Spring 2023

# Exploring the Lexical Profile of Advanced L2 Writers: Longitudinal Data From The Russian Overseas Flagship Program

William J. Comer
*Portland State University*, wcomer@pdx.edu



## Citation Details

Published as: Comer, W. J. Exploring the lexical profile of advanced L2 writers: Longitudinal data from the Russian Overseas Flagship program. The Modern Language Journal.

Accepted for Modern Language Journal, probably publication in Spring 2023 issue.

# Exploring the Lexical Profile of Advanced L2 Writers: Longitudinal Data From The Russian Overseas Flagship Program

WILLIAM J. COMER
ORCID ID: 0000-0002-8168-2810

*Portland State University, Department of World Languages and Literatures, 1855 SW Broadway Portland, OR 97201 E-mail: wcomer@pdx.edu*

<ABSTRACT>

This study explores the lexical profile of essays written by 48 advanced learners of second language (L2) Russian who participated in the Russian Overseas Flagship, an intensive year-long study abroad program, designed to help students reach Interagency Language Roundtable (ILR) Level 3 proficiency in all skills. Using the lexical frequency profile (LFP) and P–Lex as measures of vocabulary sophistication, the study found that over the 9 months of the program, students significantly increased their usage of words from the lowest frequency bands. This adds to the findings of Hacking and Tschirner (2017) that knowledge of lexical items at the 3,000–5,000 frequency levels predicts reading proficiency at the ACTFL Advanced High-Superior level in Russian. However, the increase of vocabulary sophistication was not clearly correlated with improvements in the students' writing proficiency scores, as measured on the ILR scale. A qualitative analysis of the students' low frequency vocabulary usage reveals their control of native Russian vocabulary and derivational morphology. The analyses reveal the effects of writing tasks on student vocabulary usage.

<END ABSTRACT>

*Keywords*: lexical sophistication; lexical frequency profile; L2 writing; Russian; ILR Level 3; derivational morphology

Since 2006 The Language Flagship, an initiative of the National Security Education Program, has supported the development of innovative language programs that encourage undergraduates to follow the major of their choice while studying a critical language (Murphy & Evans–Romaine, 2017). Participants are expected to reach Limited Working proficiency, Level 2 as measured on the Interagency Language Roundtable (ILR) scale (Interagency Language Roundtable, 2021) in their critical language through study in U.S. domestic language programs and then to reach General Professional proficiency (ILR Level 3) by the conclusion of a capstone year abroad. The students' proficiency is measured at multiple points using certified ACTFL OPI tests for speaking and ILR-calibrated proficiency tests developed by American Councils for reading, listening and writing. Davidson (2010) reported on the proficiency outcomes for speaking, reading, and listening from the early years of the Russian Flagship Program, and his data show that student participants regularly reach ILR Level 3 in those three modalities.

Starting in 2015, student participants in the Russian Overseas Flagship completed writing proficiency tests as part of the pre-program and post-program testing battery. This study will describe the growth of lexical sophistication (i.e., the number of low-frequency words) that Flagship students demonstrate in their written essays, and explore the relationship between their lexical sophistication and writing proficiency scores. This is the first study to explore the lexical component in the writing of English learners of Russian, and even among the research studies of lexicon in L2 production, this study is unique for its longitude (9 months elapse from pretest to posttest), and for the high level of language proficiency of the participants since ILR Levels 2 and 3 correspond roughly to ACTFL Advanced and Superior, and CEFR levels B1–C1 (Tschirner, 2011). Because of the highly selective nature of the Russian Flagship Program, only a dozen or so students participate in the Overseas program each year. Therefore, to ensure a

reasonable sample size, a corpus of essays has been aggregated from four cohorts of students from academic years 2016–2017 through 2019–2020, for a total of 48 participants. The essential elements of the academic curriculum of the Overseas program and the structure of the study abroad experience (i.e., home stays, internship with local business or non-profit organization, extracurricular activities, pledge to use only Russian) were consistent across the four cohorts.

Studies of L2 lexical development generally explore either individual lexical items (Horst & Collins, 2006; Lindqvist et al., 2011) or multiword units (Forsberg–Lundell & Lindqvist, 2012; Laufer & Waldman, 2011; Siyanova–Chanturia & Spina, 2020). In their conceptual approaches some studies examine learners' lexicon growth as a facet of evolving linguistic complexity (Vyatkina & Housen, 2021), while others tend to focus on statistical modeling and automated tools for measuring learners' lexicon in relationship to their overall proficiency (Bestgen, 2017; Crossley & Kyle, 2018; Crossley et al., 2011; Kyle & Crossley, 2015).

This study takes a mixed-methods and frequency-based approach to the students' lexical sophistication in these writing samples based on individual lexical items. The study uses a current frequency list of lemmas derived from the Russian National Corpus (Liashevskaia & Sharov, 2009) and generates two well-established vocabulary measures, the Lexical Frequency Profile (LFP; Laufer & Nation, 1995) and P–Lex (Kyle, 2020; Meara & Miralpeix, 2017). This study is important in documenting a baseline of the lexical development of very advanced L2 learners (most with English as an L1, and all completing higher education in English in U.S. institutions) writing in a language with a high level of complexity in both inflectional and derivational morphology. While learners' use of collocations and other multiword units in this corpus should be studied, it seems prudent first to explore the vocabulary growth of L2 learners of Russian at the level of lexical items. The study will rely primarily on word frequency

measures because there are no accepted psycholinguistic indices (e.g., measures of word features, such as concreteness, distinctiveness, etc.) for Russian, despite the utility of such indices in documenting the multidimensional aspects of lexical sophistication (Eguchi & Kyle, 2020; Kim et al., 2018). To make up for the limitations of a frequency-based analysis, the study is complemented with an extensive qualitative analysis of the change in learners' usage of low-frequency words.

## <A>LEXICAL FREQUENCY PROFILE

Laufer and Nation (1995) first advanced the idea of measuring the size of language learners' vocabulary through the Lexical Frequency Profile (LFP), in which the words used in learners' writing were measured against established frequency lists of general and academic vocabulary. Their LFP measure produced a set of four percentage scores indicating how much the learners' vocabulary matches the first thousand (hereafter 1K) and second thousand (hereafter 2K) most frequent word families in English, how much it matches the academic vocabulary list, and how much is not found in any lists. The LFP views a learner's vocabulary size as a line with many contours, rather than a single numerical measure.

Horst and Collins (2006) applied the LFP to a longitudinal corpus of writing by young (age 11–12) beginner-level francophone learners of English. Despite 400 hours of instruction, the LFP of their learners' texts changed very little, although there was clear growth in the learners' vocabulary. They attributed some of this mismatch to the fact that the LFP is more accurate with longer texts (over 200 words) and to the age of the learners. In a qualitative review of the learners' writing samples, Horst and Collins noticed that the use of cognates seemed to distort the LFP measure. They noticed that in the earliest writing samples their learners often used

Latinate cognates in their texts (i.e., *respond*, a very close cognate to the French *répondre*) when more basic words of Germanic origin (i.e., *answer*) would be more typical. Since Latinate words in English often have a lower frequency than words of Germanic origin, the use of cognates created the appearance that the learners had a richer vocabulary than they actually did. Indeed, over the course of instruction, their learners used fewer of these Latinate words and used a wider range of words and word forms from the 1K band.

Lindqvist et al. (2011) examined the lexical richness of Swedish learners of L2 French and Italian using samples of oral language, and they adapted the LFP to frequency lists built on lemmas, rather than word families. They found that LFPs could distinguish their low advanced French learners from high advanced learners and native French speakers; for L2 Italian, the LFP could distinguish intermediate L2 learners from advanced learners and advanced learners from native speakers. LFP was not able to distinguish the lexical richness of high advanced French learners from native speakers. In a follow-up qualitative analysis of the vocabulary in the learners' samples, Bardel and Lindqvist (2011) found (like Horst and Collins) that cognates could cause some learners' vocabulary to appear richer than it actually was. They also noted that for instructed L2 learners the thematic organization of vocabulary in textbooks may also play a role in learners acquiring relatively low frequency words.

Vedder and Benigno (2016) conducted a similar cross-sectional study of lexical richness and collocation use in the writing of native speakers of Italian and Dutch L2 learners of Italian at CEFR Levels A2–B1. Although their LFP analysis did not find differences in lexical richness based on the L2 learners' proficiency levels, they did find different patterns of lexical richness between the native speakers and L2 learners. In comparison with native speakers, L2 learners

tended to overuse words in the 1K band and underuse words at lower frequencies (3K and beyond).

While recent studies of lexical richness have focused more on learners' use of collocations and other multiword units, either cross-sectionally (Bestgen, 2017; Crossley et al., 2011; Forsberg–Lundell & Lindqvist, 2012; Garner et al., 2019; Laufer & Waldman, 2011) or longitudinally (Siyanova–Chanturia & Spina, 2020), the LFP with its focus on individual lexical items continues to be used and a recent study of its concurrent validity found that the ratios from an LFP analysis have significant correlation with writing proficiency scores of university-level ESL students (Higginbotham & Reid, 2019).

Given the very advanced nature of the learners in this study and the length of time between pretest and posttest, it seems likely that as these learners progress to or reach ILR Level 3, their vocabulary will pattern more like that of the native speakers in previous studies (Lindqvist et al., 2011; Vedder & Benigno, 2016), decreasing their use of high-frequency words and increasing the number of low-frequency vocabulary items. Horst and Cobb (2006) and Lindqvist et al. (2011) show the importance of balancing an LFP analysis with a qualitative examination of the learners' vocabulary use.

The LFP (as "a line with many contours") is a very suitable tool for documenting the frequency ranges where the learners' lexical growth occurs since almost all the essays Flagship students produced on the writing test were over 200 words in length. The LFP was adapted for this study in two ways. First, the frequency dictionary produced by Liashevskaia and Sharov (2009) on the basis of the Russian National Corpus is organized by lemmas, not word families. Liashevskaia and Sharov offer frequency lists of lemmas by subcorpora (i.e., fiction, non-fiction, oral language), but because the learners in this study wrote to a variety of prompts including

personal narratives as well as argumentative essays, it seemed best to use the general frequency list of lemmas based on the whole corpus. Second, Laufer and Nation (1995) used a general list for the 1K and 2K bands of most frequent word families, supplementing it with an academic vocabulary list, and then counting the number of words not in any list. Talalakina et al. (2020) have developed an academic vocabulary list for Russian, but that list did not seem entirely appropriate for this study, since the texts that the learners produce in the testing sessions are not pieces of academic expository writing. So, for this study, learners' LFPs were generated from the general frequency list of lemmas, divided initially into six bands based on the first (1K), second (2K), third (3K), fourth (4K), fifth (5K) thousand and beyond the fifth (5K+) thousand. Extending the LFP to these very low frequency ranges is justified given the very advanced nature of the writers in this study, and it also allows the results of this study to be compared with those from others about vocabulary knowledge and advanced Russian language skills.

## <A>LEXICAL STUDIES AND L2 RUSSIAN

Hacking and Tschirner (2017) explored the relationship between receptive vocabulary knowledge and reading proficiency for 48 English-speaking learners of Russian. They found that receptive vocabulary knowledge of the first 3,000 most frequent words in Russian predicted reading proficiency at the ACTFL Advanced Low level, while knowledge of the first 5,000 most frequent words predicted reading proficiency at the Advanced High/Superior level. These findings give some basis to expect that as Flagship students work on improving their skills in speaking and writing from ILR level 2 to 3, they are likely to need to increase their productive control of vocabulary at the 3K-5K frequency bands. Talalakina et al. (2019) examined the frequency of lexical items contained in advanced-level Russian language textbooks widely used

in the United States. They found that the LFP of the textbooks on average was 58% (of 1K words), 10% (2K), 5.4% (3K), 3.2% (4K), 2.9% (5K), and 21% (5K+). These findings suggest that texts that provide language input and models for learners across the ILR 2 to 2+ levels use primarily very high frequency words (1K-2K bands) and very low frequency words (5K+). It is possible that texts written by Flagship students on the continuum from ILR Level 2 to 3 will show a similar profile.

Several other studies have pointed in a general way to vocabulary as an important factor in distinguishing oral proficiency performances in Russian at the ILR levels 2-3. Long et al. (2012), using a battery of fine-grained perception and production tasks, found that certain kinds of vocabulary skills (including breadth of knowledge of "basic" nouns and verbs, derivational morphology, collocations and proverbs) could be used to predict oral proficiency scores at ILR 2, 2+, and 3 levels in Russian.

Golonka (2000) examined a sample of 22 OPIs of students who rated ACTFL Intermediate High before spending a semester studying abroad in Russia. Twelve of those students made no perceptible gain on their post-program OPI, and ten made gain, with five achieving ACTFL Advanced, and five achieving Advanced High. She found that the number of unique words (i.e., types) used in the pre-program OPI was a very strong predictor for the kind of progress that the students made by the end of their time abroad. Students who used a broader range of words in the pre-program OPI were more likely to score higher on the post-program OPI. Golonka does not sort the types found in the students' samples by frequency, but her findings make intuitive sense: the broader a students' vocabulary before going abroad, the more likely they are to make sense of the linguistic input that they receive from native speakers and other sources abroad.

Walsh (2014) explored the lexical diversity and sophistication in 47 OPIs produced by 18 different subjects who were studying abroad in Russia for an academic year. For lexical sophistication she calculated the percentage of word types that came from beyond the 2K band in Liashevskaia and Sharov's (2009) frequency list of oral speech. She found that at the Advanced Low level, approximately 18% of the speech consisted of types beyond the 2K level, at Advanced Mid the average percentage was 24%, at Advanced High 29%, and for Superior 27%. Walsh's findings suggest that low frequency vocabulary from bands 3K–5K may make up roughly a quarter of the words in the essays produced by Flagship students who write at the 2, 2+ and 3 levels. Walsh organized her investigation as a cross-sectional study, and so it is not possible to calculate the growth made by individual subjects over the academic year.

From these studies, we can expect that the Flagship students will demonstrate an increase in their low-frequency vocabulary use, and the increase in low-frequency productive vocabulary use might predict higher Writing proficiency scores on the ILR scale. Furthermore, Golonka's (2000) study strongly suggests that the more sophisticated the learners' vocabulary at the pretest, the more likely they are to reach higher levels of writing proficiency according to the ILR scale.

## <A>RESEARCH QUESTIONS

The specific research questions addressed in this study are:

RQ1. What is the Lexical Frequency Profile of Flagship students based on their pretest and posttest writing samples? In what frequency bands is there change over an academic year of study abroad?

RQ2. Is there a change in the students' lexical sophistication as measured by any one band of the LFP or P–Lex over time?

RQ3. Are changes in the students' lexical sophistication as measured by a specific LFP band or P–Lex predictive of gains on the holistic ILR scores on the writing test?

RQ4. What is the nature of the lexis in those LFP bands where students' vocabulary use grows over the course of the program. To what extent are learners using words that show a close similarity between their L1 and L2?

<A>METHODS

*<B>Corpus for Study*

The corpus for this study is 51, 838 words, drawn from the written essays that 48 students from four cohorts produced during the online Writing Proficiency Test at the beginning and the end of their academic year abroad. The basic directions of the writing proficiency test changed little over those four cohorts, with students being instructed to respond to specific prompts in as much detail as possible. Students wrote directly on computer, without access to outside resources such as spelling and grammar checkers. The prompts were designed to elicit language functions expected at the ILR Level 2 (i.e., a personal essay with either extended description or narration), or ILR Level 3 (i.e., argumentative essay on a social and/or political topic).

The largest change in the format of the Writing Proficiency Test during the data collection period was its expansion from a single-essay to a two-essay format, with the time limit similarly expanded from 45 minutes to 90 minutes. The change was implemented in the middle of the 2017–2018 academic year. An overview of the testing scheme, the prompts and writing genres is provided in Table 1. To maintain the balance of pretest and posttest writing for each participant, the researcher removed the second essay that the 2017–2018 cohort wrote at their posttest. The resulting corpus contained 47,979 words.

The essays were rated by trained evaluator(s), and the anonymous essays, prompts and ILR ratings were provided to the researcher by the testing organization.

<INSERT TABLE 1 ABOUT HERE>

TABLE 1

Writing Proficiency Testing Scheme by Cohort

| Cohort | N | Pretest | | | | Posttest | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | # of essays | Time (min.) | Target | Topic area; genre type | # of essays | Time (min.) | Target | Topic area; genre type |
| 2016-17 | 9 | 1 | 45 | ILR-2 | personal; description | 1 | 45 | ILR-3 | social; argumentative |
| 2017-18 | 14 | 1 | 45 | ILR-2 | personal; narrative | 2 | 90 | ILR-3 | political; argumentative |
| | | | | | | | | ILR-3 | social*; argumentative |
| 2018-19 | 9 | 2 | 90 | ILR-2 | personal; narrative | 2 | 90 | ILR-3 | political; argumentative |
| | | | | ILR-3 | political; argumentative | | | ILR-3 | social; argumentative |
| 2019-20 | 16 | 2 | 90 | ILR-2 | personal; narrative | 2 | 90 | ILR-3 | political; argumentative |
| | | | | ILR-3 | political; argumentative | | | ILR-3 | social; argumentative |

*Note.* Target refers to the ILR language functions the prompt intended. *Students' essays to this prompt were removed from the corpus to balance the number of writing samples in pretest and posttest conditions.

*<B>Pre-processing*

The student essays were preprocessed before being automatically parsed and lemmatized. Following accepted practices (Vedder & Benigno, 2016), the researcher fixed small typos and spelling errors (e.g., student's буджет with correct бюджет [budget], исползовать to использовать [to use], польный to полный [full]), and minor word choice errors (e.g., very low frequency экономия [economizing] for the more frequent and intended экономика [economics]). Grammar errors were generally not fixed, unless they might keep the automatic parser from

recognizing a word. There are two reasons to fix spelling errors during pre-processing: first, to improve the automatic parsing and lemmatization, and second, to prevent misspelled words from being wrongly identified as low frequency. A total of 1,172 words were fixed, approximately 2.4% of the corpus.

*<B>Lemmatizing and Further Processing*

The preprocessed essays were then lemmatized using the UDPIPE package in R (Version 4.0.2), which produces an ordered spreadsheet of each word from a student's text in its context along with the related lemma and grammatical information. After lemmatization, the essays were reviewed again by hand to remove proper nouns and terms (e.g., коронавирус [coronavirus], Россия1 [Russia1]) and student calques (e.g., *интеракция [interaction]) that do not exist in Russian. During the review, inaccurate lemmatization by UDPIPE was fixed. The list of lemmas was then compared with the frequency list of lemmas from Liashevskaia and Sharov (2009), which ranks lemmas by frequency from 1 to 20,000. The researcher reviewed all words whose lemmas could not be found in the frequency list. If a student had used a word that fell into one of the frequency list's inconsistencies (i.e., separate frequency ratings for a word prefixed by не- [un-/in-] or each partner in imperfective/perfective verb pairs), the researcher made small adjustments so that the word could be identified with a closely related form in the lemma list (i.e., неэффективность [ineffectiveness] was given the same frequency rating as the lemma эффективность [effectiveness]). Where there were alternative ways of adjusting a word, the researcher chose the way that would result in the student's usage being associated with the higher frequency lemma. The researcher assigned the remaining words to a frequency group "beyond 20,000."

After processing, the corpus contained 45,965 words, and since the students' writing samples varied in length, the LFP for the pretest and posttest essays were generated by dividing the number of words at each frequency band by the total number of words in the writing sample.

*<B>P–LEX*

Despite the utility of LFP for documenting lexical sophistication, the interrelated scores of an LFP analysis pose problems for statistical tests (Kyle, 2020). For this reason, a second measure of vocabulary sophistication was added to this study. P–Lex provides a single measure for how often lower frequency words appear in a text. To determine the P–Lex value a student's writing sample is divided into 10-word segments, removing proper nouns, numbers and punctuation. The number of words in each segment that come from a frequency level beyond the first 1K band are counted and the P–Lex value (lambda) is calculated by taking the mean of the number of low frequency words in each segment (Meara & Miralpeix, 2017). The resulting P–Lex measure can range from 0 (no low frequency words used) to a theoretical 10 (only low frequency words used), but the typical range is 0 to 4.5, and a normal English language text produces a P–Lex value of 2.5 (Milton, 2004). The advantage of P–Lex is that it provides a single measurement of lexical sophistication that is independent of text length.

*<B>Student Outcomes*

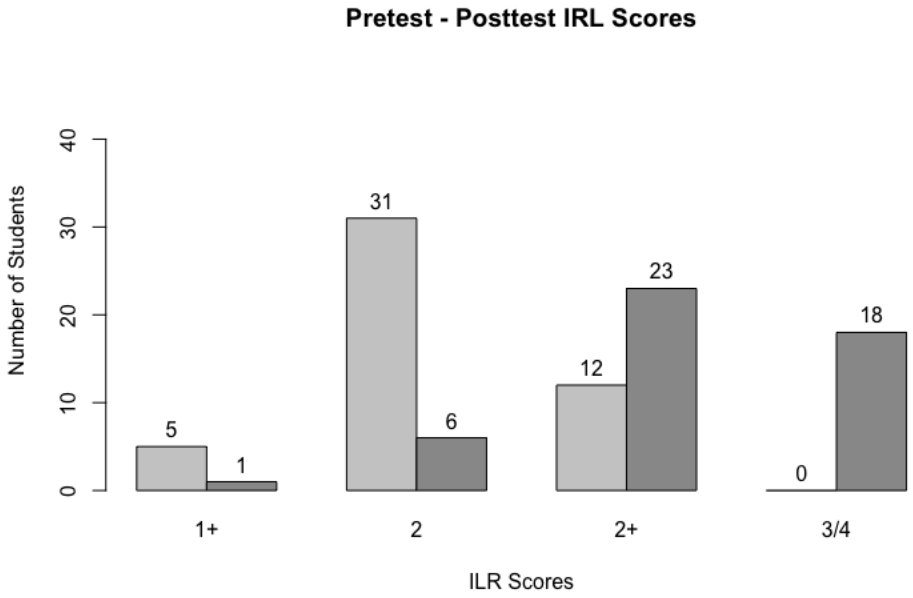
The ILR ratings for the writing samples were provided by the testing organization and student progress from pretest to posttest is summarized in Figure 1. The median writing score was 2 on the pretest, and 2+ on the posttest, although approximately a third of students reached

Level 3 or higher at the posttest. The one writer whose posttest sample was rated ILR 4 was grouped together with the Level 3 scores.

<INSERT Figure 1 ABOUT HERE>

FIGURE 1

Distribution of ILR Ratings




*Note*. For each score level, pretest data is in the left, light grey column; Posttest data is on the right, dark grey column.

To avoid statistical problems with the uneven distribution of test scores and to associate changes in students' lexical sophistication with their improvement in their ILR scores, the ILR scores were recalibrated as the categorical variable Outcome, with four levels of significance: students who reached/stayed at ILR Level 2 ($n$ = 6); those who stayed at Level 2+ ($n$ = 4), those who improved to Level 2+ ($n$ = 19); those who improved to Level 3 or higher ($n$ = 18). Data

from one student whose writing proficiency score decreased over the year from ILR 2 to ILR 1+ was excluded from the analysis in RQ 3.

## <A>RESULTS

The descriptive statistics for the length of student writing samples are presented in Table 2. The difference in number of words from pretest to posttest was normally distributed (D = 0.081, p = .61) according to a Lilliefors test, and a one-tailed *t*-test for dependent samples showed that this difference is significant ($t = -5.304$, $df = 47$, $p_{1\text{-tailed}} < .0001$, ). The students were able to write longer essays at the posttest, increasing their texts by a little over one hundred words on average, and that difference is significant, although the effect size is small (Cohen's $D = 0.77$).

<INSERT TABLE 2 ABOUT HERE>

TABLE 2

Number of Words in Student Writing Samples

| | Words | | | |
|---|---|---|---|---|
| | Mean | SD | Min. | Max. |
| Pretest | 422 | 201 | 67 | 818 |
| Posttest | 535 | 244 | 165 | 1176 |

### *<B>RQ1: Lexical Frequency Profiles*

The first research question concerned the LFP for the pretest and posttest essays.

Table 3 presents the students' word use by frequency band on both the pretest and posttest. The table presents both a raw word count, and the percentage of words in each band. From pretest to posttest there is a decrease in the percentage of 1K words used in the essays (from approximately 81% to 77%), and a small increase in words used from the 2K band (7.8% to 8.7%). There is no change from pretest to posttest at the 3K level, and a very small change at the 4K level. There is an increase of over 1% in words drawn from both the 5K and 5K+ bands. Boxplots representing the LFP for 1K words are presented in Figure 2, while the LFP for the remaining frequency levels is found in Figure 3.

<INSERT TABLE 3 ABOUT HERE>

TABLE 3

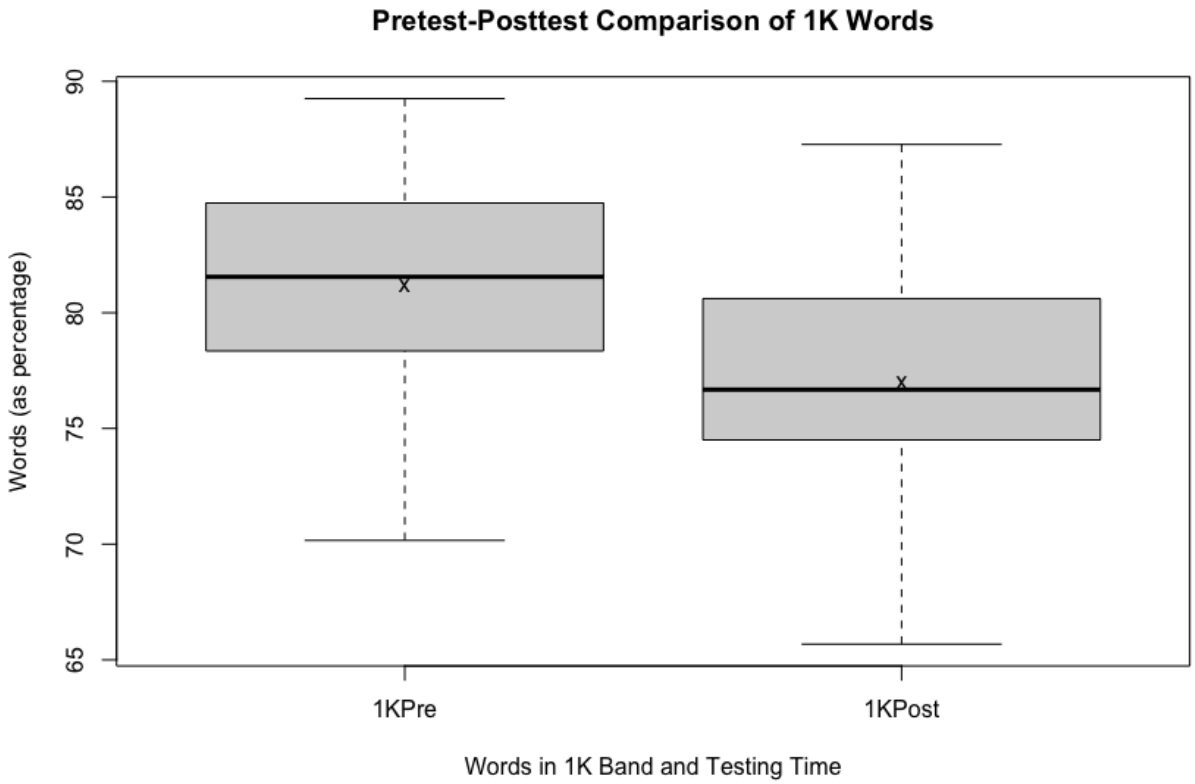
Mean Number of Words by 1,000-Word Frequency Bands (Count and as Percentage)

| Band | Pretest | | | | Posttest | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Min. | Max. | Mean | SD | Min. | Max. |
| | | | | Word count | | | | |
| 1K | 343 | 162 | 54 | 695 | 413 | 188 | 124 | 868 |
| 2K | 34 | 20 | 4 | 88 | 47 | 25 | 7 | 107 |
| 3K | 15 | 9 | 3 | 43 | 19 | 11 | 3 | 45 |
| 4K | 8 | 5 | 0 | 24 | 13 | 8 | 0 | 43 |
| 5K | 4 | 4 | 0 | 14 | 12 | 6 | 2 | 30 |
| 5K+ | 18 | 12 | 2 | 52 | 31 | 18 | 6 | 98 |
| | | | | As percentage | | | | |
| 1K | 81.2 | 4.2 | 70.2 | 89.3 | 77.0 | 3.9 | 65.7 | 87.3 |
| 2K | 7.8 | 1.8 | 4.9 | 12.1 | 8.7 | 2.1 | 4.2 | 14.3 |
| 3K | 3.5 | 1.5 | 1.1 | 8.4 | 3.5 | 1.3 | 1.1 | 6.4 |
| 4K | 2.1 | 1.3 | 0 | 6.5 | 2.5 | 1.2 | 0 | 6.4 |
| 5K | 1.0 | .7 | 0 | 3.3 | 2.3 | 1.0 | .8 | 4.9 |
| 5K+ | 4.2 | 1.9 | .8 | 9.5 | 5.8 | 2.0 | 2.2 | 11.5 |

<INSERT FIGURE 2 ABOUT HERE>

FIGURE 2

Boxplot of 1K Words in Pre- and Posttest Essays, as % of Total Word Count

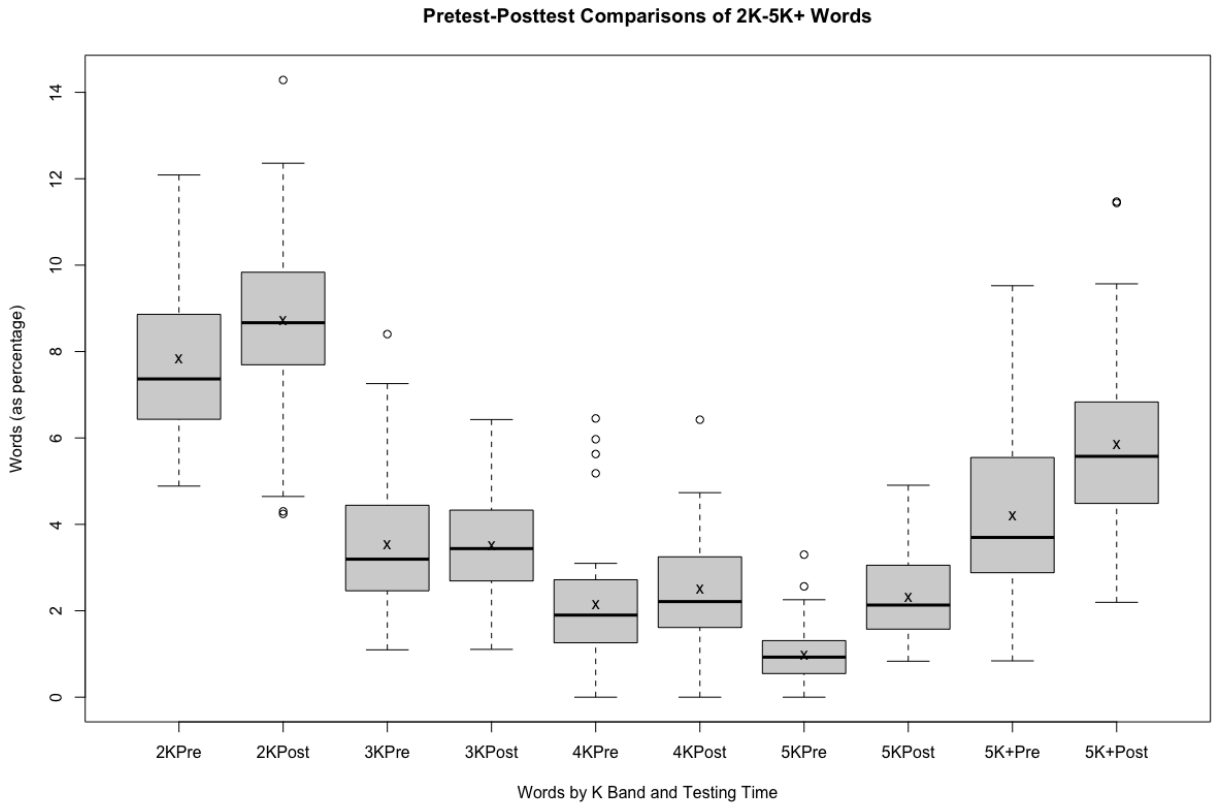



*Note.* Thick line marks the median; x marks mean.

<INSERT FIGURE 3 ABOUT HERE>

FIGURE 3

Boxplot of 2K-5K+ Words in Pre- and Posttest Essays, as % of Total Word Count




*Note*. Thick line marks the median; x marks mean. Pre = pretest; Post = posttest

Since the measures in the LFP are interrelated, only differences in one band can be used in statistical tests (Kyle, 2020). To see which band makes the most sense to include in the analysis for RQs 2 –4, the pretest essays were analyzed according to writing prompt, since as Yoon (2016) and Yoon and Polio (2017) note, vocabulary usage can be dependent on genre, with argumentative essays featuring greater lexical sophistication than narratives. At the pretest all cohorts of students wrote a narrative essay, and two cohorts wrote a second argumentative essay. Table 4 presents the students' word use by frequency band on the pretest grouped by essay type.

In the narrative pretest essays students use more words from the 1K band and fewer from the 2K band (82% and 7%, respectively) than in argumentative essays (79% and 9%, respectively).

The percentages of words from the 1K and 2K bands in the pretest argumentative essays (79% and 9%, respectively) are similar to posttest averages for the 1K and 2K bands (77% and 9% respectively), where all participants wrote two argumentative essays. So the task of writing argumentative essays may be related to the fluctuation in the use of words from the 1K and 2K bands. However, the percentages of low frequency vocabulary (3K to 5K+ bands) are virtually the same in the pretest narrative and argumentative essays. This suggests that the change in the LFP at the 5K and 5K+ bands is better choice for exploring RQs2-4, since increased use of these very low frequency words is more likely to reflect students' vocabulary growth over the course of the year abroad.

<INSERT TABLE 4 ABOUT HERE>

TABLE 4

Mean Number of Words by 1,000-Word Frequency Bands for Pretest Writing Samples by Prompt Type (Count and as Percentage)

| Band | Narrative ($N = 47$) | | | | Argumentative ($N = 25$) | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Min. | Max. | Mean | SD | Min. | Max. |
| | | | | Word count | | | | |
| 1K | 248 | 91 | 54 | 617 | 192 | 74 | 86 | 314 |
| 2K | 23 | 10 | 4 | 56 | 22 | 10 | 5 | 51 |
| 3K | 10 | 5 | 3 | 24 | 8 | 5 | 2 | 19 |
| 4K | 6 | 3 | 0 | 14 | 5 | 4 | 0 | 17 |
| 5K | 3 | 3 | 0 | 12 | 3 | 2 | 0 | 8 |
| 5K+ | 12 | 7 | 2 | 38 | 12 | 8 | 2 | 34 |
| | | | | As percentage | | | | |
| 1K | 81.8 | 4.3 | 70.2 | 89.3 | 78.8 | 5.0 | 63.6 | 85.8 |
| 2K | 7.4 | 1.9 | 3.4 | 11.6 | 9.3 | 2.8 | 3.1 | 15.5 |
| 3K | 3.5 | 1.6 | 1.1 | 8.4 | 3.4 | 1.4 | .7 | 6.6 |
| 4K | 2.1 | 1.5 | 0 | 6.5 | 2.0 | 1.1 | 0 | 4.7 |
| 5K | 1.0 | .9 | 0 | 5.1 | 1.3 | 1.1 | 0 | 3.6 |
| 5K+ | 4.1 | 2.1 | .7 | 10.5 | 5.0 | 2.6 | .7 | 13.6 |

*<B>RQ2: Change in Lexical Sophistication Over Time*

Since the LFP revealed that most of the modest growth in lexical sophistication was seen in the 5K and 5K+ bands, the student results for these two neighboring bands were conflated into a single "5K and Beyond" band. The descriptive statistics for the conflated band and the change in this conflated band (posttest-pretest) are given in Table 5.

<INSERT TABLE 5 ABOUT HERE>

TABLE 5

Conflated 5K and Beyond at Pretest and Posttest and Change of 5K and Beyond

| | Band 5K and Beyond (percentages) | | | |
|---|---|---|---|---|
| | Mean | SD | Min. | Max. |
| Pretest | 5.2 | 2.25 | .8 | 12.1 |
| Posttest | 8.1 | 2.53 | 3.6 | 14.0 |
| $\text{Change}_{\text{Posttest-pretest}}$ | 2.96 | 2.67 | −.7 | 10.9 |

The difference between the pretest and posttest percentage of words from the conflated bands is normally distributed (D = 0.091, $p = .40$) according to a Lilliefors test and a one-tailed *t*-test for dependent samples showed that this difference is significant ($t = -7.705$, $df = 47$, $p_{1\text{-tailed}} < .0001$) and the effect size is medium (Cohen's $D = 1.11$). Based on their usage of words from the 5K and Beyond band, there was significant growth in the students' lexical sophistication over the academic year.

The descriptive statistics for the P–Lex values are listed in Table 6. The difference in P–Lex values from Pretest to Posttest were normally distributed (D = .084, $p = .538$) according to a Lilliefors test, and a one-tailed *t*-test for dependent samples showed that this difference is

significant ($t = -6.788$, $df = 47$, $p_{1\text{-tailed}} < .0001$) and the effect size is small (Cohen's $D = 0.72$). From the P–Lex data, there was significant growth in the students' lexical sophistication over the academic year in words from all bands beyond the first thousand. The average increase of 0.432 indicates that at the posttest, students were adding roughly one lower frequency word for every 20 words of running text.

<INSERT TABLE 6 ABOUT HERE>

TABLE 6

P–Lex Values at Pretest and Posttest and Change of P–Lex Values from Pretest to Posttest.

| | P–Lex Values | | | |
|---|---|---|---|---|
| | Mean | SD | Min. | Max. |
| Pretest | 1.866 | 0.416 | 1.03 | 2.85 |
| Posttest | 2.298 | 0.415 | 1.25 | 3.48 |
| $\text{Change}_{\text{Posttest-pretest}}$ | 0.432 | 0.441 | -0.66 | 1.72 |

*<B>RQ3: Change in Lexical Sophistication and Gains in ILR Scores*

The third research question asks if changes in the students' lexical sophistication as measured by P–Lex values or gains in specific LFP bands are predictive of gains on the holistic ILR scores on the writing test.

The P–Lex measures by Outcomes are presented in Table 7, and a boxplot of the change in P–Lex values can be found in Figure 4. A linear regression model was fit with the Change in P–Lex values as the response and the categorical variable Outcomes as the only predictor (Table

8). The model indicated insignificant fit ($F = 1.596$, $df = 3$, $df = 43$, $p = .20$). The model fit was weak (Mult. $R^2 = 0.10$, Adj. $R^2 = 0.037$), and change in P–Lex values does not seem to be a useful predictor of the kinds of changes in the ILR scores on the writing proficiency posttest.[1]

<INSERT TABLE 7 ABOUT HERE>

TABLE 7

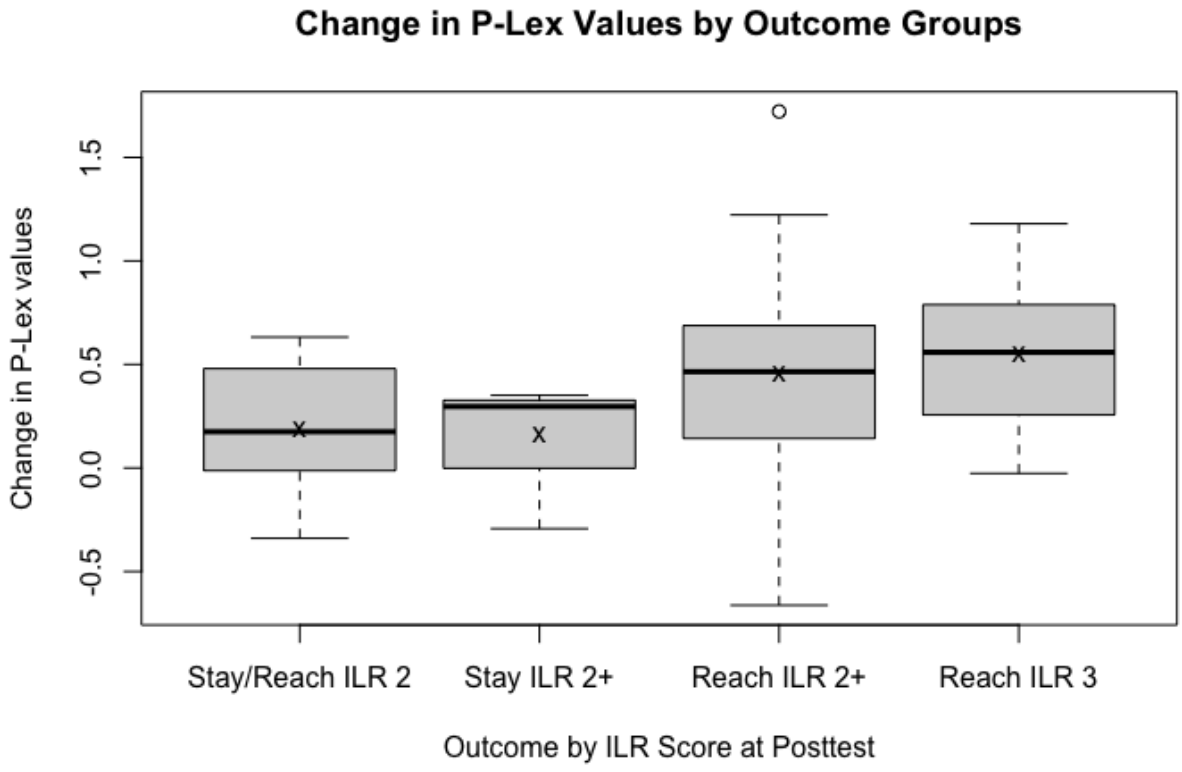
P–Lex Values at Pretest, Posttest, and Change Grouped by Outcomes

| Outcomes | Pretest | | Posttest | | $\text{Change}_{\text{posttest-pretest}}$ | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| Reach/Stay 2 ($N = 6$) | 1.858 | 0.371 | 2.039 | 0.284 | 0.185 | 0.348 |
| Stay 2+ ($N = 4$) | 2.261 | 0.410 | 2.424 | 0.139 | 0.163 | 0.305 |
| Reach 2+ ($N = 19$) | 1.808 | 0.372 | 2.267 | 0.484 | 0.458 | 0.535 |
| Reach 3 ($N = 18$) | 1.846 | 0.047 | 2.395 | 0.401 | 0.549 | 0.357 |

<INSERT FIGURE 4 ABOUT HERE>

FIGURE 4

Boxplot of Change in P–Lex, Grouped by Outcomes




*Note*. Thick line marks the median; x marks mean.

<INSERT TABLE 8 ABOUT HERE>

TABLE 8

Coefficients from the Linear Model: Change of P–Lex as Response; Outcome as Predictor

| | Estimate | 95%-CI | *se* | *t* | $p_{2\text{-tailed}}$ |
|---|---|---|---|---|---|
| Intercept (Outcome: Reach/Stay 2) | 0.185 | [−0.175, 0.545] | 0.178 | 1.037 | 0.306 |
| Outcome: Stay 2+ | -0.022 | [−0.591, 0.547] | 0.282 | −0.078 | 0.938 |
| Outcome: Reach 2+ | 0.274 | [−0.139, 0.686] | 0.205 | 1.337 | 0.188 |
| Outcome: Reach 3 | 0.364 | [−0.052, 0.779] | 0.206 | 1.766 | 0.085 |

The percentage of words from the conflated band 5K and Beyond grouped by Outcomes is presented in Table 9, with a boxplot of the change in the conflated 5K and Beyond values presented in Figure 5.

<INSERT TABLE 9 ABOUT HERE>

TABLE 9

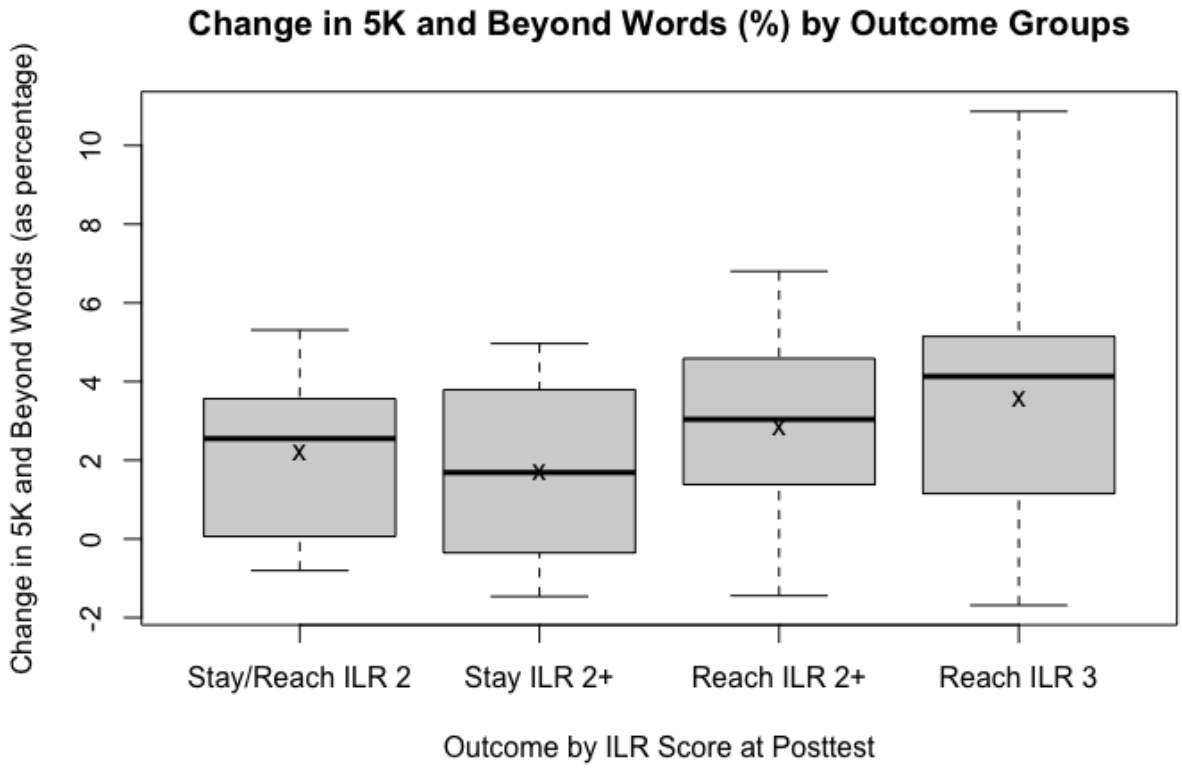
Change in 5K and Beyond (Posttest-pretest, as Percentage), Grouped by Outcomes

| Outcomes | Pretest | | Posttest | | Change 5K and Beyond (as percentage) | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| Reach/Stay 2 (*N* = 6) | 4.7 | 2.66 | 6.9 | 1.44 | 2.2 | 2.27 |
| Stay 2+ (*N* = 4) | 7.0 | 3.46 | 8.7 | 1.66 | 1.7 | 2.73 |
| Reach 2+ (*N* = 19) | 4.5 | 1.80 | 7.3 | 2.09 | 2.8 | 2.47 |
| Reach 3 (*N* = 18) | 5.7 | 2.19 | 9.2 | 3.03 | 3.6 | 3.05 |

<INSERT FIGURE 5 ABOUT HERE>

FIGURE 5

Change in 5K and Beyond Words (%), Grouped by Outcome Groups




*Note.* Thick line marks the median; x marks mean.

A linear regression model was fit with the change in the 5K and Beyond band as the response and the categorical variable Outcomes as the only predictor. The model (Table 10) indicated insignificant fit ($F = 0.76$, $df = 3$, $df = 43$, $p = .522$), and the model fit was weak (Mult. $R^2 = 0.05$, Adj. $R^2 = -0.016$). The change in students' use of very low frequency vocabulary from the 5K and Beyond band does not seem to predict the outcome on the posttest.

<INSERT TABLE 10 ABOUT HERE>

TABLE 10

Coefficients from the Linear Model: Change of 5K and Beyond as Response; Outcome as Predictor

| | Estimate | 95%-CI | *se* | *t* | $p_{2\text{-tailed}}$ |
|---|---|---|---|---|---|
| Intercept (Outcome: Reach/Stay 2) | 0.022 | [−0.0003, 0.0444] | 0.011 | 1.991 | 0.0529 |
| Outcome: Stay 2+ | −0.005 | [−0.0402, 0.0305] | 0.018 | −0.277 | 0.7828 |
| Outcome: Reach 2+ | 0.006 | [−0.0192, 0.0320] | 0.013 | 0.503 | 0.6176 |
| Outcome: Reach 3 | 0.014 | [−0.0120, 0.0396] | 0.013 | 1.077 | 0.2876 |

*<B>RQ4: Qualitative Analysis of Low Frequency Vocabulary*

Since the LFP reveals that much of the student's growth in vocabulary comes from words at the lowest frequency ranges (5K and Beyond), and that the overall growth at these ranges is significant, a qualitative analysis was made of all the words from these frequencies from both the pretest and posttest essays to examine the nature of the words that the writers chose to use.

In the whole corpus of essays, there are 3,134 tokens of 1,310 types which come from the 5K and Beyond frequency range; 1,080 tokens occur in the pretest essays, and 2,054 in the posttest (Table 11).

<INSERT TABLE 11 ABOUT HERE>

TABLE 11

Tokens and Types from 5K and Beyond Bands, Grouped by Test Time and Category

| | Pretest | | Posttest | |
|---|---|---|---|---|
| | Count | Percentage | Count | Percentage |
| ***Tokens*** | | | | |
| Close to English | 386 | 37.5 | 664 | 32.3 |
| Task-related | 89 | 8.2 | 592 | 28.8 |
| Hobby | 41 | 3.8 | 0 | 0 |
| Technology | 22 | 2.0 | 6 | .3 |
| Academic | 84 | 7.8 | 15 | .7 |
| Uncategorized | 459 | 42.5 | 777 | 37.8 |
| Total | 1,080 | | 2,054 | |
| ***Types*** | | | | |
| Close to English | 149 | 29.8 | 211 | 26.0 |
| Task-related | 8 | 1.6 | 37 | 4.6 |
| Hobby | 23 | 4.6 | 0 | 0 |
| Technology | 12 | 2.4 | 5 | .6 |
| Academic | 30 | 6. | 7 | .9 |
| Uncategorized | 278 | 55.6 | 550 | 67.9 |
| Total | 500 | | 810 | |

For words that are close to English, there are 386 (37.5%) tokens in the pretest, and 664 (32.3%) tokens in the posttest. These tokens include recent borrowings (e.g., аккаунт [account] in online transactions), recent borrowings modified for Russian parts of speech (e.g., креативный [creative], with adjective ending -ный [-nyi], блокировать [to block], with verb suffix -ировать [-irovat']), or words borrowed from a common source (e.g., география [geography]). The number of tokens of task-related words (i.e., коррупция [corruption], a Russian equivalent of a word that figures prominently in the English prompt) grew from 89 (8.2%) at the pretest to 592 (28.8%) at the posttest.

Of the 1,310 types, 149 (29.8%) in the pretest and 211 (26%) in the posttest have a close similarity between English and Russian. There were only 8 (1.6%) types closely related to the task in the pretest data, and 37 (4.6%) in the posttest. In the pretest, 65 (13%) types cover aspects of students' hobbies, technology use, and academic life which appear in the personal narrative and descriptive essays. Only 12 such types (1.5%) are present in the posttest essays.

Of these 1,310 low-frequency types, 828 appear once and only once in the whole corpus. In the pretest essays, there are a total of 316 unique types, and 44 out of the 48 writers contribute at least one of these unique types; 13 writers contribute 10 or more unique types, and one writer makes the largest contribution of 25 unique types. The posttest essays feature 512 unique types. All but one of the 48 writers contribute at least one of these unique types, 23 contribute 10 or more such words, and one writer makes the largest contribution of 47 unique words. The largest contributions at the pretest and posttest were made by different writers.

The 1,314 low-frequency types were analyzed by parts of speech (Table 12). As a point of comparison, the POS breakdown for the analogous frequency levels from the Russian frequency list is also presented in Table 12. Nouns account for the largest portion of the distinct lemmas (212 [42.4%] in pretest; 304 [37.5%] at posttest), followed by verbs, adjectives and adverbs, and other. The number of low-frequency verbs almost doubles from pretest (149, 29.8%) to posttest (288, 35.6%). This increase is quite marked since in the frequency dictionary as a whole only 28% of the entries at these levels are verbs.

<INSERT TABLE 12 ABOUT HERE>

TABLE 12

Types From K5 and Beyond by Part of Speech

| POS | Pretest | | Posttest | | Frequency Dictionary | |
|---|---|---|---|---|---|---|
| | Count | Percentage | Count | Percentage | Count | Percentage |
| Nouns | 212 | 42.4 | 304 | 37.5 | 6,904 | 43.2 |
| Verbs | 149 | 29.8 | 288 | 35.6 | 4,523 | 28.3 |
| Adjectives | 104 | 20.8 | 156 | 19.3 | 3,394 | 21.2 |
| Adverbs | 32 | 6.4 | 56 | 6.9 | 930 | 5.8 |
| Other | 3 | .6 | 6 | .7 | 249 | 1.6 |
| Total | 500 | | 810 | | 16,000 | |

The types were analyzed from the point of view of Russian word formation, which can give a sense of the complexity of the words as well as indicate their basic semantic field. Compound nouns and adjectives, which are typical of written Russian (e.g., благополучие

[welfare], взаимопонимание [mutual understanding], вышеупомянутый [aforementioned], are rarer in the pretest writing than the posttest essays (56/84 items, respectively). In Russian, specific word suffixes are strongly related to certain kinds of semantic notions. Nouns ending in suffixes like -ец [-ets], -ник [-nik], -тель [-tel'], -анин [-anin] are strongly associated with human actors, while nouns suffixed with -ость [-ost'], -ность [-nost'], -ация [-atsiia], -ство [-stvo], -ствие [-stvie], -ание [-anie], -ение [-enie] are strongly associated with abstract notions or processes (Townsend, 1975). Writers more than double the number of words with abstract suffixes from pretest (66 such nouns) to the posttest (162 such nouns), while the number of nouns associated with human actors remains roughly the same (24 pretest, 25 posttest).

Writers increase the number of verbs derived with the suffixes -овать [-ovat'] and -ировать [-irovat'] (22 pretest, 52 posttest), and they increase the number of imperfective verbs derived with the infixes -ыва- [-yva-] or -ива- [-iva-] (10 pretest, 30 posttest). Writers more than double the number of low-frequency reflexive verbs from 40 at pretest to 90 at posttest, showing greater facility with verbs whose transitivity changes with the presence or absence of the reflexive -ся [-sya] particle. For example, transitive forms of the lemma уменьшать/уменьшить [to reduce, to shrink] appear in the pretest and posttest 5 times each. The intransitive forms of the lemma уменьшаться/уменьшиться [to reduce, shrink, be reduced] appear once in the pretest, but 8 times in the posttest.

## <A>DISCUSSION

### *<B>Lexical Frequency Profiles*

The change in the lexical frequency profiles in this study shows that Flagship students during their year abroad make broad gains in their vocabulary knowledge, particularly at the

lowest frequency levels. The change in uses in the 1K and 2K bands from pretest to posttest are harder to interpret. There may be a task effect, as Yoon (2016) and Yoon and Polio (2017) have found. It may also be a function of proficiency level. Just as Vedder and Benigno (2016) observed that native speakers in their study used fewer words from the 1K frequency band than non-natives at CEFR level B1, the Flagship students use more 1K words in the pretest (when the majority score at ILR Level 2, very similar to CEFR B1) but pattern more like native speakers in using fewer 1K words at the posttest, when the majority score at ILR levels 2+ and 3.

The increase in use of words from the 5K and Beyond band as students reach the 2+ and 3 levels in writing seems to fit with the findings of Hacking and Tschirner (2017) for reading; namely, that word knowledge at the 5K and Beyond band appears associated with the ability to understand and compose texts that feature language functions such as hypothesis, argumentation and supported opinions. The overall contours of the students' posttest LFP are not unlike the patterns that Talalakina et al. (2019) found in their examination of advanced Russian language textbooks, where low frequency words from the 5K and Beyond band made up 24% of the books' vocabulary. The writing samples in the current study show less use of low frequency words than the range of lexical sophistication that Walsh (2014) found in students' speech across the ACTFL Advanced and Superior levels. In the writing samples, words belonging to the 3K and beyond bands on average account for 10% of the text at the pretest and 14% at posttest, about half the quantity found in Walsh's sample (18%–29%). One possible reason for this difference may be that at the advanced-superior levels in spontaneous speech students might be more likely to use words that they know orally, but whose exact spelling and morphology they do not control. In a writing proficiency test, students may limit their risk taking, gravitating towards words that they can spell with confidence. Indeed, the typos that were fixed in this

study's data were generally keyboarding errors or very small phonetic misspellings that do not suggest that students were taking large risks in their word choice.

*<B>Change of Lexical Sophistication over Time*

There is statistically significant growth in use of lower frequency vocabulary from the pretest essays to the posttest essays as measured by both by the use of words from the 5K and Beyond band and by the P–Lex statistic, although the effect sizes are moderate and small, respectively. In many ways this is the expected result given the extensive amount of language input that students have in a year-long study abroad context, where they have approximately 4 hours of class in Russian daily, individual tutoring, home-stay interactions, and a Russian-language internship. Furthermore, the nature of the prompts in the posttest likely pushed students to produce longer texts, and the additional volume of writing increases the likelihood of a broader range of vocabulary.

*<B>Change in Lexical Sophistication and Gain in ILR Scores*

Growth of students' vocabulary as measured by change in 5K and Beyond words and P–Lex values showed no significant correlations with the writing proficiency outcomes. A number of factors may contribute to the lack of significance here. First, the 5K and Beyond band and P–Lex are rather blunt measures of low frequency vocabulary. Changes in them over time may not be sufficiently fine-grained to capture the potential relationship between the students' lexical growth and their ILR scores. Second, the ILR scores are a holistic evaluation of the students' writing, where vocabulary plays a part, but so do other factors such as meeting specific linguistic functions, demonstrating morphological and syntactic precision, and having enough

content knowledge to discuss the political and social topics dictated by the prompt. Third, the small sample size ($N = 47$) is close to the minimum data points needed for regression modeling, and the number of students in each outcomes group varies considerably. Furthermore, the data in this study are examined longitudinally, so that changes in lexical sophistication are being compared with changes in ILR scores. A cross-sectional study of independent samples of writing tests may indeed find that lexical sophistication can be a significant predictor of ILR scores, or at least may be capable of distinguishing major proficiency levels, such as ILR 2 from ILR 3.

Despite the lack of significance, a closer look at the data shows two interesting trends: first, the variance in the P–Lex measures for students who reach the ILR 2+ level in writing is very large (.286, Figure 4). This suggests that students earning a score of 2+ in writing can have rather different lexical profiles, and some can compensate for a lower level of lexical sophistication with other means (i.e., circumlocution) to meet the expected linguistic functions at the 2+ level. However, a student might use a very broad range of sophisticated vocabulary for the 2+ level, but fail to meet the linguistic functions and/or control of morphology and syntax expected at ILR 3. This might also explain the lexical profile of those students who scored ILR 2+ at both the pretest and posttest. At the pretest their average P–Lex score (2.261) was much higher than that of the students in all the other outcomes groups (range 1.80–1.858). At the posttest, the students who remained at the 2+ level again had the highest average P–Lex score (2.424), slightly higher than the P–Lex average for students who reached ILR 3 (2.395). The students who stay at 2+ had a very sophisticated vocabulary to start with, and they continued to build on it, but that did not compensate for whatever weaknesses that kept their samples from being rated ILR 3 at either testing time. The situation with the group that remained at the 2+ level is different from Golonka's (2000) findings that a broad vocabulary range at the pretest OPI

in Speaking was predictive of students making gains into the Advanced and Advanced High levels.

*<B>Qualitative Analysis of Low Frequency Vocabulary*

A qualitative examination of the words from the low-frequency bands that the students used in their essays indicates how broadly students' vocabulary and language awareness grow over the academic year abroad. When required to produce extended discourse and argumentation typical of the formal written Russian, students can use a wide range of low frequency vocabulary. They increase their use of compound words and nouns marked with suffixes expressing abstract notions or processes, vocabulary that is typical of written Russian and the argumentative essay. The increase in verbs in the posttest writing samples, at the same time as the number of nouns expressing agent remains the same, suggests that students over time either become more facile at structuring their sentences to use more abstract agents as subjects of these verbs, or they can deploy agentless constructions using intransitive main verbs marked by the reflexive particle.

That a quarter to a third of the low-frequency words that students draw on have a close relationship with English does not suggest that they are relying too much on cognates to inflate their lexical sophistication. Comer (2021) documented that about 21% of the words in the first 5,000 words on the Liashevskaia and Sharov's frequency list can be classified as international words. Thus, international words are a significant and unavoidable feature of Russian vocabulary. Moreover, at least two thirds of the low frequency words that students use draw on Russian roots, which suggests a broad knowledge of native Russian roots and derivational morphology principles. Indeed, the writers show flexibility in deploying words built on a single

stem, but that differ in part of speech. For example, the noun вражда [enmity], the adjectives вражеский [enemy] and враждебный [hostile] are all found in the corpus. Students are also sensitive to how the derivational morphology principles work with borrowed words, deploying the related words оппонент [opponent], оппозиция [opposition], оппозиционер [person in the opposition], оппозиционный [opposition (adjective)].

The qualitative analysis of the low-frequency lexical items used in the Flagship students' writing also shows evidence of task and prompt effects. In the pretest personal essays a number of low frequency words reflect students' hobbies or academic backgrounds. In the posttest essays, Russian equivalents of English words that appeared in the very lengthy prompts for argumentative essays make up over a quarter of the low-frequency tokens in the students' writing.

## <A>IMPLICATIONS, LIMITATIONS AND CONCLUSIONS

Several pedagogical implications arise from the very broad vocabulary knowledge that learners need to develop to reach ILR Levels 2+/3 in writing. Gaining controld of that range of vocabulary does present a significant learning burden, and teachers and curriculum designers should provide every affordance for learners to make both intentional and incidental vocabulary learning efficient and effective. As learners work with more and more varied input from the late novice stage onward, teachers can implement vocabulary learning strategies that help them notice the regularity of Russian derivational morphology (Barcroft 2012, Comer 2021). Teachers should also train students to use Russian word formation principles to increase their recognition vocabulary and improve their reading fluency. For students at higher levels of proficiency, teachers will need to deepen learners' knowledge of derivational morphology, which can help

them recognize fine distinctions between closely related words. Vocabulary instruction at that level should also draw learners' attention to individual words' specific features, such as the transitivity of verbs and their capacities to be made intransitive and passive, since constructions where grammatical subjects show a range of agency are frequent in argumentative essays. Finally, vocabulary instruction, particularly for students trying to build facility in argumentative writing, will need to balance the presentation of individual lexical items with multiword units that also play an important part in framing arguments and in mustering supporting points. Nevertheless, learners' vocabulary breadth alone seems unlikely to change the proficiency rating of any particular piece of their writing, and so while they expand their vocabularies, learners also need to develop strong monitoring skills to notice and correct surface level errors, particularly in subject-verb and adjective-noun agreement, especially as sentence length expands.

A significant limitation for this study of the development of lexical sophistication in the writing of very advanced L2 learners of Russian is the genre-based writing prompts used in the proficiency tests, and the fact that the genres are not balanced at the pretest and posttest. The deployment of these different prompts makes sense from a testing perspective: at the pretest the narrative essay acts as a level check for ILR-2, while the argumentative essay probes for Level 3 functions; at posttest the two argumentative essays act as checks for Level 3 functions over a range of topics. Nevertheless, for studying linguistic growth or vocabulary development, this scheme is problematic for several reasons. Students have been shown to write more in response to prompts for argumentative essays, and that increased output is thought to help students make proficiency gains in their writing (Brown et al., 2009). Yoon (2016) and Yoon and Polio (2017) have both shown lexical differences between narratives and argumentative essays whether they were written by ESL students or native English speakers. For both types of writers,

argumentative essays have greater lexical sophistication (i.e., longer words and lower frequency words), whereas narratives have greater lexical diversity (i.e., more varied word use). Furthermore, participants relied more on formulaic phrases in the argumentative essays than in narratives. A future study might restrict analysis just to the genre of argumentative essay, selecting one essay from the pretest and one from the posttest.

This study has looked at vocabulary in the writing of advanced L2 Russian learners from a frequency perspective. It shows the growth of students' vocabulary over time at the word level and documents some aspects of the nature of that growth, particularly students' use of words from the lowest frequency levels. This study does not try to qualify or quantify the appropriateness of those lexical choices in the students' essays, although ILR scores stand as an indirect indicator that the lexical choices are generally appropriate, since writing at ILR Level 2 and above must be "understandable to a native reader not used to reading the writing of foreigners." The study also highlights that lexical sophistication by itself seems to play only one part in the holistic scoring rubric of the ILR scale, at least when looking at dependent pretest-posttest scores. Future studies should explore a complex of features in the students' vocabulary usage in relationship to ILR outcomes in a cross-sectional design. Further research should also examine students' lexical sophistication in relationship to their ability to deploy multiword units in their argumentative writing, since such linguistic formulae or chunks are quite noticeable in the posttest argumentative essays.

<A>ACKNOWLEDGMENT

The author is grateful to Jane Hacking and Lynne deBenedette for their thoughtful comments on earlier drafts of this manuscript. This study has made use of data made available for this purpose by agreement with American Councils for International Education, that is responsible for the

Russian Writing Proficiency Test. The author would like to thank Nadra Garas and Dan Davidson of American Councils for their help in obtaining the student writing samples used in this research. The Language Flagship Russian Program is sponsored by the Language Flagship at the Defense Language and National Security Education Office (DLNSEO), through a grant administered by The Institute of International Education (IIE) in cooperation with Bryn Mawr College. This research article does not necessarily reflect the position or policy of American Councils for International Education, IIE or the Government; and no official IIE or Government endorsement should be inferred.

<A>NOTES

[1]In describing the regression formulae, I follow the usage of Gries (2021) who uses the terms "predictor" variable for what one could think of as the independent variable, and the term "response" for the dependent variable.

<A>REFERENCES

Barcroft, J. (2012). *Input-Based Incremental Vocabulary Instruction*. TESOL International Association.

Bardel, C., & Lindqvist, C. (2011). Developing a lexical profiler for spoken French and Italian L2. The role of frequency, cognates and thematic vocabulary. In L. Roberts, G. Pallotti, & C. Bettoni (Eds.), *EUROSLA Yearbook 11* (pp. 75–93). John Benjamins. https://doi.org/10.1075/eurosla.11.06bar

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System, 69*, 65–78. http://dx.doi.org/10.1016/j.system.2017.08.004

Brown, N. A., Bown, J., & Eggett, D. L. (2009). Making rapid gains in second language writing: A case study of a third-year Russian language course. *Foreign Language Annals*, *42*, 424–452. https://doi.org/10.1111/j.1944-9720.2009.01038.x

Comer, W. (2021). Russian's most frequent words and implications for vocabulary instruction. *Russian Language Journal, 71*, 115–138.

Crossley, S. A., & Kyle, K. (2018). Assessing writing with the tool for the automatic analysis of lexical sophistication (TAALES). *Assessing Writing, 38*, 46–50. https://doi.org/10.1016/j.asw.2018.06.004

Crossley, S. A., Salsbury, T. & McNamara, D. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing, 29*, 243–263. https://doi.org/10.1177/0265532211419331

Davidson, D. E. (2010). Study abroad: When, how long, and with what results? New data from the Russian front. *Foreign Language Annals, 43*, 6–26. https://doi.org/10.1111/j.1944-9720.2010.01057.x

Eguchi, M. & Kyle, K. (2020). Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *Modern Language Journal, 104*, 381–400. https://doi.org/10.1111/modl.12637

Forsberg–Lundell, F., & Lindqvist, C. (2012). Vocabulary aspects of advanced L2 French: Do lexical formulaic sequences and lexical richness develop at the same rate? *LIA: Language, Interaction and Acquisition, 3*, 73–92. https://doi.org/10.1075/lia.3.1.05

Golonka, E. (2000). *Identification of salient linguistic and metalinguistic variables in the prediction of oral proficiency gain at the advanced-level threshold among adult learners of Russian* (Publication No. 9995979). (Unpublished doctoral dissertation). Bryn Mawr College, Bryn Mawr, Pennsylvania.

Gries, S. (2021). *Statistics for Linguistics with R: A Practical Introduction*, (3rd ed.). De Gruyter Mouton.

Hacking, J. F., & Tschirner, E. (2017). The contribution of vocabulary knowledge to reading proficiency: The case of college Russian. *Foreign Language Annals, 50*, 500–518. https://doi.org/10.1111/flan.12282

Higginbotham, G. & Reid, J. (2019). The lexical sophistication of second language learners' academic essays. *Journal of English for Academic Purposes, 37*, 127–140. https://doi.org/10.1016/j.jeap.2018.12.002

Horst, M., & Collins, L. (2006). From *faible* to strong: How does their vocabulary grow? *The Canadian Modern Language Review, 63*, 83–106. https://doi.org/10.1353/cml.2006.0046

Interagency Language Roundtable. (2011). *Language Skill Level Descriptions – Writing*. https://www.govtilr.org/Skills/ILRscale5.htm

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *Modern Language Journal, 102*, 120-141. https://doi.org/10.1111/modl.12447

Kyle, K. (2020). Measuring lexical richness. In Webb, S. (Ed.), *The Routledge handbook of vocabulary studies* (pp. 454–76). Routledge.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 4*, 757–786. https://doi.org/10.1002/tesq.194

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

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*, 647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Liashevskaia, O. N., & Sharov, S. A. (2009). *Chastotnyi slovar' sovremennogo russkogo iazyka (na materialakh Natsional'nogo korpusa russkogo iazyka)* [Frequency dictionary of contemporary Russian (based on the Russian National Corpus)]. Azbukovnik. http://dict.ruslang.ru/freq.php

Lindqvist, C., Bardel, C., & Gudmundson, A. (2011). Lexical richness in the advanced learner's oral production of French and Italian L2. *IRAL, International Review of Applied Linguistics in Language Teaching 49*, 221–240. https://doi.org/10.1515/iral.2011.013

Long, M. H., Gor, K., & Jackson, S. (2012). Linguistic correlates of second language proficiency. Proof of concept with ILR 2-3 in Russian. *Studies in Second Language Acquisition, 34*, 99–126. https://doi.org/10.1017/S0272263111000519

Meara, P. M. & Miralpeix, I. (2017). *Tools for researching vocabulary*. Multilingual Matters.

Milton, J. (2004). Comparing the lexical difficulty of French reading comprehension exam tests. *Language Learning Journal*, *30*, 5–11. https://doi.org/10.1080/09571730485200171

Murphy, D., & Evans-Romaine, K. (Eds.). (2017). *Exploring the U.S. Language Flagship Program: Professional competence in a second language by graduation*. Multilingual Matters.

Siyanova–Chanturia, A., & Spina, S. (2020). Multi-Word expressions in second language writing: A large-scale longitudinal learner corpus study. *Language Learning, 70*, 420–463. https://doi.org/10.1111/lang.12383

Talalakina, E., Brown, A. N., & Kamrotov, N. (2019). Lexical profiles of L2 Russian textbooks. *Russian Language Journal, 69*, 195–212.

Talalakina, E., Stukal, D., & Kamrotov, M. (2020). Developing and validating an academic vocabulary list in Russian: A computational approach. *Modern Language Journal, 104*, 618–646. https://doi.org/10.1111/modl.12664

Townsend, C. (1975). *Russian Word-Formation*. Slavica Publishers.

Tschirner, E. (2011). Reasonable expectations: Frameworks of reference, proficiency levels, educational standards. *Studie z aplikovane lingvisitiky, 1*, 101–119.

Vedder, I., & Benigno, V. (2016). Lexical richness and collocational competence in second-language writing. *International Review of Applied Linguistics in Language Teaching, 54*, 23–42. https://doi.org/10.1515/iral-2016-0015

Vyatkina, N., & Housen, A. (2021). Complexity. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge handbook of SLA and corpora* (pp. 318–331). Routledge.

Walsh, I. K. (2014). *Mapping the internalization of parts of speech in the spoken interlanguage of adult American learners of Russian from level 1 to level 3* (Publication No. 3644812). (Unpublished doctoral dissertation). Bryn Mawr College, Bryn Mawr, Pennsylvania.

Yoon, H. (2016). Association strength of verb-noun combinations in experienced NS and less experienced NNS writing: Longitudinal and cross-sectional findings. *Journal of Second Language Writing, 34*, 42–57. https://doi.org/10.1016/j.jslw.2016.11.001

Yoon, H., & Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly, 51*, 275–301. https://doi.org/10.1002/tesq.296