of *D* has been demonstrated to measure developmental trends in lexical diversity for short texts and across multiple types of language, including that produced by children and adults, L2 and native speakers, and in academic and non-academic contexts (Durán et al., 2004). Although claims that *D* is suitable for texts of varying lengths have been called into question, it has been shown to outperform many other indices of lexical diversity (McCarthy & Jarvis, 2007).

Lexical density is a measure of the proportion of lexical words to the total number of words in a text (Ure, 1971) and represents one aspect of systemic complexity. Early studies on lexical density compared written and spoken texts, generally finding that writing contains a higher degree of lexical words (e.g., Ure, 1971; Halliday, 1989). Density measures, both calculating a summary density and examining the proportion of individual lexical parts of speech, have also been shown to distinguish registers of both speech (e.g., Stubbs 1986) and writing (e.g., Biber, 1991). Biber et al. (2006) elaborate on differences in informational density and the proportion of parts of speech in Spanish. Their multidimensional analysis revealed that lexical parts of speech vary according to register and text function (narration, informational reports of past events, *irrealis*, etc.). Based on these findings, we posit that lexical density is likely to differ between task types in L2 writing. However, previous research has consistently found little to no statistically significant relationship between lexical density and L2 proficiency level (e.g., Crossley & McNamara, 2009; Lu, 2011).

Lexical sophistication is defined as the proportion of low-frequency words in a text, “rather than just general, everyday vocabulary” (Read, 2000, p.200). Lexical sophistication informs our understanding of both systemic complexity (breadth) and structural complexity (depth), as measures of sophistication by necessity capture information about both the variety and level (advanced vs. common) of words represented in a text. Results of measurements of lexical sophistication are highly dependent on researchers’ definitions of advanced words. Most research in this area relies on frequency lists, such as the General Service Wordlist for English (West, 1953), to define common lexical items, and has generally found correlations between lexical sophistication and proficiency scores (e.g., Crossley & McNamara, 2009; Laufer & Nation, 1995).

Lexical density, diversity, and sophistication are frequent components in the proficiency scales of major international standardized language tests, including the AAPPL, where performance descriptors refer to the extensiveness of a learner’s vocabulary (ACTFL, 2012). Much recent research has focused on establishing effective, reliable measurements of lexical complexity and exploring the extent to which different measures correlate with writing quality (e.g., Crossley et al., 2012; McCarthy & Jarvis, 2007). Researchers have also focused on the development of tools for the automatic measurement of these indices (e.g., Kyle & Crossley, 2015) and explored their application for automatic grading (e.g., Crossley & McNamara, 2013).

Recent research has additionally focused on expanding the types of lexical information that is measured and considered as part of lexical complexity as it relates to learner proficiency and development. Researchers have investigated multiword units and their frequencies (Bestgen & Granger, 2014; Staples & Reppen, 2016), fiction-like versus academic vocabulary as a component of lexical sophistication (Durrant & Brenchley, 2019), and psycholinguistic word properties such as concreteness and familiarity (Crossley & Skalicky, 2019). While indices based on these lines of research are included in some automatic lexical assessment tools, most notably Kyle et al.’s (2018) Tool for the Automatic Analysis of Lexical Sophistication, at this time, such tools are largely limited to processing English texts.

Task Effects

An issue that has complicated the interpretation of results of many studies on the development of lexical complexity is the effect of task complexity, which we will refer to as task effect. Task complexity has been primarily considered from two different theoretical perspectives: Robinson’s (2011) Cognition Hypothesis and Skehan’s (2014) Limited Attentional Capacity Model. Both hypothesize how different features of a task may influence output by directing attention to or away from specific aspects of language production. However, both models differ crucially as to whether they postulate the existence of only one or several possible sources of attention available to the learner. As its name implies, the Limited Attentional Capacity Model follows a single-source view of attention. In this model, a learner’s limited attentional resources will be taxed by more complex tasks. As a result, learners will focus their attention on the content of the task and less attention will be available to focus on linguistic form. The model therefore predicts that more complex tasks will elicit less complex language. The Cognition Hypothesis, on the other hand, posits the existence of different pools of attentional resources that learners may draw on as they focus on form and meaning. Consequently, an increase in task complexity will not necessarily have a negative effect on language output. In this model, Robinson posits the existence of two types of variables that can affect the complexity of a task:

Resource-dispersing variables: This dimension refers to variables that place procedural demands on the learner, such as planning time or familiarity with the task or topic. An increase in these variables forces the learner to direct attentional resources away from the language code, which results in less complex output.

Resource-directing variables: The resource-directing dimension includes variables that pose conceptual demands on the learner, for example whether the task requires learners to describe past or present events, or whether it requires referencing few or many elements. These variables push learners to pay attention to forms needed to meet task demands, which results in higher levels of lexical and syntactic complexity and better accuracy, although often at the expense of fluency.

There have been a number of studies that analyze the effects of task complexity on L2 speech production (Iwashita et al., 2001; Rahimpour, 1999), but only limited research to date has studied task effects on writing, specifically on measures of lexical complexity. Researchers have measured the effects of task modifications to test the predictions made by Robinson’s and Skehan’s theories, but the results do not lead to any clear conclusions. Ishikawa (2007) investigated the effect of manipulating task complexity along Robinson’s [±Here-and-Now] dimension (a Resource-Directing variable) on the level of complexity (lexical and syntactic), accuracy and fluency of L2 learners’ written production. The measures of lexical complexity included lexical density and type/token measures. Syntactic complexity was measured using T-Unit related measures. The author worked with 54 Japanese high school students of English who were divided into two groups of comparable proficiency levels and asked to write a narrative based on a cartoon in either present [+Here-And-Now] or past [-Here-And-Now] tense. Learners in the more complex [-Here-And-Now] condition wrote responses that showed higher levels of lexical and syntactic complexity.

Kormos (2011) examined the effect of task complexity on several lexical, syntactic and cohesive features of writing. The participants in her study, secondary-school L2 learners of English, were divided into two groups that were given writing tasks requiring the production of a past narration based on a cartoon. One group was asked to write a narrative based on a series of pictures that formed a coherent story line. Since the plot of the story was already provided by the pictures and students only had to describe it, this task was considered low in complexity. The second group had to create a story based on six unrelated pictures. In this case, in addition to the writing task, students had the added cognitive demand of developing a plot which connected the pictures, making this task more cognitively demanding. Results showed that the level of task complexity did not affect the students' accuracy or syntactic complexity, a finding that the author attributes perhaps to the fact that both tasks required learners to write in the same genre (narration). One measure of lexical complexity showed a significant difference between the two groups: the group that produced the picture narration used more complex vocabulary than the group that produced the cartoon description. This finding might lend support to the Cognition Hypothesis although the other measures of lexical complexity used in the study (*D*-value, frequency of content words and concreteness of content words) showed very similar levels between the two groups, which runs counter to the predictions of the Cognition Hypothesis.

Frear and Bitchener (2015), in a partial replication of previous studies, examined the effects of task complexity on lexical variety and syntactic complexity. Their subjects were a group of 34 English learners of different L1s who were studying in New Zealand. The learners were given three writing tasks of increasing complexity by manipulating the type and amount of information provided in the task instructions. Lexical variety was measured using type-token ratio, while the ratio of dependent clauses to T-Units was used as a measure of syntactic complexity. The authors found that lexical variety increased when the level of task complexity was increased, but there was no change in syntactic complexity.

Working in the context of a university in Iran, Rahimi and Zhang (2018) studied the written production on an argumentative task of 80 L2 learners of English of upper-intermediate proficiency and good writing ability. Learners were randomly assigned to four groups. Two writing tasks were used that varied in complexity depending on the level of reasoning required and the number of elements included in the task description. Each writing task was administered under two pre-task planning conditions: One with ten minutes of pre-task planning and one with no pre-task planning time. The results of lexical complexity measures, which included lexical diversity and sophistication, showed that sophistication improved significantly in the complex task under both pre-task conditions, but neither task complexity, nor pre-task planning had any effect on learners' level of lexical diversity. As the authors point out, some of the results lend support to the Cognition Hypothesis while other findings are more in line with the predictions of the Limited Attentional Capacity Model.

As this brief review of research indicates, the results of studies that look at task effects on writing have been mixed. This may be because there are several factors that can affect output including the level of cognitive complexity of a task, the planning conditions, and the type of writing elicited. In fact, there may be an interaction between task complexity and writing genre that would need to be isolated in order to make any valid claims about task effects. For example, Skehan (2009) discusses the validity of Robinson's (2007) claim that the [–Here-And-Now] condition is less complex than the [+Here-And-Now] condition and argues that "there is something of a genre difference between the two conditions which significantly complicates their comparison". Writing genre, or the type of writing elicited by a task, has also been identified as a variable that can influence writing complexity. For example, Staples and Reppen (2016) emphasize "the importance of comparing student writing across various genres to understand students' use of language for functions such as argumentation, narration, and conveying information" (p. 19). However, much research in this area has focused on syntactic complexity rather than lexical (e.g., Lu, 2011) along with other measures of writing quality such as accuracy and fluency (e.g., Way et al., 2000).

In sum, this body of research has posited several models of task effect on language production. Research on lexical complexity has established the relationship between diversity and sophistication and proficiency, and the relationship between density and text register, but the effects of task type and genre on lexical complexity have not yet been fully explored. By investigating its relationship to both proficiency and task type in combination, we can more fully interpret measures of lexical complexity.

The goal of the present study is to investigate lexical complexity, both in its relationship to proficiency level and task type. Specifically, we address the following research questions:

1. To what extent do measures of lexical complexity vary between texts with different proficiency scores? Is there an identifiable pattern in this variation as writers progress from Novice High to Advanced?
2. To what extent does task type impact lexical density and lexical sophistication? Is there an identifiable pattern across levels?
3. To what extent can information about the frequencies of lexical word classes elaborate our understanding of difference in lexical complexity both by level and task?

Methods

Corpus Collection

This study is based on the analysis of the Spanish subcorpus within the Corpus of Utah Dual Language Immersion (CUDLI; Rubio & Schnur, 2019-). CUDLI is a new corpus that comprises almost 75,000 written texts (approx. 5 million words) produced by 12,339 students enrolled in Utah's Dual Language Immersion (DLI) program in four foreign languages: Spanish, German, Portuguese, and Chinese. The texts are responses to the presentational writing portion of the AAPPL assessment. The corpus was collected from results of testing in the 2017-2018 school year and contains all responses to the written portion of the test. Because the corpus represents an entire population, rather than a sample, it is not balanced for level, grade, etc.

AAPPL Presentational Writing Test

The presentational writing portion of the AAPPL includes six prompts. Prompts are determined by administration of one of two test forms (see Table 1). Writing prompts were static for each test form so that, during the test event from which the corpus data was collected, all students who took each test form answered the same six prompts. For this study, we targeted the Form B subset of the Spanish subcorpus of CUDLI, which was administered to students in grades six, eight, and nine. This enabled us to focus on differences among and between Novice High, Intermediate, and Advanced learners across the same six tasks. We excluded texts produced by learners whose writing proficiency was scored at the "below N4" level because these texts represent the entire range of proficiency levels that cannot be determined by test form B (from below N1 to N3). The subset of the corpus that remained was comprised of 20,915 texts produced by 3,486 learners, however this number was reduced by corpus cleaning methods.

Corpus Cleaning

Because CUDLI represents a new data set, the process of cleaning the corpus files is still underway. Text files in the corpus were collected directly from student-typed data and therefore contained not only spelling and grammar errors, but also responses typed in English (counter to test instructions), responses that contained significant code-switching, and nonsense responses (e.g., when a student hit random keys on their keyboard in lieu of a response).

We undertook the following procedures to identify and remove as many of these files as possible from our data set. Blank and nonsense texts were identified by their word count. All files containing zero words were removed, and files containing fewer than ten words were manually inspected to determine if they represented language or random characters. A Perl program was used to identify texts containing common English words (*the*, *and*, and *this*), which were then examined manually and removed if they were comprised primarily of English.

Finally, the most frequently occurring spelling and typographic errors were identified and corrected. This process involved part of speech (POS) tagging using the Spanish dictionary included in the Child Language Analysis (CLAN) software, developed by the Child Language Exchange System (CHILDES) program, which was used because of its integration with other CLAN programs, such as morphological analysis and VOCD calculations. The *freq* program within CLAN identified all words that the tagger could not recognize. These words were sorted by frequency count, and untaggable words occurring more than 100 times in the corpus were manually examined. Cases where the target word form could clearly be identified and the untaggable word represented a misspelling (ablar → hablar), misuse of diacritic(s) (tambien → también), or other typographic errors (megusta → me gusta, méxico → México) were identified and corrected throughout the corpus.

Final Corpus Composition

Once the cleaning process was completed, our final dataset included 20,102 texts produced by 3,128 learners, totaling 2,353,386 words. Table 2 summarizes the final subcorpus by proficiency level.

Table 2. *Distribution of Subcorpus by Proficiency Level*

Proficiency Level	**Learners**	**Texts**	**Words**
NH	211	1,153	63,805
IL	380	2,277	152,256
IM-Low	223	1,338	99,480
IM-Mid	244	1,459	114,419
IM-High	1,264	7,583	901,443
IH	615	3,686	540,229
A	435	2,606	481,754
Totals	3,128	20,102	2,353,386

Corpus Analysis

Identification of Task Types

In order to investigate the effect of task type on lexical complexity, all six test form B prompts were examined and classified according to the assumed cognitive complexity of the task and the type of writing they elicited. One of the prompts elicited questions from learners, one combined both descriptive and narrative elements, and two each were determined to be primarily narrative or descriptive. In examining task effect, only responses to these four prompts were used for analysis, creating a smaller dataset for this research question (Table 3).

For the purposes of this study, we considered the descriptive tasks to be less complex than the narrative tasks due to differences in both procedural and conceptual demands. The descriptive tasks required learners to respond to prompts with a basic description in the present tense followed by a brief explanation. The narrative tasks required a brief description in present tense, a past tense narration, and a future tense narration. Narrative tasks were considered more complex from a procedural perspective because they included more elements and covered topics that were more removed from the students' immediate experience (e.g., talking about future plans). From a conceptual perspective, the narrative tasks were also more complex because they required control of three different tenses and the corresponding morphological markers. In sum, following Robinson's (2011) Cognition Hypothesis, the narrative tasks involved more resource-directing variables and more resource-dispersing variables. This differentiation also makes sense considering the construct of language proficiency that informs the ACTFL proficiency guidelines. As the ACTFL guidelines indicate, Intermediate learners “write primarily in present tense,” while Advanced learners “can narrate and describe in the major time frames of past, present, and future” (ACTFL, 2012). Crucially, both types of tasks covered topics that related to students’ personal lives and immediate experience.

Table 3. *Dataset by Type of Writing*

	Narration		Description	
Proficiency Level	**Texts**	**Words**	**Texts**	**Words**
NH	395	26,302	382	19,009
IL	770	66,837	754	42,773
IM-Low	450	40,183	447	30,323
IM-Mid	490	45,224	485	36,958
IM-High	2537	387,368	2520	254,044
IH	1230	238,211	1233	146,620
A	868	214,722	871	124,224
Total	6740	1,018,847	6692	653,951

Lexical Complexity Measures

Lexical diversity

Because our texts vary in length, we have adopted the VOCD method of calculating lexical diversity (Malvern, et al., 2004). The CLAN software calculates VOCD by comparing the actual TTR against tokens of a curve sample (based on randomly chosen words throughout the text) and uses a curve-fitting procedure to find the best fit. The best-fit value (reported as the VOCD value) is the index of lexical diversity for that text, with higher values representing greater diversity.

VOCD has been shown to produce a measurement of lexical diversity that is stable across different text lengths, however it requires a minimum of 35 words in each text. Because approximately 25% of students’ responses to individual prompts fell below this threshold, for this analysis, we combined each student’s prompt responses into a single text. The CLAN software was then used to generate a VOCD value for each learner. A mean of these values was then generated by proficiency level. Because texts could not be kept separate by prompt for this analysis, the distinction between narrative writing and descriptive writing was lost, and we were unable to examine lexical diversity as it relates to task type.

Lexical Density

Lexical density (LD) measures the proportion of lexical words (nouns, verbs, adjectives, and adverbs) in a text. In order to calculate lexical density values for each text in our corpus, a Perl program was used to count the frequency of each lexical part of speech for each text in the POS tagged files. These values were used to investigate differences in specific lexical parts of speech across level and task, and were summed and divided by the total number of tokens in the text to generate a single lexical density score for each text. Mean scores were then calculated by proficiency level and task type (narrative, descriptive).

Lexical Sophistication

In calculating lexical sophistication, we define advanced words based on frequency lists generated from the Corpus del Español, which were formulated by Davies (2002) based on a 20-million-word corpus which is balanced by genre and accurately tagged. The analysis was run using both the 2,000 most frequently occurring lemmas and the 2,000 most frequently occurring word forms to identify non-advanced words in our corpus. Because results for these two analyses were parallel, only word form results are presented and discussed here.

Simple proportions of advanced words to non-advanced words have been shown to be highly unstable when texts vary greatly in length. In order to obtain a standardized measure that is stable across texts of varying lengths, we calculated Guiraud Advanced (GA) (Daller et al., 2003). The formula for GA is the number of advanced types divided by the square root of the number of tokens.

In order to calculate GA, a Perl program was used to process the CLAN files, matching words and their parts-of-speech (to distinguish homographs) to the high frequency word list. Results were then used to calculate GA for each text using the formula above. Mean GAs by proficiency level and task type were then calculated.

In addition to a summary GA score, sophistication scores for individual lexical parts of speech were calculated. These scores were simple proportions of, for example, number of advanced nouns (defined as nouns not on the 2k most frequent word forms list) divided by total number of nouns. This calculation was performed for all lexical word classes (nouns, verbs, adjectives, and adverbs) for each text, enabling mean scores to be determined by level and task type. We chose to focus on lexical parts of speech because they are open classes and offer learners the most constituent words to choose from. This decision had the added advantage of providing complementary part of speech information for our analyses of density and sophistication.

Statistical Analyses

Mean scores and standard deviations were computed for each of the three measures at each proficiency level. Because we are interested in examining changes in each lexical measure from each proficiency level to the next, a series of one-way ANOVAs was run after assumptions were checked for all three measures. Levene's test showed that all measures met the homogeneity of variances assumption. Due to the large sample size, Kolmogorov-Smirnov tests were used to test for normality and revealed that none of our measures were normally distributed, however as ANOVA testing is robust to violations of normality with large sample sizes, it was still used.

In our analyses of differences between task types, both for individual parts of speech and for summary measures, a series of t-tests were run. Again, assumptions were checked for all data sets. Levene's test showed that all measures met the homogeneity of variances assumption. A series of Shapiro-Wilk tests of normality revealed that data was not normally distributed, however, as with ANOVA, t-tests are robust to violations of this assumption when sample sizes are large.

Results

Lexical Complexity Development Across Proficiency Levels

The first goal of this study was to determine whether the three components of lexical complexity differed systematically by proficiency level. Table 4 presents descriptive statistics for lexical diversity (VOCD, by student), lexical density (LD, proportion of lexical words, by text), and lexical sophistication (GA, by text) by proficiency level.

As indicated in previous sections, our overall hypothesis was that VOCD and GA would increase with learner's writing proficiency score. In fact, all three measures increased, with a notable exception at the IM-Mid level, where VOCD and LD decreased, and GA flattened before all three measures rose again at the IM-High level. This seeming anomaly at IM-Mid, along with patterns in the general rate of change between levels, is discussed in the following sections.

Table 4. *Lexical Complexity by Proficiency Level*

Proficiency level	**VOCD**		**LD**		**GA**	
	M	***SD***	***M***	***SD***	***M***	***SD***
NH	60.88	22.13	.52	.17	.26	.10
IL	65.27	17.68	.55	.15	.31	.11
IM-Low	68.17	18.60	.57	.13	.33	.11
IM-Mid	65.08	16.78	.56	.15	.33	.11
IM-High	77.15	16.44	.58	.11	.42	.13
IH	86.44	15.89	.59	.09	.49	.14
A	92.72	14.31	.59	.07	.66	.19

Lexical Diversity

Lexical diversity, as measured by VOCD, increased as proficiency level increased except at proficiency level IM-Mid where it fell slightly (Figure 1).

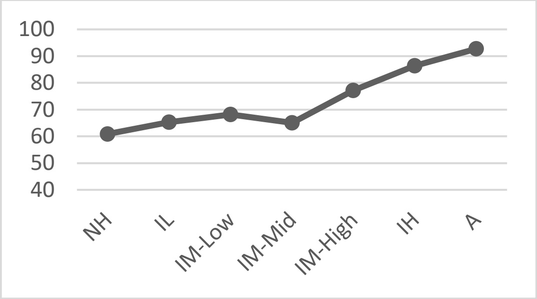

Figure 1. Mean VOCD by proficiency score.

There was a statistically significant difference in VOCD score between groups according to a one-way ANOVA (F(6,3380) = 172.20, *p* = .000, η^2 = .24). Despite this, Figure 1 illustrates that learners' lexical diversity remained relatively level between the NH and IM-Mid levels and then increased in larger increments from IM-Mid to A. A Tukey post hoc test bears out this pattern, revealing no significant differences between IL and IM-Low (*p* = .61), IL and IM-Mid (*p* = .99), and IM-Mid and IH (*p* = .67). Differences between all other groups were significant.

Lexical Density

As with VOCD, lexical density, as measured by the proportion of lexical words to tokens, increased with proficiency score except at the IM-Midlevel (Figure 2).

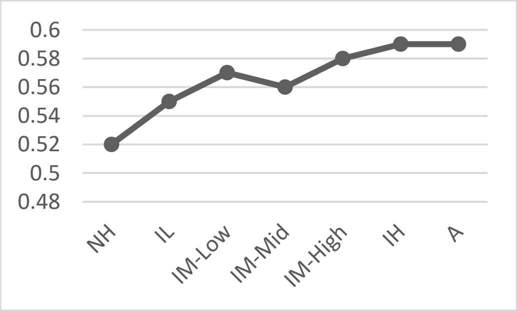

Figure 2. Mean LD by proficiency score.

Results of a one-way ANOVA revealed a statistically significant difference between groups (F(6,20438) = 93.79, p = .000, η^2 = .03). Although lexical diversity (VOCD) and lexical density both increased with proficiency level and revealed the same dip in values at the IM-Mid level, they otherwise exhibited approximately opposite patterns in terms of when the largest gains occurred. As can be seen in Figure 2, learners' lexical density increased in larger increments between NH and IM-Low, fell at IM-Mid, and then rose more slowly, eventually levelling off at the IH and A levels. A Tukey post hoc test confirms this analysis, showing a nonsignificant difference between groups IL and IM-Mid (p = .20) and groups IM-Low and IM-Mid (p = .34), as well as between groups IH and A (p = .73).

It is important to note, however, that although ANOVA results for LD were significant, the overall variation in LD spans a relatively small range and the effect size is small. An analysis of the proportion of each lexical part of speech to total tokens (Figure 3) illustrates that the larger differences between NH and IM-Low texts are influenced by an increase in the proportion of noun use. The dip in LD at IM-Mid is attributable to a slight decrease in the use of verbs, adjectives, and adverbs, and a larger decrease in the use of nouns. However, the overall trend for all parts of speech displays a small, steady, and relatively unexciting rise in proportion across levels.

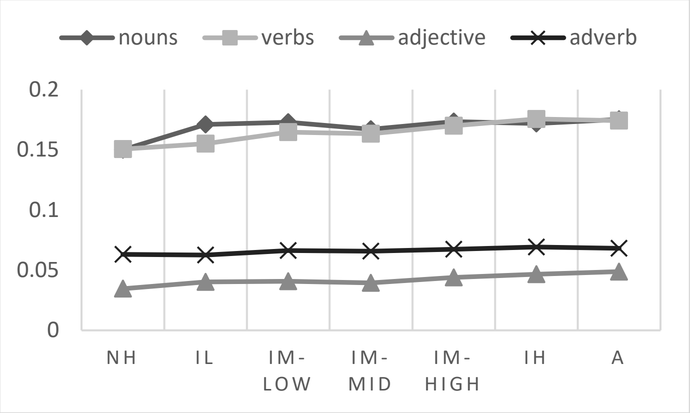

Figure 3. Mean proportion of lexical items by part of speech and level.

Lexical Sophistication

Lexical sophistication, as measured by GA, also increased with proficiency rating across all levels except between IM-Low and IM-Mid (Figure 4). Unlike VOCD and LD, GA did not fall at the IM-Mid level but rather remained stable, with texts produced by learners at both IM-Low and IM-Mid levels displaying a mean GA of 0.33.

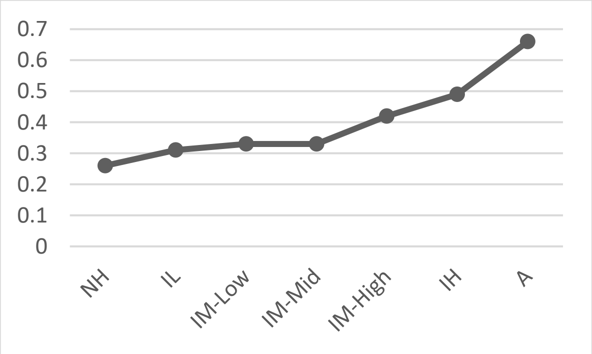

Figure 4. Mean GA by Level.

A one-way ANOVA found that differences in GA between levels were statistically significant (F(6, 20438)

= 1572.26, p = .000, η^2 = .32). A Tukey post hoc analysis revealed significant differences between all groups except for IM-Low and IM-Mid (p = .57). Mean GA scores followed the same general pattern as VOCD scores, exhibiting relatively small increases between Novice High and Intermediate Mid learners, and greater increases between Intermediate Mid and Advanced learners. The largest increase in mean score occurred between Intermediate High and Advanced students.

Sophistication scores for lexical parts of speech (Figure 5) revealed that the proportion of advanced nouns to all nouns rose and fell slightly between NH and IM-Mid before leveling off at higher proficiency levels. Verbs, adjectives, and adverbs all followed the same pattern as overall GA score, with modest increases from NH to IM-Mid and larger increases from IM-Mid to Advanced. The proportion of advanced verbs to all verbs exhibited the greatest change, beginning to rise sharply at the IM-Low level. While this change is likely partially attributable to an increase in the number of different advanced verbs used, it is important to note here that this data was generated using the 2,000 most frequent word forms for Spanish, rather than lemmas. The verbs labelled "advanced" will have also contained less common forms (e.g., conjugations) of common verbs, and so the increases in frequency of advanced verbs here likely represents a combination of both increased lexis and an increasing repertoire of grammatical structures such as tenses.

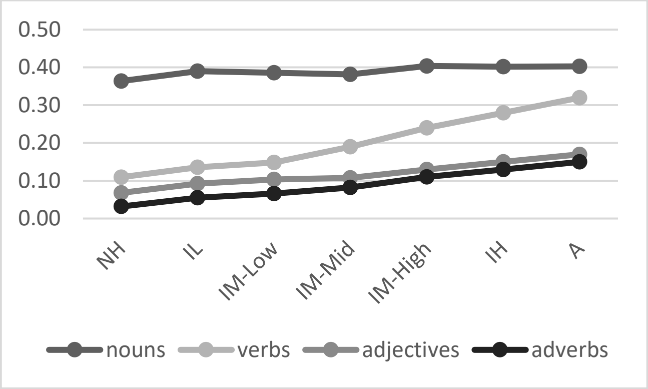

Figure 5. Proportion of advanced nouns/all nouns, etc.

Lexical Complexity and Task Type

A second goal of this study was to determine if measures of lexical complexity differ according to task type (narrative or descriptive). Because VOCD was calculated based on the entirety of each student's written output (six tasks combined), results for individual task types could not be calculated. Therefore, our analysis of task effect on lexical complexity is based on only two measures: density and sophistication.

Lexical Density

Table 5 presents the means and standard deviations of lexical density values by level and task. Proficiency levels with a significant difference in LD between narrative and descriptive writing are indicated with an asterisk.

Descriptive writing produced higher LD values at every level, indicating that learners were using a higher rate of lexical words to non-lexical words in descriptive than in narrative writing.

In order to determine if differences in LD between task types were significant at each level, and in total, a series of independent t-tests was performed with an adjusted significance level of p < .01. Results indicated that descriptive writing had significantly higher LD than narrative writing for all texts combined (t(13629) = -31.37, p = .000). In comparisons between task type by level, significant differences were found at the IM-Low level (t(910) = -7.18, p = .000) and at the IM-Mid level (t(998) = -11.47, p = .000). Results at all other levels were nonsignificant.

Figure 6 shows the density of each lexical part of speech by task type. Nouns and verbs both represent a higher proportion of descriptive texts than narrative, while adjectives and adverbs comprise a higher proportion of narrative texts.

Table 5. *Lexical Density by Level and Task*

Proficiency level	Narrative		Descriptive	
	M	***SD***	***M***	***SD***
NH	0.48	0.18	0.56	0.15
IL	0.51	0.15	0.58	0.12
IM-Low*	0.53	0.16	0.59	0.11
IM-Mid*	0.50	0.15	0.60	0.10
IM-High	0.55	0.11	0.60	0.09
IH	0.56	0.09	0.61	0.08
A	0.57	0.06	0.61	0.07
Total	0.53	0.12	0.59	0.10

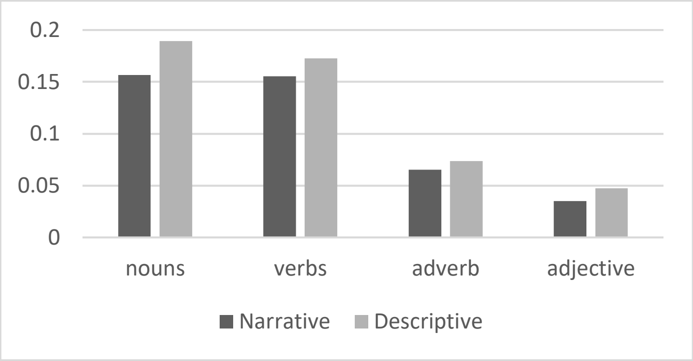

Figure 6. Density of each lexical POS by task type (proportion of POS to all tokens).

Results of a series of independent t-tests indicated that all differences were statistically significant (Table 6).

Table 6. *Results of t-tests on POS by Task Type*

Task types	Narrative		Descriptive		*df*	*t*	*p*
	M	***SD***	***M***	***SD***			
Nouns	0.16	0.05	0.19	0.05	13629	-36.09	0.000
Verbs	0.16	0.06	0.17	0.06	13629	-17.01	0.000
Adjectives	0.05	0.03	0.03	0.02	13629	24.94	0.000
Adverbs	0.07	0.04	0.05	0.04	13629	12.63	0.000

Overall, the analysis of lexical density shows that the two types of tasks elicited different degrees of density, but also different density profiles when POS is taken into account.

Lexical Sophistication

Table 7 presents the means and standard deviations for GA values by level and task type and indicates that narrative tasks elicited higher lexical sophistication at all levels except for IM-Mid. Proficiency levels with a significant difference in GA between narrative and descriptive writing are indicated with an asterisk.

Table 7. *GA Values by Level and Task*

Proficiency Level	**Narrative**		**Descriptive**	
	M	***SD***	***M***	***SD***
NH	0.28	0.11	0.26	0.10
IL*	0.32	0.12	0.30	0.10
IM-Low	0.33	0.12	0.32	0.10
IM-Mid	0.33	0.12	0.34	0.10
IM-High*	0.45	0.13	0.39	0.11
IH*	0.56	0.14	0.45	0.13
A*	0.66	0.14	0.53	0.14

A series of t-tests was performed to determine if differences in GA by task type were significant at each proficiency level (again, with an adjusted significance level of $p < .01$). Results indicated that the difference in GA was significantly different for the two task types at the IL ($t(1561) = 3.56$, $p = .000$), IM-High ($t(5118) = 17.33$, $p = .000$), IH ($t(2486) = 18.06$, $p = .000$), and A ($t(1742) = 20.09$, $p = .000$) proficiency levels. Results at all other levels were nonsignificant.

Figure 7 shows the sophistication score for each lexical part of speech (proportion of advanced part of speech to all tokens of that part of speech). Results indicate that advanced nouns occurred more frequently in descriptive texts, while all other lexical parts of speech were more frequent in narrative texts. Differences between task types for all POS were statistically significant (Table 8).

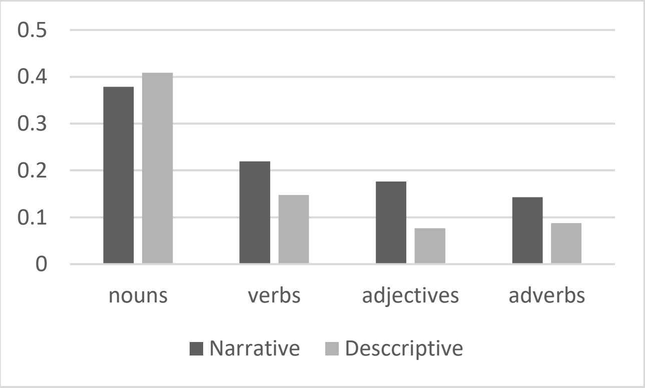

Figure 7. Mean proportion of each POS that is advanced by task.

Table 8. *Results of t-test on GA POS Sophistication Measures*

Task Types	**Narrative**		**Expository**		***df***	***t***	***p***
	M	***SD***	***M***	***SD***			
Nouns	0.38	0.15	0.41	0.16	13629	-10.21	0.000
Verbs	0.22	0.14	0.15	0.10	13629	35.06	0.000
Adjectives	0.18	0.16	0.08	0.06	13629	40.78	0.000
Adverbs	0.14	0.11	0.09	0.07	13629	21.63	0.000

Discussion

This study had three purposes: To investigate variation in lexical complexity in the writing of Spanish K-12 DLI students across proficiency levels (RQ1), to ascertain to what extent the writing task affects lexical complexity (RQ2), and to determine whether considering complexity data for lexical word classes elaborates on our understanding of lexical complexity (RQ3).

Lexical Complexity Across Proficiency Levels

Regarding our first research question, our results indicate an overall increase in the levels of lexical diversity, density and sophistication that parallels students' progression through proficiency levels. The growth is registered across major levels (Novice to Intermediate to Advanced) and also between sublevels within a major level (e.g., IM-Mid to IM-High to IH). Lexical diversity and lexical sophistication demonstrated the same basic pattern, with smaller increases across the NH to IM-Mid range, and larger increases from IM-Mid to Advanced. However, the transition from IM-Low to IM-Mid clearly and consistently defies this upward trend in that all three indicators of complexity decrease or remain stable. There are several possible explanations for this anomaly. Although further analysis of the data set would be needed in order to rule out competing explanations, we want to contemplate two possibilities.

Lexical complexity represents only one part of the suite of factors that comprise a learner's proficiency level and impact proficiency scores. Although the effect sizes reported in this study for measures of variation (.24) and sophistication (.32) were high, indicating that lexical complexity accounts for a substantial portion of variation between proficiency levels, we have not yet investigated the role of syntactic complexity, fluency, or accuracy in this corpus, nor did we evaluate appropriacy and content of response. One potential explanation for the increase in ACTFL proficiency score between IM-Low and IM-High in the absence of increased lexical complexity is that students at that level are making substantial gains in other areas.

Alternatively, the nature of the ACTFL proficiency scale may help explain the dip in lexical complexity from IM-Low to IM-Mid The ACTFL scale was originally developed with four major levels (Novice, Intermediate, Advanced, and Superior). When the AAPPL test was developed, ACTFL created a specific rating scale in which the lower levels were divided into four sublevels for Novice and five for Intermediate in order to capture and report the smaller, incremental, and more fine-grained changes that are likely to happen at these levels, but are not registered by the regular ACTFL proficiency scale. It is conceivable that the IM-Mid anomaly that surfaced in our study is the result of trying to convert the original scale into one that is too granular to register actual differences. The upper two sublevels are assigned to those Intermediate learners that can perform sometimes (IM-High) or often (IH) at the next major level (Advanced), while the lower two sublevels are reserved for those learners that exhibit basic performance (IL) or strong performance (IM-Low) at Intermediate with no evidence of the ability to perform at the next level (Advanced). This would leave the IM-Mid sublevel in a sort of "no-man's-land" position that is too indeterminate for raters to pinpoint accurately and consistently. Additionally, because each response is rated separately by a human rater and an algorithm uses those ratings to determine a global proficiency score, it is difficult to know how factors such as inconsistency across responses might impact a student's rating. From the perspective of our corpus, however, and calculations based on mean score by text, it is easy to imagine that prompts that were barely or inadequately answered might skew data, even if the same student responded meticulously and well to other prompts before facing time and/or fluency constraints.

Task Effect on Lexical Density and Sophistication

Our second research question explored the relationship between writing task and lexical density and sophistication of language produced. The two types of tasks differed in terms of the number of resource-directing and resource-dispersing variables that they involved. We considered narrative tasks to be more complex as they involved a larger number of both types of variables. However, compared to the tasks typically used in previous studies, ours were relatively low in both procedural and conceptual complexity.

The data analysis revealed differences between the two task types. Descriptive tasks resulted in higher levels of lexical density overall than more complex narrative tasks. This finding could be attributed to the predictions of Skehan's (2014) Limited Attentional Capacity Model, where more complex tasks tax learners' attentional resources causing the lexical complexity of their output to decrease. However, we hypothesize that the lexical density results can be explained as an effect of genre rather than of task complexity. Previous research indicates that narrative discourses produce more third person personal pronouns, imperfect and preterit tense verbs, and possessives, all of which may contribute to sophistication, but do not increase density.

The results of the lexical sophistication analysis paint a different picture. Here, the more complex narrative tasks elicited consistently higher degrees of sophistication; a result that cannot be explained under the assumptions of the Limited Attentional Capacity Model. The difference in sophistication between tasks is particularly significant at higher levels of proficiency. Crucially, and in contrast to the findings for lexical density where growth between levels happens in small and consistent increments, gains in sophistication accelerate once learners pass the IM-Mid level. It is likely that students rated below IM-High did not have the linguistic resources needed to produce the more sophisticated language required by the narrative task; they struggled with a task that required them to describe and narrate using different tenses. At these lower levels of proficiency, the beneficial effects of conceptually more complex tasks are not realized precisely because learners do not have the linguistic resources that would be tapped into through those resource-directing variables. Learners at higher levels of proficiency have the necessary resources to produce more complex language, and the right task can push them to do so. As we pointed out earlier, the analysis of sophistication was done at the word (not the lemma) level. Consequently, the ability to produce text in present, past and future tense will elicit multiple forms of the same verbs, some of which are likely to fall outside the 2,000 most common words in Spanish. Only learners that have the morphosyntactic resources to produce past and future narrations will be able to produce those more sophisticated verb forms. The obvious conclusion from this interpretation of the results is that the impact of task effects on lexical sophistication is, to some extent, relative to the level of proficiency of the learner. This finding could still be explained from the perspective of the Cognition Hypothesis, but it would represent an anomaly for Skehan's Limited Attentional Capacity Model since the less complex tasks consistently generated less sophisticated vocabulary.

Lexical Word Classes, Proficiency, and Task Effect

Our third research question aimed to elaborate on the results found by the previous two sets of analyses (variation by level and variation by task) by investigating differences in the use of lexical words by part of speech. By level, results for LD revealed little change in the proportion of any lexical word class. In combination with ANOVA results for total LD, which were statistically significant but had a small effect size and a small range of variation, these results seem to indicate that lexical density is the least informative of our three lexical complexity measures. This may be, in part, because lexical density is a better discriminator between registers (academic versus non-academic language) and modes (speech versus writing) than between proficiency levels. This is supported by the fact that the density of lexical parts of speech did exhibit significant differences by task type, with more complex tasks eliciting a higher proportion of adjectives and adverbs. These results do indicate that register and/or genre impact the expression of LD, especially considering that all tasks responded to in our corpus cover relatively "everyday" topics.

Part of speech data for sophistication was more illuminating. While the proportion of advanced nouns remained relatively stable across proficiency levels, the proportion of advanced verbs rose dramatically beginning at IM-Low and continuing through Advanced. Adjectives and adverbs both also rose very modestly across lower proficiency levels, with a sharper rise after the IM-Mid level. This analysis indicates that an increased use of advanced verbs characterizes higher proficiency levels. Again, this increase is likely to be due both to an expanding repertoire of verbs and to an expanding mastery of grammatical forms such as tense. The part of speech sophistication data for task indicated that easier tasks elicit more nouns than more complex tasks, and that advanced verbs, along with smaller proportions of adjectives and adverbs, are required for more complex writing.

Taken together, our results indicate that development across lower levels of writing proficiency is characterized by modest increases in lexical complexity with respect to diversity and sophistication. Students at these levels are largely relying on more repetition and common words, with only small changes between proficiency levels. This indicates that differences in proficiency as students move from NH to IM are likely characterized by development in other areas, such as syntax, accuracy, and fluency. After a period of little change (or even loss) in lexical complexity at the IM level, lexical diversity and sophistication both rise dramatically as students move from Intermediate to Advanced. This indicates that both breadth and depth of lexical knowledge are a distinguishing factor between the writing of students at these higher proficiency levels.

Implications

The results of this study clearly point to the importance of attention to lexical development, particularly in terms of diversity and sophistication, as both indicators show a strong correlation with writing proficiency levels. Although explicit attention to vocabulary development in the second/foreign language classroom has received some level of increased attention over the past two decades (e.g., Hinkel 2002; Nation, 2005), it is not clear that this emphasis has also spread to teaching practices overall and in particular, to the immersion context that is the focus of this study. Unlike other models that integrate the teaching of language and content, the DLI model in the United States is particularly (often exclusively) present at the elementary and middle school levels. At those levels, teachers place a strong emphasis on teaching the content on which students will be tested in standardized assessments (e.g., math and science) and teachers’ subject-matter expertise is often stronger than their knowledge of second language acquisition or pedagogy. The dual demands of attention to language and content in the DLI classroom often resolve in favor of content. One of the consequences that this focus on content has for the acquisition of vocabulary is the assumption that learners will acquire most of their vocabulary incidentally through reading. However, as research has demonstrated, unless incidental acquisition can be complemented with more explicit, intentional approaches, learners are unlikely to develop the kind of lexical depth and breadth required to reach advanced levels of proficiency (see Hulstinjn, 2003, for a review of research). The fact that lexical sophistication showed the largest effect size, that is, connection with writing proficiency, lends even more support to this call for explicit attention to vocabulary development. Our data show that, while lexical sophistication does not exhibit strong gains between Novice and the lower ranges of the Intermediate level, the ability to use lower frequency vocabulary is key to writing at higher levels of proficiency. This is an area where the immersion model arguably provides an advantage to language learners. A context in which language is taught through academic content is likely to incorporate a wider variety of vocabulary and more instances of lower frequency words than the traditional foreign language classroom, which would explain the relatively high levels of lexical sophistication that the students in our study exhibited.

The analysis of task effects also results in important implications for classroom teachers, specifically the differing effects that task complexity and genre have on the written output. The data show that it is genre, rather than task complexity that affects lexical density. Teachers need to keep in mind that writing tasks that require what would a priori be considered an “easier” genre, such as description, may still be beneficial to students. Additionally, the effects of task complexity on lexical sophistication that are discussed above indicate that more complex tasks are better suited to elicit the kind of complex vocabulary that is a trademark of the Advanced level. Learners need to be given tasks that include what Robinson (2011) terms resource-directing variables so that they are pushed to produce vocabulary beyond what is typical in most routine communicative situations.

In addition to the immediate applications for classroom teaching discussed above, this study has wider implications for our field. With increasingly easier access to very large sets of learner language data provided by corpora such as CUDLI, researchers are able to investigate learning processes at a much deeper level than what is normally possible in the classroom or through small scale testing. As Godwin-Jones

(2017) points out, in language teaching and learning, just like in any other field, "the availability of large data sets can lead to evidence-based questioning of accepted theories and practices" (p. 11). But the advantages of this big data approach to second language acquisition will only be possible if sources of big data for language research are free and easy to access and if teachers and researchers are trained to use them.

Conclusion

This study integrated an examination of lexical complexity across proficiency with an investigation of task effects to better interpret the relationship between lexical characteristics and proficiency. By using a large corpus of writing produced under highly standardized conditions and containing varying tasks, we were able to identify patterns in both the development of lexical complexity and the effect of task. Our results indicate that a simple view of lexical complexity where diversity and sophistication are expected to rise as proficiency increases may be insufficient to interpret lexical development. In particular, task characteristics play a role in determining the lexical features of writing. Applications of lexical analyses that do not account for task run the risk of evaluating only a partial picture of learner proficiency. As lexical complexity measures are increasingly used to inform language pedagogy and assessment, considerations of the effect of task on learner output can offer additional and critical data to researchers and educators.

References

ACTFL. (n.d.). *Assigning CEFR ratings to ACTFL assessments*. https://www.actfl.org/sites/default/files/reports/Assigning_CEFR_Ratings_To_ACTFL_Assessments.pdf

ACTFL. (2012). *ACTFL proficiency guidelines—writing*. http://www.actfl/org

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*(1), 28–41.

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

Biber, D., Davies, M., Jones, J. K., & Tracy-Ventura, N. (2006). Spoken and written register variation in Spanish: A multi-dimensional analysis. *Corpora*, *1*(1), 1–37.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 23–46). John Benjamins.

Crossley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, *18*(2), 119–135.

Crossley, S., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, *17*(2), 171–192.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*(2), 243–263.

Crossley, S. A., & Skalicky, S. (2019). Examining lexical development in second language learners: An approximate replication of Salsbury, Crossley & McNamara (2011). *Language teaching*, *52*(3), 385–405.

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied linguistics*, *24*(2), 197–222.

Davies, M. (2002). *Corpus del Español: 100 million words, 1200s-1900s*. Retrieved from https://www.corpusdelespanol.org/

Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics, 25*(2), 220–242.

Durrant, P., & Brenchley, M. (2019). Development of vocabulary sophistication across genres in English children's writing. *Reading and Writing, 32*(8), 1927–1953.

Frear, M. W., & Bitchener, J. (2015). The effects of cognitive task complexity on writing complexity. *Journal of Second Language Writing, 30*, 45–57.

Godwin-Jones, R. (2017). Scaling up and zooming in: Big data and personalization in language learning. *Language Learning and Technology, 21*(1), 4–15. Retrieved from http://llt.msu.edu/issues/february2017/emerging.pdf

Halliday, M. (1989). *Spoken and written language*. Oxford University Press, USA.

Hulstinjn, J. (2003). Incidental and intentional learning. In C. J. Doughty & M. H. Long (Eds), *The Handbook of Second Language Acquisition* (pp. 349–381). Blackwell.

Ishikawa, T. (2007). The effect of manipulating task complexity along the [+/-Here-and-Now] dimension on L2 written narrative discourse. In M. del P. Garcia Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 136–156). Multilingual Matters.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language learning, 51*(3), 401–436.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19*(1), 57–84.

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing, 20*, 148–161, doi: 10.1016/j.jslw.2011.02.001.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, *49*(4), 757–786.

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior research methods*, *50*(3), 1030–1046.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics*, *16*(3), 307–322.

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

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development.* Palgrave Macmillan.

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing, 24*(4), 459–488.

Polio, C., & Park, J.-H. (2016). Language development in second language writing. In R. Manchón & P. K. Matsuda (Eds.), *Handbook of second and foreign language writing* (pp. 287–306). Mouton de Gruyter.

Rahimi, M. & Zhang, L. J. (2018). Effects of Task Complexity and Planning Conditions on L2 Argumentative Writing Production. *Discourse processes, 55*(8), 726–742, DOI: 10.1080/0163853X.2017.1336042

Rahimpour, M. (1999). Task complexity and variation in interlanguage. In N. Jungheim & P. Robinson (Eds.), *Pragmatics and Pedagogy: Proceedings of the 3rd Pacific Second Language Research Forum, Vol 2* (pp.115–134). PacSLRF.

Read, J. (2000). *Assessing Vocabulary*. Cambridge University Press.

Robinson, P. (2007). Criteria for classifying and sequencing pedagogic tasks. In M. del P. Garcia Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 7–26). Multilingual Matters.

Robinson, P. (2011). Second language task complexity, the Cognition Hypothesis, language learning, and performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (pp. 3–37). John Benjamins.

Rubio, F. & Schnur, E. (2019-). *The Corpus of Utah Dual Language Immersion (CUDLI)*.

Skehan, P. (2003). Task-based instruction. *Language teaching, 36*, 1–14.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied linguistics, 30*(4), 510–532.

Skehan, P. (2014). Limited attentional capacity, second language performance, and task-based pedagogy. In P. Skehan (Ed.), *Processing perspectives on task performance (task-based language teaching)* (pp. 211–260). John Benjamins.

Staples, S. & Reppen, R. (2016). Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. *Journal of Second Language Writing 32*, 17–35.

Stubbs, M. (1986). Lexical density: A technique and some findings. In M. Coulthard (Ed.) *Talking about text* (pp. 27–42). English Language Research.

Ure, J. (1971). Lexical density and register differentiation. In G. Perren & J. L. M. Trim (Eds.), *Applications of Linguistics* (pp. 443–452). Cambridge 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*(ii), 171–184.

West, M. (1953). *A general service list of English words*. Longman, Green and Co.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied linguistics, 31*(2), 236–259.

About the Authors

Erin Schnur is a curriculum developer at Cambly. Her research interests include corpus linguistics methodology and pedagogical applications of corpora.

E-mail: erin@cambly.com

Fernando Rubio is Professor of Spanish Linguistics at the University of Utah, where he is also Director of the Second Language Teaching and Research Center. His research interests are in the areas of Applied Linguistics and Teaching Methodologies including technology-enhanced language learning and teaching, and language proficiency assessment.

E-mail: fernando.rubio@utah.edu