Exploring Lexicogrammatical Features as Predictors of Writing Quality in Second Language Spanish: A Multivariate Analysis

by

Carla Herminia Consolini

A dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in Linguistics

Dissertation Committee:

Kristopher Kyle, Chair

Keli Yerian, Core Member

Melissa Bease-Berk, Core Member

Robert Davis, Institutional Representative

University of Oregon

Spring 2025

© 2025 Carla Herminia Consolini

DISSERTATION ABSTRACT

Carla Herminia Consolini

Doctor of Philosophy in Linguistics

Title: Exploring Lexicogrammatical Features as Predictors of Writing Quality in Second Language Spanish: A Multivariate Analysis

This dissertation investigates the relationship between the use of lexicogrammatical features and L2 Spanish writing quality using a multivariate modeling approach. While a substantial body of research has demonstrated that lexical diversity and sophistication are strong predictors of L2 writing quality in English (Besten & Granger, 2014; Kim et al., 2018; Kyle & Crossley, 2016, inter alia), comparatively little is known about if and how these features can predict L2 Spanish writing quality in the form of human judgements. The current study addresses this gap by analyzing a corpus of 400 graded essays written by L2 Spanish learners and examining the predictive power of various indices of lexical richness, including lexical diversity, word frequency, bigram strength of association, and bigram dependencies, on human judgments of writing quality. To do so, the study compares multiple operationalizations of morphological normalization (e.g., raw forms, lemmas, and lemmas tagged with partial or full verbal information) to evaluate their impact on the calculation of lexical indices. The results show that operationalizations retaining more verb morphology tend to improve the predictive power of certain indices, especially for lexical diversity and bigram dependencies. Final models combining multiple features account for approximately 30–34% of the variance in human ratings of writing quality, indicating the value of a multivariate approach. Findings from this dissertation contribute to our understanding of productive lexical proficiency in L2 Spanish and highlight the importance of tailoring computational indices and morphological strategies to the typological characteristics of the target language. Implications are discussed for second language acquisition theory, writing assessment practices, and pedagogy.

ACKNOWLEDGMENTS

There are many people whom I would like to acknowledge and thank for the existence of this dissertation. This project would not have come to fruition if it were not for the time, expertise, and support of many.

First and foremost, I want to thank my advisor, Dr. Kristopher Kyle, without whose support and guidance this project would not have been possible. He spent countless hours brainstorming with me, listening to my ideas, and making them possible by grounding them in reality. Kris has also been the most patient and understanding academic I have had the pleasure to learn from and collaborate with. He has dedicated a lot of his time to guiding me through statistical analyses, complex concepts, and the intricacies of the academic world. To him, I will be eternally grateful for allowing me to be myself and pursue what I was interested in, while always ensuring that I was learning and developing my skills as a professional. He has taught me that leaders can be kind without jeopardizing professionalism and excellence, and I promise to practice and share what I have learned from him with others.

I would also like to thank everyone at the Department of Linguistics at the University of Oregon for inspiring me in their classes, for teaching me how to be a better academic and professional, and for always being open to discussions that helped me think outside the box. Particularly, to my committee members: Dr. Keli Yerian, for the many warm and intellectually stimulating conversations; Dr. Melissa Bease-Berk, for her willingness to participate in my committee and her openness to answer any questions; and Dr. Robert Davis from the Department of Romance Languages, for his Spanish expertise; and to all of them, for their time and dedication to see me through this dissertation.

I would also like to thank Brittany, Mery, and Irene for their collegial academic support. These amazing colleagues and friends have helped me brainstorm when I've been stuck, have cheered me on, supported me, and accompanied me through the years it has taken to get here, and this dissertation exists, partly, because of them. Similarly, I want to thank the rating team, Ángela and Mery, who made possible the creation of an open-source sample of human-graded essays in Spanish. Their expertise was invaluable in creating quality work, and this dissertation would not have been possible without them.

I want to thank Johanna Jimenez, who was the first person to welcome me to Oregon and whose daily support and familiar warmth made it possible for me to feel at home away from home since 2019. This dissertation partly exists, most definitely, because of her. Thank you. I also want to thank, profoundly, my parents, whose support throughout the years and the distance has added love and encouragement to this journey away from home. Their belief in me became a belief in myself.

Humans do not live, and human innovation does not happen, in a vacuum. As social beings, we need community to thrive and be the best version of ourselves, something that makes projects like this one, and many others far bigger, possible. Thank you to my community.

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION 14

II. SECOND LANGUAGE WRITING AND FEATURES OF LEXICOGRAMMATICAL USE 21

Second Language Writing Research 21

Features of Lexicogrammatical Use 24

Word Level 24

Lexical Diversity 24

Lexical Sophistication 28

Beyond Word Level 31

Bigrams and trigrams 31

Bigram dependencies 37

Spanish Second Language Writing Research 40

Lexicogrammatical Differences Between English and Spanish 41

Inflectional Systems 43

Declension 43

Conjugation 44

Pronouns 46

Subject Pronouns 47

Object Pronouns 48

Reflexive Pronouns and Clitics 48

Prepositions 50

III. OPERATIONALIZATIONS OF MORPHOLOGICAL NORMALIZATION AND LEXICAL RICHNESS 54
Operationalizations of Morphological Normalization and Indices 55
Lexical Diversity 55
Lexical Sophistication 57
Word Frequency 57
Bigram Frequency and SOA 58
Bigram Dependencies 61
Morphological Normalizations 61
Raw Words 62
Lemmas with Part of Speech 63
Lemmas with Verbal Information 64
Study One: Lexical Diversity 65
Learner Corpus 65
References Corpus 68
Human Judgements 68
Linguistic Analysis 71
Lexical Diversity 73
Statistical Analysis 74
Results 78
Discussion 83
Study Two: Lexical Sophistication 87
Lexical Sophistication 87

Statistical Analysis …. 88
Results …. 89
Word Frequency …. 90
Bigram SOA …. 95
Bigram Dependencies …. 102
Discussion …. 111
Word Frequency …. 111
Bigram SOA …. 114
Bigram Dependencies …. 119
IV. LEXICAL RICHNESS AS PREDICTOR OF L2 SPANISH WRITING QUALITY: A MULTIVARIATE APPROACH 125
Methods 128
Final Models 128
Lexical Diversity 129
Word Frequency 129
Bigram SOA 130
Bigram Dependencies 131
Results 133
Discussion 137
Vocabulary Scores 139
Language Use Scores 147
V. CONCLUSION AND OUTLOOK 151
Operationalization Of Morphological Normalization and Lexical Richness 151

Summary of Findings 152
Research questions 1a-1d 152
Research questions 2a-2d 154
Lexical Richness as Predictor of L2 Spanish Writing Quality 157
Summary of Findings 158
Research question 3 158
Contributions 160
Implications 162
Second Language Acquisition 162
Second Language Spanish Writing Assessment 162
Second Language Spanish Pedagogy 163
Limitations 163
Outlook 165
APPENDICES 166
A. ESSAY RATING RUBRIC 166
REFERENCES CITED 172

LIST OF FIGURES

Figure Page

1. List of verbal times and unique conjugations in Spanish and English 46
2. Examples of short text showing the use of subject pronouns 47
3. Rater severity from ICC 77

LIST OF TABLES

Table Page

1. Descriptive Statistics for Descriptive Essays on COWSL2H before sampling 66
2. Descriptive Statistics for Descriptive Essays on COWSL2H after sampling 67
3. Summary of Many-Facet Rasch Model for Calibration Data 70
4. Distribution of essays per rater 71
5. Lemmatization of sample sentence 73
6. Inter-reliability results from Intraclass correlation coefficients 76
7. Rater severity scores from ICC 76
8. MTLD regression models comparison for total scores 79
9. MTLD Regression models comparison for vocabulary 80
10. MTLD Regression models comparison for language use 80
11. MATTR regression models comparison for total scores 81
12. MATTR Regression models comparison for Vocabulary 82
13. MATTR Regression models comparison for Language Use 82
14. Lemma + TMPN comparison between MTLD and MATTR 83
15. All words frequency regression models comparison for total scores 91
16. All words frequency regression models comparison for Vocabulary 91
17. All words frequency regression models comparison for Language Use 92
18. Content words frequency regression models comparison for total scores 93
19. Content words frequency regression models comparison for Vocabulary 93
20. Content words frequency regression models comparison for Language Use 94

21. Lemma Operationalization comparison between All Words and Content Words. 94

22. T'-Score Regression models comparison for total scores 95

23. T'-Score Regression models comparison for Vocabulary..... 96

24. T'-Score Regression models comparison for Language Use 96

25. MI Regression models comparison for total scores..... 97

26. MI Regression models comparison for total Vocabulary 97

27. MI Regression models comparison for total Language Use..... 98

28. Delta P1 Regression models comparison for total scores..... 98

29. Delta P1 Regression models comparison for Vocabulary 99

30. Delta P1 Regression models comparison for Language Use 99

31. Delta P2 Regression models comparison for total scores..... 100

32. Delta P2 Regression models comparison for Vocabulary 100

33. Delta P2 Regression models comparison for Language Use..... 101

34. Raw word Operationalization comparison between all SOAs..... 101

35. T-scores Object dependencies summary..... 104

36. MI scores Object dependencies summary..... 104

37. DeltaP1 Object dependencies summary..... 105

38. DeltaP2 Object dependencies summary..... 105

39. T'-scores Adv dependencies summary..... 106

40. MI scores Adv dependencies summary 106

41. DeltaP1 scores Adv dependencies summary 107

42. DeltaP2 scores Adv dependencies summary 107

43. T'-scores Adj dependencies summary..... 109

44. MI scores Adj dependencies summary 109

45. DeltaP1 scores Adj dependencies summary 110

46. DeltaP2 scores Adj dependencies summary 110

47. Summary of predictors for final model..... 132

48. Summary of predictors for final model..... 133

49. Variance Inflation Factors (VIF) for each predictor in the full model 134

50. Optimal linear model for Vocabulary scores 136

51. Optimal linear model for Language Use scores..... 137

52. Example of an Essay from a Beginner learner..... 141

53. Example of an Essay from an Intermediate learner 143

54. Example of an Essay from an Advanced learner 145

1 INTRODUCTION

Research has demonstrated the important role that productive lexical proficiency plays in successful second language (L2) communication (e.g., Nation, 2001). Consequently, several studies have examined the relationships between lexical use, language proficiency, and language development. Early research in this area (e.g., Engber, 1995; Laufer & Nation, 1995; Meara & Bell, 2001) showed a clear connection between two aspects of lexical use (lexical diversity and reference corpus word frequency) and writing proficiency. More recent research (e.g., Crossley & McNamara, 2012; Kim et al., 2018; Kyle et al., 2018; Kyle & Crossley, 2015) has indicated that word frequency is one of several word-level characteristics that are correlated with productive lexical proficiency. Additional research (Bestgen & Granger 2014; Eguchi & Kyle, 2020; Garner et al., 2019; Kyle & Eguchi, 2021) has also shown that features of lexicogrammatical use (e.g., the strength of association between contiguous words, and word combinations in particular grammatical relationships) are important predictors of L2 writing competence and quality. Importantly, this research has indicated that productive lexical proficiency scores are most accurately modeled when multiple indices of lexical and lexicogrammatical use are considered. Furthermore, research has demonstrated that a combination of different indices of lexical and lexicogrammatical use (e.g., lexical diversity, bigram dependencies, lexical sophistication, strongly associated collocations, grammatical relationships) are significant predictors of second language (L2) writing quality in the form of human judgments (Crossley et al., 2016; Garner et al., 2019; Guo et al., 2013; Jung et al., 2019; Kim et al., 2018; Lee et al., 2021). Overall, this body of research has indicated that more proficient writers tend to use a wider variety of words, less frequent words, words that are less concrete, words that are more contextually distinct, word combinations that are more strongly

associated, and more strongly associated words in particular grammatical dependencies (inter alia). The evidence provided by these studies is relevant since modeling the relationship between indices of lexicogrammatical features and human judgments of L2 writing can provide tools for the automatic evaluation of L2 writing at a large scale (e.g., TOEFL iBT).

One limitation of the existing research is that it has focused primarily on L2 English learners, a tendency likely influenced by the widespread availability of L2 learner corpora and text analysis tools specifically tailored for English. This is particularly evident regarding multivariate approaches to analyzing L2 writing features that may predict L2 writing proficiency in the form or human judgments. One commonly taught L2 that is understudied in this area is Spanish. Spanish speakers and learners continue to grow in the US (Fernandez Vítores, 2023), and as 2023 over a third of the world's L2 Spanish learners are in the US. Consequently, it is imperative to have a better understanding of linguistic features that are characteristic of productive proficiency in Spanish as an L2. The continued need for developing professional skills in L2 Spanish (Fernandez Vítores, 2023) prompts the investigation of academic writing development as one of the key components of successful professional communication. In order to gain a better understanding of which features of lexicogrammatical use are most relevant to predicting L2 Spanish writing quality, in-depth multivariate approaches to analyzing lexicogrammatical use in L2 Spanish writing are needed.

Although the findings in English as an L2 might indeed apply to other languages, there's also a possibility that different target languages or L1-L2 combinations may yield different results. This is possible due to the variability in the way lexicogrammatical features operate in different languages, namely, inflectional and derivational systems, word order, clitic pronoun

systems, and agreements among words classes. For instance, emerging research suggests that while the same lexicogrammatical features predicting L2 writing quality in English may serve as reliable predictors for other L2s, some indices may be more relevant for a given language than others (e.g., Hao et al., 2023). This may also be the case with L2 Spanish, which differs substantially from English in terms of word order, pronoun clitic type and use, and inflectional systems (Whitley, 2002).

The morphosyntactic rules that govern Spanish differ substantially from those governing English since Spanish has more complex nominal and verbal inflectional systems (Montrul, 2004). For instance, nouns, past participles, adjectives, interrogative, relative, and personal pronouns are inflected for both number and gender. Moreover, the use of these pronouns varies in position, sometimes even used as particles that can be attached to verbs; such is the difference between English and Spanish clitic pronoun use. While English only allows clitic pronouns after verbs (enclisis), Spanish allows clitic pronouns before verbs (proclisis) in simple structures, while complex structures allow for either enclisis or proclisis. However, the difference between Spanish and English is most noticeable when examining verbs. Spanish verbs are inflected for not only tense (e.g., future, past) and person (e.g., first, second, plural, singular), but also mood (indicative, subjunctive, and imperative) and aspect (perfective and imperfective). While English inflection for verbs exists, the system is less complex. For example, where English inflects verbs for person, this occurs based on number only (I, you, they, we vs. he, she, it), while in Spanish, each person (first, second, and third; singular and plural) has its own distinct verb ending, varying according to the conjugation pattern of the verb, the tense, the mood, and in some instances the degree of formality. This is a relevant aspect of Spanish language because the use of particular combinations of time, aspect, mood, and person may be an indicator of proficiency

(Kanwit, 2019). Therefore, these differences between the English and Spanish verb systems have the potential to produce different results when investigating lexicogrammatical features of L2 written Spanish and their relationship with productive language development.

Although a few studies have begun to investigate some aspects of productive lexical proficiency in L2 Spanish (e.g., Castañeda-Jiménez & Jarvis 2013; Consolini & Kyle, 2024; Diez-Ortega & Kyle, 2024; Tracy-Ventura, 2017) and their relationship with writing quality (e.g., de la Torre García et al., 2021), a multivariate approach to lexical use has not yet been taken with graded written L2 Spanish samples. Therefore, this study continues to build on previous research that uses multivariate approaches to modeling lexical and lexicogrammatical use but extends this research to explore indices of lexicogrammatical features in Spanish as an L2 and their predictive power in relation to human judgments of L2 Spanish writing. Therefore, this dissertation will focus on the analysis of indices of lexicogrammatical use in L2 writing through the examination of lexical diversity, lexical sophistication, and dependency bigrams. Additionally, these indices will also be analyzed in relation to measurements of L2 writing quality in the form of human judgments.

In order to guide this research study, the following research questions have been established:

1- To what extent are different operationalizations of morphological normalization when calculating lexical diversity more or less predictive of L2 Spanish writing quality in the form of human judgments?

a. To what extent is lexical diversity predictive of L2 Spanish writing quality in the form of human judgments when tokens are morphologically normalized as raw words?
b. To what extent is lexical diversity predictive of L2 Spanish writing quality in the form of human judgments when tokens are morphologically normalized as lemmas with parts of speech information?
c. To what extent is lexical diversity predictive of L2 Spanish writing quality in the form of human judgments when all tokens are morphologically normalized as lemmas, except for mood and tense information in verbs?
d. To what extent is lexical diversity predictive of L2 Spanish writing quality in the form of human judgments when all tokens are morphologically normalized as lemmas, except for mood, tense, person, and number information in verbs?

2- To what extent are different operationalizations of morphological normalization when calculating lexical sophistication more or less predictive of L2 Spanish writing quality in the form of human judgments?
    a. To what extent is lexical sophistication predictive of L2 Spanish writing quality in the form of human judgments when tokens are morphologically normalized as raw words?
    b. To what extent is lexical sophistication predictive of L2 Spanish writing quality in the form of human judgments when tokens are morphologically normalized as lemmas with parts of speech information?

c. To what extent is lexical sophistication predictive of L2 Spanish writing quality in the form of human judgments when all tokens are morphologically normalized as lemmas, except for mood and tense information in verbs?

d. To what extent is lexical sophistication predictive of L2 Spanish writing quality in the form of human judgments when all tokens are morphologically normalized as lemmas, except for mood, tense, person, and number information in verbs?

3- To what extent are features of written lexical and lexicogrammatical use in the form of lexical sophistication and lexical diversity predictive of L2 Spanish writing quality in the form of human judgments?

This dissertation is organized as follows: First, there is a review of the relevant literature that contextualizes the rationale for the studies in this dissertation. Secondly, two studies are conducted to address the first two sets of research questions in order to examine the most suitable operationalizations for calculating indices of lexical richness. Next, the third research question is addressed in a separate study that is built from the previous two studies. Finally, in the last chapter, the results of all the studies are summarized guided by the research questions and implications are discussed.

Chapter 2 involves a discussion of the literature regarding two main areas: first, previous research in second language writing and multivariate approaches to modeling lexical richness and its relationship with writing quality. Second, operationalizations of morphological normalization and their relevant evaluation in the context of Spanish as a second language, by

contemplating the differences with English which may cause different operationalizations to have an impact on indices of lexical richness.

Chapter 3 describes two studies that evaluate four different operationalizations of morphological normalization in Spanish, in relation to 20 different indices of lexical richness, which include lexical diversity indices, and lexical sophistication indices in the form of word frequency, bigram strength of association measures, and bigram dependencies. This chapter addressed the first two sets of research questions in this dissertation.

Chapter 4 addresses the third research question by evaluating multivariate models comprised of various indices of lexical richness to predict human judgment scores of writing quality in L2 Spanish in essays from the COWSL2H corpus. Two models are created for two different scoring categories based on the results of the previous two studies on operationalizations of morphological normalization, and then the most optimal model for each is found and their results examined and discussed.

Chapter 5 consists of a summary of the results of the studies in the previous chapters. In this chapter, the implications of the findings of this dissertation are discussed in the context of Spanish second language writing, second language assessment, and second language acquisition.

2 SECOND LANGUAGE WRITING AND FEATURES OF LEXICOGRAMMATICAL USE

This chapter addresses key concepts in the field of second language writing assessment from a corpora lens, which involves two main concepts: second language writing quality in the form of human judgments, and second language assessment in the form of mathematical indices of lexical and lexicogrammatical use. Additionally, this chapter addresses the gap in this body of research by discussing the application of indices of lexicogrammatical use to second-language Spanish writing, and for this the key lexicogrammatical features that differentiate Spanish from English are discussed in detail. Therefore, the chapter will be divided into three sections. The first section reviews literature regarding second language writing research, which will discuss findings originated primarily from research investigating writing performance by English as a Second or Foreign Language (ESL/EFL) learners. The second section will discuss in detail the different indices that measure lexical and lexicogrammatical features of second-language writing. The third section will discuss how this dissertation builds on previous research in ESL/EFL to examine the relationship between the aforementioned indices and writing quality scores of human judgments in the context of Spanish as a second or foreign language. This last section will include the lexicogrammatical features that differentiate Spanish from English to clarify and exemplify the current gap in the field of second language writing assessment.

Second Language Writing Research

In academic settings, second language (L2) writing skills are important to develop a strong competency to communicate in the L2 (Kellogg & Raulerson, 2007). Research has shown that L2 writing skills are critical in global contexts for successful cross-cultural communication

(Hyland, 2003) and can affect academic and professional opportunities and outcomes (Leki et al., 2008; Manchón, 2011). To better understand the development of second language writing, it is necessary to examine the elements that comprise it, and second language lexical and lexicogrammatical knowledge are the building blocks of successful L2 writing. Particularly, productive lexical proficiency plays an important role in successful L2 communication and writing (e.g., Nation, 200; Kellogg & Raulerson, 2007). Much research has been carried out in the areas of L2 writing development (e.g., Silva, 1993; Laufer,1994; Kobayashi & Rinnert, 2012), and assessment of second language writing (e.g., Polio, 1997; Biber & Gray, 2010; Plakans et al., 2019) in order to better understand the stages of development across time and proficiency levels, and to develop assessment criteria and methodologies that accurately reflect a learner's developmental stage in their second language writing journey.

Particularly, attention has been paid to the acquisition and productive use of lexical and lexicogrammatical items in academic writing contexts. Early research was focused predominantly on lexical frequency measures (e.g., Laufer & Nation, 1995), and measures of lexical diversity (e.g., Engber, 1995). More recently, research has focused on a wider range of features that relate to lexical and lexicogrammatical use, such as strength of association measures between sequences of given numbers of adjacent words (i.e., n-grams), different indices of lexical diversity, average word and n-gram frequencies by means of a reference corpus, and collocations dependency relationships (e.g., Granger & Bestgen 2014; Jarvis, 2013; Kim et al., 2018; Kyle & Eguchi, 2021; Paquot, 2018; Rubin et al., 2021). Additionally, research that incorporates a wider range of indices to examine features of lexical and lexicogrammatical use has done so by implementing multivariate approaches to the analysis of written productions. Applying multivariate approaches to features of second language writing provides a clearer

picture of the different aspects that make up a proficient piece of writing, how they relate to each other, and how they may relate to the human perception of second language writing proficiency.

Accordingly, an important addition to this body of research is the integration of quality writing scores in the form of human judgments. Research in this area has added to the information provided by indices of lexicogrammatical use by integrating the quality scores of trained raters, who evaluate academic essays based on a rubric. These scores are examined in relation to indices of lexicogrammatical features to evaluate if these indices can be effective predictors of human judgment scores. Research on this area has indicated that the combination of diverse indices of lexical and lexicogrammatical use (e.g., grammatical dependencies, lexical diversity, strongly associated collocations, lexical sophistication) are important predictors of second language writing quality in the form of human judgments (Crossley et al., 2016; Garner et al., 2019; Guo et al., 2013; Jung et al., 2019; Kim et al., 2018; Lee et al., 2021).

Furthermore, these research studies have indicated that overall, more proficient writers tend to use more strongly associated words, a wider range of different words, less frequent words and n-grams, and more strongly associated words in particular dependency relationships, such as verb–adverb, and verb–direct-object relationships. Most importantly, these studies have highlighted the importance of combining different features of lexical and lexicogrammatical use to create multivariate models that can provide clearer insight into the linguistic features of successful second language writing, and that can be used to model and predict the writing quality of second language writers.

Features of Lexicogrammatical Use

Different terms have been used to refer to features of lexical and lexicogrammatical use. For instance, Bulté and Housen (2012) have referred to these features as 'lexical complexity'. However, possibly the most common term is 'lexical richness' (e.g., Yule, 1944; Laufer & Nation, 1995; Read, 2000). Lexical richness can be better understood when separated into three different but related subconstructs: lexical diversity, lexical sophistication, and lexical density. In this research study, the focus is on lexical diversity and sophistication since early research (Ure, 1971; Shohamy, 1994) suggests that lexical density may be helpful for discriminating between registers or modalities, but it is not as helpful for measuring lexical proficiency.

Word Level

Lexical diversity. Lexical diversity is commonly defined as the variance of the lexical choices made by a writer (Jarvis, 2013; Zenker & Kyle, 2020). More proficient writers presumably hold or have acquired a bigger lexicon, and can therefore produce a wider array of lexical items or employ a broader range of lexical choices than less proficient writers. Indices such as the type-token ratio (TTR) and its subsequent alternation in Guiraud's (1960) root TTR (RTTR) have historically been used to measure this type of textual diversity. However, research has demonstrated that the results from these metrics tend to vary with the length of the text analyzed, complicating their reliability in comparative analysis (e.g., Jarvis, 2002; McCarthy & Jarvis, 2010; Zenker & Kyle, 2020). For this reason, much research has been devoted to developing, validating, and testing indices of lexical diversity that are reliable predictors and independent of text length. For example, Zenker and Kyle (2020) showed that in English essays written by L2 learners, indices such as the measure of textual lexical diversity (MTLD;

McCarthy & Jarvis, 2010) and the moving average TTR (MATTR; Covington & McFall, 2010) were stable in argumentative essays extending from 50 to 200 words, with 50 words being the minimum possible length for a reliable measurement. The scores provided by the MTLD index represent the average length of factors in a text (McCarthy & Jarvis, 2010). Factors in the context of MTLD refer to segments within a text where the type-token ratio falls below a certain threshold, commonly set at 0.72. To ensure accuracy and validity, each factor includes a minimum of 10 words. Factors are outlined in a non-overlapping manner; once a boundary is established, subsequent factors are identified following the same criteria until the text ends. This process is performed from the beginning to the end of the text and then repeated in reverse, from end to beginning. The MTLD score is then calculated as the average length of all identified factors, providing insight into the lexical variety and repetition within the text. Since the index looks at 10-word factors at a time, it provides an assessment of diversity in short portions, instead of a global look at the textual diversity. Similarly, when we look at a text, we may assess it as repetitive if the repetitions happen close to one another, but not if they repeat globally in the text but with fair distance between one another. Therefore, this method can provide a more nuanced view of how words are diversified or repeated across different sections of the text, reflecting potential variations in how humans perceive and process language diversity.

Similarly, the scores provided by the MATTR index represent the average type-token ratio (TTR) calculated across overlapping segments, or "windows," of fixed length within a text (Covington & McFall, 2010). Each window includes a predetermined number of words (commonly 50) and slides through the text one word at a time. For every position, a local TTR is calculated by dividing the number of unique word types by the total number of words in that window. These TTR values are then averaged across all possible windows in the text, yielding a

single MATTR score. Unlike global TTR (Richards, 1987), which is highly sensitive to text length, MATTR provides a more stable and reliable estimate of lexical diversity, particularly when working with short to medium-length texts. The sliding-window approach allows for continuous monitoring of lexical variation throughout the text, offering a detailed view of how varied the vocabulary is at different points of the text. Since this method can help see changes in repetition at the local level, MATTR can also align closely with the human perception of lexical richness, where repetition in a few sentences tends to be more noticeable than repetition across distant parts of the text.

Consequently, using both the MTLD and MATTR indices in lexical diversity analysis could provide a more comprehensive and balanced view of variation within learner texts. While MTLD captures broader patterns by identifying how long a learner can sustain lexical diversity before falling below a defined threshold, MATTR offers a finer-grained, localized view of repetition and variation by averaging diversity across continuously shifting windows. Each measure is sensitive to different aspects of lexical behavior: MATTR is good at detecting short-range repetition and formulaic language, whereas MTLD highlights longer-term variation and the learner's ability to maintain diverse vocabulary over extended passages. By combining both indices, researchers can identify patterns that may be missed when relying on a single measure. For instance, a learner who uses a wide range of vocabulary overall but relies heavily on repetitive sentence frames may show high MTLD but low MATTR. Together, the two measures offer a more nuanced understanding of how learners deploy vocabulary across different spans of text, aligning more closely with both computational and human judgments of lexical richness.

Furthermore, research has examined the evidence for the validity of MATTR and MTLD (Jarvis, 2017; Kyle et al., 2021). For instance, Kyle et al. (2021), showed that MATTR and MTLD were positively correlated (r = .566 for each) with human judgements scores of lexical diversity in argumentative essays written by L2 English learners. Nevertheless, very little research has been conducted into lexical diversity measures in other languages. This is particularly relevant to the study of second language written assessment, since different L2s may have different morphological patterns that may impact the assessment of proficiency level when analyzed through the lenses of lexical diversity. This is important to consider when it comes to deciding what word forms to analyze when assessing the lexical diversity of a text. For example, most research examining lexical diversity indices in English tend to lemmatize all tokens. This means to simplify the total number of analyzed words in a text, or tokens, to their "dictionary form". For instance, the lemma "run" in English includes all instances of "runs, ran, running". Whereas the lemma for "hablar" in Spanish would include all instances of "hablo, hablas, habla, hablan, hablamos, habláis, hablé, hablaste, hablaron, habló, hablasteis, hablaré, hablaremos, hablarán, hablaréis", and many more. It is logical, therefore, to reduce tokens to lemmas in English, since the inflectional system most likely wouldn't provide a significant difference in lexical diversity measures. This is due to the fact that most, if not all, forms of "run" (ran, runs, running) are learned relatively early in the development of second language writers of English, and therefore may add little information to the development stage of the learner or their overall proficiency level. However, in Spanish, the complexity of different inflectional systems is greater, presenting a bigger challenge for English speakers that are learning Spanish. Moreover, since the inflectional systems are more complex, the exposure to them and their subsequent acquisition are more gradual over time and may be therefore indicators of a learner's

developmental stage in the target language. This in turn, reflects on their proficiency level, since the fluent use of these inflectional systems is necessary for successful communication (Kanwit, 2019).

However, it is necessary to emphasize that measurements of lexical diversity are specific to textual range and variety. Consequently, they do not provide a clear indication of the sophistication of the words used since they are not sensitive to the features that define the sophistication of the words used. In this sense, a phrase such as 'a hot summer day' would receive the same diversity score as 'quite an intricate argument'. For this reason, research has suggested (e.g., Jarvis, 2013) that a more accurate and appropriate approach to measuring lexical knowledge in second language learners is through a multivariate approach; that is to say, a combination of several measures where each examines a different feature of lexical richness.

Lexical Sophistication. Lexical sophistication is commonly defined as the use of advanced words in a text (Read, 2000). Typically, writers who have reached a higher level of proficiency in their L2 tend to use what are considered more difficult words. Whether a word is thought to be sophisticated depends on who is evaluating the word. From the learners' perspective, lexically sophisticated words include those with characteristics that make them harder to learn (e.g., length, spelling, lower frequency of use, obscurity of meaning). However, from the readers' perspective, lexical sophistication is conceptualized as words with features that are perceived as advanced (e.g., nuanced meanings, subtle connotations, and proper contextual use). Traditionally, there have been two main approaches to measuring lexical sophistication. The earliest of these approaches, still in some use today, is the lexical frequency profile (LPF; Laufer, 1994). In this approach, words from a reference corpus are grouped into different

frequency bands (i.e., 1,000 first most frequent words, 1,000 second most frequent words, etc.) and then a series of scores are provided based on the number and percentage of words used in the L2 text that belong in each band. Research studies that have used this approach have shown that more proficient L2 writers tend to use more words from the lower frequency bands, and fewer words from the bands with the most frequent words, such as the first and second 1,000 most frequent word bands (Laufer, 1994; Laufer & Nation, 1995; Crossley et al., 2013). Nonetheless, there are some limitations to this approach. Most notably, the analysis of the frequency of words is large-grained. This means that the difference between the most and least frequent words is the same as the difference between the 1,000th and the 1,001st words. This makes it very difficult to clearly see small advances in lexical use, while also providing related but different scores that are difficult to process statistically.

However, other approaches of calculating frequency scores are more fine-grained. For instance, calculating the average frequency score for each word. For this approach, measures are calculated individually for each word used in the L2 essay instead of grouping a set number of words into different frequency bands. This process results in an average frequency score for each word, which in turn can provide an average frequency score for each essay. Consequently, the score reflects more detailed information about the small advances in lexical development and the scores are more manageable statistically, which allows for more complex multivariate models. Research that has used this approach and examined L2 written text has found that, on average, more proficient writers tend to use lower-frequency words (e.g., Guo et al. 2013; Kyle & Crossley, 2015; Kyle et al., 2018).

Similarly to the case of lexical sophistication, little research has applied this concept on languages other than English. The few studies that have researched lexical sophistication in L2 Spanish and have found similar correlations to the ones in L2 English studies. These studies focusing only or mostly on L2 Spanish lexical sophistication have examined this index in relation to other variables, such as task-type and passive vs active vocabulary learning (Berton & Sánchez, 2023), after studying abroad (Tracy-Ventura, 2017), or in the context of immersion or dual language schools (Schur & Rubio, 2021; Waldvogel, 2014). Additionally, most of this research focuses on oral production (Berton & Sánchez, 2023; Waldvogel, 2014) or on a combination of both written and spoken texts (Tracy-Ventura, 2017). More recent research has included lexical sophistication in multivariate approaches to L2 Spanish lexical development and its predictive potential of language proficiency (Consolini & Kyle, 2024; Díez-Ortega and Kyle, 2024), demonstrating that multivariate approaches are more effective as predictive models of proficiency than isolated measures of lexical richness. However, when analyzing lexical sophistication in these studies, little to no attention is paid to the part of speech in detail, and tokens are almost always lemmatized, losing some of the morphological richness of the language. This is another key component when analyzing second language writing in a variety of contexts, since what makes a word more advanced for a learner of reader may be different from language to language. For instance, it could be argued that the features that describe lexical sophistication (e.g., length, spelling, lower frequency of use, nuanced meanings, subtle connotations) can be more salient in the verbal conjugation system where morphology not only changes within and between tenses to mark gender, number and person, but also withing and between markers of verbal mood to convey nuanced meanings and subtle connotations (Montrul, 2004).

Beyond Word Level

While investigating lexical sophistication through the application of measurements of lexical frequency has proven useful (e.g., Crossley et al. 2016; Kyle et al. 2018), a key feature of productive lexical proficiency is the use of words in their appropriate lexicogrammatical contexts (e.g., Sinclair 1991; Nation 2001). In order to measure this, Delta P (Allan, 1980) is used to calculate the strength of association (SOA) between contiguous by quantifying the degree to which the occurrence of one word is predictive of another (Ellis, 2006). Additionally, a reference corpus is used to create a list of word combinations and their frequency of use when used together to explore which combinations of words are more or less frequently used in the target language. Research that utilizes this technique, along with other measurements in a multivariate approach, has shown that more proficient English L2 writers, on average, use more frequent and more strongly associated bigrams (two-word combinations) and trigrams (three-word combinations) (Bestgen & Granger, 2014; Kim et al., 2018; Kyle et al., 2018; Garner et al., 2019). Furthermore, these research studies showed that the frequency and SOA measures are strong predictors of language proficiency and development.

Bigrams and trigrams. Bigrams are defined as two contiguous words in a text. Some word combinations are very frequent (e.g., the house, a cat, that day), and some are strongly associated (e.g., views of, due to, based on). This means that while some combinations of words are more frequent, or common, some are not only frequent but happen mostly only in that order or that combination. Such is the case of fixed expressions or strong collocations (e.g. give up, look forward, carry out). While there are many frequent combinations of two words in the form of, for example, article + noun (e.g., a book, the car, an idea), or noun + adjective (e.g., cold

weather, great opportunity, high risk), there are also some combinations of words that are very strongly associated, meaning that they almost always happen together (e.g., fast food, pay attention). This is more clearly exemplified by trigrams, or three contiguous words in a text, for instance, *in spite of,* or *as well as*.

The analysis of the frequency and strength of association between these n-grams provides a clearer insight into the proficiency of learners since they reflect the productive use of more advanced word combinations in the target language. Research in this area has demonstrated that more proficient writers tend to use more frequent and more strongly associated combinations of words, and that measurements of bigram frequency of use are more predictive of proficiency than isolated tokens (e.g., Kyle & Crossley, 2015; Garner et al., 2019). Conversely, less frequent combinations of words (e.g., unexpected happiness, cautious excitement, subtle influencing) and less strongly associated words (e.g., consider options, reflect in depth, pursue the opportunities) reflect a less developed understanding of the lexicogrammatical rules of the target language. Consequently, much research has been devoted to refining the methodology to calculate, as accurately as possible, the strength of association between contiguous words.

The most commonly used and studied measures of strength of association are T-score, MI score, and Delta P. T-scores (Allan, 1980) are used to measure the SOA between bigrams or trigrams based on their frequency of co-occurrence. To calculate T-scores, the following formula is used:

$$T = \frac{(O - E)}{\sqrt{O}}$$

where O is equivalent to how many times the words occur together in a given corpus, and E is equivalent to the expected frequency of the bigram if words were independent, which is give by:

$$E = \frac{(f_1 \ X \ f_2)}{N}$$

where f_1 is the frequency of the first word in the bigram, f_2 is the frequency of the second word in the bigram, and N is the total number of words in the corpus.

Consequently, a higher T-score indicates that the words that comprise the bigram have a stronger strength of association, while the lower the T-score goes it may indicate that the co-occurrence could be due to randomness rather than linguistic relationship. This measurement commonly gives more importance to frequent bigrams, making it most useful for identifying common collocations.

Alongside T-scores, it has been common to use MI scores to provide a more complete picture of strength of association between bigrams or trigrams. MI (Mutual Information; Church & Hanks, 1989) evaluates how much more often a pair of words occur together than it would be probable to occur. The key components to calculate MI are the frequency and probability of each word in a bigram in a given corpus, as illustrated in this formula:

$$MI = \ log_2 \left(\frac{P \ (W_1, W_2)}{P(W_1), P(W_2)}\right)$$

where $P\ (W_1, W_2)$ is the probability of the bigram occurring, and $P\ (W_1)$ and $P\ (W_2)$ are the probability of the first and second words in the bigram occurring, respectively. To reach probability estimates, the following formulas are used:

$$P\ (W_1, W_2) = \frac{O}{N}\ ; P(W_1) = \frac{f_1}{N}\ ; P(W_2) = \frac{f_2}{N}$$

where O is the number of times the bigram appeared in the given corpus, f_1 is the frequency of the first word, f_2 is the frequency of the second word and N is the total number of words in the corpus.

When MI scores are higher, they indicate a stronger association between the words in a bigram, while low MI scores closer to 0 indicate a lower association closer to random co-occurrence. Research (e.g., Evert, 2005) has shown that MI tends to overestimate rare bigrams due to the fact that even low-frequency pairs can get a high MI score if the words appear separately on rare occasions. Consequently, attempts have been made to use T-scores and MI scores together to get a better understanding of n-grams' strength of association.

While both MI and T-scores measure lexical strength of association, MI is prone to assign higher scores to less expected word pairings that co-occur frequently relative to their individual occurrences (e.g., asymptotic analysis); in contrast, T-score is likely to assign higher scores to high-frequency word pairings with strong associations relative to the number of occurrences (e.g., good morning) (Evert, 2005). Therefore, researchers started implementing Delta P instead, or complementarily. Unlike T-score or Mutual Information (MI), which relies on frequency-based statistical significance, Delta P (ΔP; Allan, 1980; Ellis, 2006) assesses the degree to which the occurrence of one word predicts the occurrence of another. There are two ways to calculate ΔP, depending on which word is considered the predictor:

ΔP (Word 1 → Word 2) measures how the occurrence of the first word increases or decreases the probability of the second word occurring at all, using the following formula:

$$P_2 = P(W_1) - P(W_2|\neg W_1)$$

and ΔP (Word 2 → Word 1) measures how much the occurrence of the second word increases or decreases the probability of the first word occurring at all, using a similar formula:

$$P_2 = P(W_2) - P(W_1|\neg W_2)$$

where $P(W_2|\ W_1)$ is the probability of W_2 occurring given W_1, and $P(W_2|\ \neg W_1)$ is the probability of W_2 occurring when W_1 is absent. Similarly, $P(W_1|\ W_2)$ is the probability of W_1 occurring given W_2, and $P(W_1|\ \neg W_2)$ is the probability of W_1 occurring when W_2 is absent. To reach probability estimates, the following formulas are used:

$$P(W_1) = \frac{O}{f_1}\ ;\ P(\neg W_1) = \frac{f_1 - O}{N - f_1}\ ;\ \text{and}$$

$$P(W_2) = \frac{O}{f_2}\ ;\ P(\neg W_2) = \frac{f_2 - O}{N - f_2}$$

where O is the observed co-occurrence frequency of the bigram, f_1 is the frequency of the first word, f_2 is the frequency of the second word, and N is the total number of bigrams in the corpus.

The range of the results from ΔP range between -1 and 1. A higher ΔP score indicates a stronger predictive relationship between the two words, while a negative score would indicate an inverse relationship where one word appears less often given the other word.

ΔP is often considered a better measure of bigram strength compared to T-score and Mutual Information (MI) because it captures the directionality of association and provides a more intuitive interpretation of predictability between words. Unlike T-score and MI, which measure general co-occurrence strength, ΔP quantifies how well one word predicts the other.

This is especially useful in linguistic relationships where direction matters, such as verb-object relationships (e.g., "eat → apple" vs. "apple → eat"), cause-effect expressions, or fixed collocations (e.g., "take a break" vs. "break a take"). For example, considering the bigram “coffee shop”, T-score and MI would give the same value to both word orders, while ΔP would likely be higher for "coffee → shop" than for "shop → coffee", reflecting the natural linguistic tendency. Moreover, T-score favors high-frequency bigrams, so it often overlooks meaningful but less frequent collocations; and MI favors rare bigrams, which leads to overestimation of uncommon word pairs. Meanwhile ΔP balances both frequent and rare word pairs by considering their relative probabilities in context. For example, the bigram "asymptotic analysis" (low frequency, but strong association) would shield a high MI score, and the bigram "good morning" (very frequent but not rare) would shield a Low MI score even though it's a strong collocation. Instead, ΔP avoids this issue because it focuses on predictability rather than raw frequency or rarity.

Additionally, ΔP provides a better measure of the many asymmetrical linguistic relationships between words. MI and T-scores treat bigrams as interchangeable, but ΔP identifies which word in the pair is the stronger predictor. For example, in the bigram “heavy rain”, the word "heavy" strongly predicts "rain," but "rain" does not necessarily predict "heavy" (it could be "light rain," "torrential rain," etc.). ΔP will show a stronger association for "heavy → rain" than "rain → heavy." This is key when we are looking at writing proficiency in language learners, since L1 and L2 pairs may have different rules for word order, creating chances of negative transfer while acquiring the target language. In this sense, ΔP may align better with how humans process word associations: We don’t think about raw frequency like T-score, and we

don’t compute logarithmic probabilities like MI. Instead, we make predictive connections (if I hear "strongly," I expect "recommend" or “worded” more than “see” or “done”).

Research studies employing a multivariate approach to examining lexical sophistication have found that, on average, more advanced L2 English writers tend to use higher-frequency and more strongly associated bigrams and trigrams (Bestgen & Granger, 2014; Kim et al., 2018; Kyle et al., 2018; Garner et al., 2019). As a result, collocation frequency and strength of association (SOA) measures have been identified as reliable indicators of language proficiency and development. Regarding L2 Spanish writers, little research has been conducted, but emerging studies suggest a similar trend, where more proficient writers tend to use more strongly associated bigrams (Consolini & Kyle, 2024; Díez-Ortega & Kyle, 2024). However, the studies in L2 Spanish lack the component of writing proficiency scores in the form of human judgments, and instead rely on placement testing or self-assessment to predict overall language proficiency.

Bigram dependencies. As measures of bigrams’ strength of association (SOA) provides us with information about the frequency and level of co-occurrence strength, bigram dependencies go beyond these metrics to consider the grammatical relationship between the two words. Examining bigram dependencies in any language can shed light on the way words are used in a language; in turn, this can provide valuable information on the writing development of second language learners and partially helps us predict their proficiency. For example, in English, a verb-object dependency such as “make a decision” is common and grammatically correct, while “do a decision” is not. In Spanish, a noun-adjective such as “casa grande” is common and grammatically correct, while “grande casa” is not. Examining these relationships in

L2 writing can help us identify common L1 transfer errors, measure writing fluency and accuracy, and potentially improve language teaching methodologies.

It is valuable to note that bigrams' strength of association measures also evaluates bigram dependency, but they do so as semantic dependencies or meaning-based relationships between words. For instance, while "big success" and "great success" are both syntactically correct uses of adjective-noun constructions, "great success" has a higher SOA score than "big success". Evaluating semantic dependencies provides information about the level of writing fluency of the learner. For instance, while "make a decision" is a common collocation, "do a decision" is an uncommon verb choice. Or while "strong evidence" is also a common collocation, "heavy evidence" may provide a similar meaning, but it is an odd choice of adjective. Similarly, "fast food" is a very strongly associated colocation, while "quick food" probably conveys the same meaning in context, but it appears awkward or odd to the reader.

Comparatively, a syntactic dependency analysis of "interested to learn a language" vs "interested in learning a language" would show an incorrect use of a preposition-noun dependency. The analysis of bigrams' syntactic dependencies can provide information about the level of writing accuracy in terms of grammar and syntax. For example, "a difficult task" follows English Adjective-Noun order, "a task difficult" doesn't. While "She goes to school" shows standardized academic subject-verb agreement in English, "She go to school" does not (although there are English varieties that do make use of these and similar structures). Similarly, "she depends on her parents" follows English S-V-O pattern with the correct preposition, but "she depends of her parents" follows the same patters but with an incorrect preposition.

In this sense, we could say that semantic dependencies focus on meaning, evaluating how well words collocate with one another in terms of frequency of use, while syntactic dependencies focus on grammar and syntax, evaluating word order, agreement, and prepositions. These are concepts easily applied to Spanish as well, with the key differences that Spanish has a more layered or complex system of agreement, is more flexible with word order (although some order patterns are more frequent or commons that others outside of literary writings), and has an array of prepositions that commonly do not translate one-to-one from English (Montrul, 2004). These are valuable differences to understand and examine when evaluating the writing of English learners of Spanish.

Research using bigram dependencies to evaluate and/or predict levels of proficiency of L2 English writings has shown that certain types of dependency relationships (e.g., noun-adjective and verb-direct object; Paquot, 2019) account for a sizeable amount of the variance in predictive models of proficiency. Other studies have found small correlations between L2 English quality writing scores and dependency bigram SOA measures of verb–adverb, noun–adjective, verb–direct object dependencies (e.g., Kyle & Eguchi, 2021). Studies in other L2s (e.g., L2 Dutch; Rubin et al., 2021) examined the predictive power of the use of dependency bigrams and found certain types of dependency relationships (noun–adjective, verb–adverb, and verb–direct-object) to be predictive of pass/fail assessments of L2 Dutch writing. However, this examination of dependency bigram types and their predictive relationship with L2 writing proficiency has not been explored in Spanish, where the pattern of use of collocations, in addition to word order and agreement rules, differs significantly from other Germanic languages such as English and Dutch.

Spanish Second Language Writing Research

The research discussed up to this point indicates that L2 writers with higher proficiency tend to use a wider variety of words, less frequent words, words that are more contextually distinct, word combinations that are more strongly associated; and it suggests that certain dependency relationships between parts of speech in frequently used bigrams can be predictive factors of writing proficiency. Additionally, this body of research has found a link between indices of lexicogrammatical use and scores of human judgments, suggesting that these indices can be used to create statistical models that assist in predicting quality scores of human judgments when it comes to second language writing assessment. In turn, these models can be used for not only gaining a clearer understanding of proficiency development but for creating tools to assist in the automatic assessment of certain features of second language writing.

To date, however, the majority of the research investigating lexical and lexicogrammatical features in the context of successful second language writing have done so focusing primarily on English as a second or foreign language (ESL, EFL). The reason for this is most likely due to the large amounts of user data from both English expert speakers and English learners, which allow for the creation of massive corpora inventories. These inventories, along with tools designed specifically for the analysis of English language, allow for the study of lexical and lexicogrammatical features at a large scale, which in turn can provide insightful linguistic information that can be used to better teach, learn, and assess second language writing in English. The limitations of existing research focusing primarily on L2 English learners are most evident regarding the use of multivariate approaches to model the relationship between the diverse indices of lexicogrammatical use and scores of human judgments.

A commonly unexplored L2 in this subfield is Spanish. The number of Spanish speakers and learners in the United States continues to grow (Fernandez Vítores, 2024), and as of 2024, over one-third of the world's L2 Spanish learners are in the US. This highlights the need for a deeper understanding of the linguistic characteristics that define productive proficiency in Spanish as a second language. Moreover, the ongoing demand for professional skills in L2 Spanish (Fernandez Vítores, 2023) emphasizes the importance of studying academic writing as a key element in effective professional communication. To better understand which lexicogrammatical features contribute most significantly to writing quality in L2 Spanish, in-depth multivariate analyses are needed to investigate how these features are used in L2 Spanish writing. While findings from research on English as an L2 may provide insights into other languages, it's also likely that different target languages or specific L1-L2 combinations will yield distinct outcomes. This is because the lexicogrammatical features of each language vary considerably in aspects like inflectional and derivational systems, word order, clitic pronoun systems, and grammatical agreement. For example, emerging research indicates that while certain lexicogrammatical features that predict writing quality in L2 English may be applicable across different languages, some features are more pertinent to a specific language (e.g., Hao et al., 2023). This may also apply to L2 Spanish, which differs significantly from English in terms of word order, the type and use of clitic pronouns, and the complexity of its inflectional system (Montrul, 2004; Whitley, 2002).

Lexicogrammatical Differences Between English and Spanish

The morphosyntactic rules governing Spanish differ significantly from those of English, largely due to the more intricate inflectional systems in Spanish (Montrul, 2004). In Spanish,

elements like nouns, adjectives, past participles, interrogative, relative, and personal pronouns are marked for both number and gender. Additionally, the placement of pronouns varies in Spanish, and they can sometimes function as particles attached to verbs or other words, leading to notable differences from English in the use of clitic pronouns. In English, clitic pronouns exclusively occur as contracted auxiliary verbs, negations, or pronouns in older forms of the language (enclisis; e.g., contractions like *I'm, aren't*, or possession such as *Mary's*); whereas Spanish allows them both after verbs (enclisis; e.g., Díme*lo*, Tóma*la*, Escúcha*me*.), but also before verbs (proclisis; e.g., *Me* vio ayer, *Te* quiero, *Lo* compré). Therefore, the pronominal system in Spanish is not only larger than the English one when accounting for number, gender, and person (*me, te, lo, la, le, nos, os, los, las, les, se*), but it also has an added layer of complexity in the form of word order (enclisis vs. proclisis) and in the modification of verbs. This is relevant for second language learners since they play a key role in the correct representation of verb-argument structures, which in turns contributes to the creation of meaning (Goldberg, 2006) and therefore successful communication. Consequently, it is important to understand the challenges second language learners are faced with when learning Spanish, particularly when it comes to each language's verb system where the most pronounced differences between Spanish and English can be seen.

Spanish verbs are inflected to express tense (e.g., past, future), person (e.g., first, second, third), mood (e.g., indicative, subjunctive, imperative), and aspect (e.g., perfective, imperfective). While English verbs are also inflected to encode grammatical information, the morphological system that encodes meaning in one given verb is far less complex. Although English exhibits complexity in the use of auxiliaries that encode tense information, person-related inflection typically depends only on number (e.g., *I, you, we* vs. *he, she, it*), whereas

Spanish verb forms vary across all combinations of person (first, second, third; singular and plural), depending on the conjugation pattern for a given set of verbs (e.g., verbs ending in *-ar* vs. *-er* and *-ir*), tense, mood, and sometimes the degree of formality. This complexity in Spanish is a key feature that can signal linguistic proficiency as mastery of the combinations of tense, aspect, mood, and person can contribute to the construction of precise and/or layered meaning, demonstrating advanced language skills (Kanwit, 2019). These differences in verb inflection between Spanish and English is not only a challenge for learners, but it can also lead to distinct outcomes when assessing the lexicogrammatical features of L2 written Spanish and their relationship with language development. In the following section, this becomes clearer.

Inflectional Systems. Inflectional systems comprise the rules of a language by which different grammatical categories such as gender, number, tense, mood, and aspect are expressed. In Spanish, this system is more complex than in English, and it plays a key role in meaning construction and sentence structure. There are two types of inflectional systems in Spanish: declension and conjugation. In this section, these systems are first described along with examples, and compared to the English inflectional system to underscore the key differences that could pose a challenge for English learners of Spanish, which could potentially be a factor when automatically assessing features of lexicogrammatical use.

***Declension*.** Declension refers to the inflectional changes that modify nouns, adjectives, and pronouns in order to indicate number, gender, and grammatical case. In Spanish, only number and gender are marked with declension, while grammatical relationships like case are established through the use of prepositions and word order. Most notably, Spanish nouns and adjectives are marked for number (plural or singular) and for gender (feminine or masculine).

For example, while in English we have *cat* and *dog*, which can be marked for number as *cats* and *dogs*, in Spanish these can be marked for both number and gender, as in *gato, gata, gatos, gatas*, and *perro, perra, perros, perras*. Similarly, adjectives in Spanish also change for both gender and number—unlike English, which has no inflectional changes for adjectives. For instance, in English we could see *black cat* and *black cats*, or *small dog* and *small dogs*. However, in Spanish we could see *gato negro, gata negra, gatos negros, gatas negras*, and *perro pequeño, perra pequeñas, perros pequeños, perras pequeñas*.

Conjugation. The declension system is slightly more complex than in English, but not overwhelmingly, only doubling the production of unique tokens in each case. However, the most challenging difference for Spanish learners that speak primarily English is the inflectional system for verbs, or the conjugation system. Conjugation refers to the inflectional changes that verbs undergo in a language to express grammatical categories such as tense, mood, aspect, person, and number. All of these categories are expressed through conjugation in Spanish, while only tense, aspect, and number (only in the present tense) are marked in English. Although the expression of mood can be achieved in English through different grammatical combinations of tense and aspect, there is not a morphological marking that establishes the mood of a verb, such as what we see in Spanish. Finally, when it comes to person, English marks only one form in the present tense (e.g., *they eat, she eats*), but Spanish marks verbs differently for person in all tenses and moods (e.g., *ellos comen, nosotros comemos*).

Conjugations are the verbal inflections with which we express all the grammatical categories that verbs can express in a language: time (or tense), mood, and aspect. Verbal time is the category that grounds events in a particular time, relative to the moment of speaking, such as

past, present, and future. Verbal mood provides information about speakers' attitude in relation to what they are saying. In general, the mood of a verb shows that the phenomena presented or uttered by the speaker can be conceived as known, imaginary, certain, hoped for or desired, among others. Finally, verbal aspect provides information on the internal structure of events; it allows the hearer or reader to know if the events are emergent, repeating, finalized or ongoing.

Overall, there are 72 unique verbal forms in Spanish: 49 in the indicative mood, 21 in the subjunctive mood, and 2 for impersonal verb conjugations. In contrast, there are approximately 10 unique verbal forms in English, depending on whether we count the subjunctive mood or the continuous tenses (in both languages the continuous or progressive tense or aspect is achieved with periphrastic constructions). It is important to highlight that in English we construct most of its times or tenses with a combination of auxiliary verbs and regularly conjugated verbs. This reduces the number of unique conjugated forms for each verbal time. On the other hand, in Spanish there are fewer number of auxiliary verbs and a higher number of uniquely conjugated verbs. Figure 1 exemplifies these differences more clearly.

Language	Mood	Tense	Number of unique forms	Examples
English	Indicative	Present Simple	2	She eats, they eat
		Past Simple	1	She ate, they ate
		Future Simple	1	She will eat, they will eat
		Present Perfect	1	She has eaten, they have eaten
		Past Perfect	1	She had eaten, they had eaten
		Future Perfect	1	She will have eaten, they will have eaten
	Subjunctive	Present	1	I suggest she be on time
		Past	1	If I were you, ….

	Imperative	--	1	Eat! Walk!
Spanish	Indicative	Present	6	Hablo, hablas, habla, hablamos, habláis, hablan
		(Past) Preterit	6	Hablé, hablaste, habló, hablamos, hablastis, hablaron
		(Past) Imperfect	6	Hablaba, hablabas, hablaba, hablabámos, hablabais, hablaban
		Future	6	Hablaré, hablarás, hablará, hablaremos, hablaréis, hablarán
		Conditional	6	Hablaría, hablarías, hablaría, hablaríamos, hablaríais, hablarían
		Present Perfect	6	He hablado, has hablado, ha hablado, hemos hablado, habéis hablado, han hablado
		Past Perfect	5	Había hablado, habías hablado, habíamos hablado, habíais hablado, habían hablado
		Future Perfect	6	Habré hablado, habrás hablado, habrá hablado, habremos hablado, habréis hablado, habrán hablado
		Conditional Perfect	2	habría hablado, habrías hablado, habría hablado, habríamos hablado, habríais hablado, habrían hablado
	Subjunctive	Present	5	Hable, hables, hablemos, habléis, hablemos
		Past	6	Hablara, hablaras, hablara, habláramos, hablarais, hablaran / hablase, hablases, hablase, hablásemos, hablaseis, hablasen
		Present Perfect	5	Haya hablado, hayas hablado, hayamos hablado, hayáis hablado, hayarámos hablado
		Past Perfect (pluperfect)	5	Hubiera hablado, hubieras hablado, hubíeramos hablado, hubieiras hablado, hubiéramo hablado
	Imperative	--	2	Habla (other forms are taken from the present subjunctive)

Figure 1. List of verbal times and unique conjugations in Spanish and English, separated by mood.

Pronouns. Pronouns are another area of differences between English and Spanish that may prove challenging for learners of Spanish and that may play a role in the assessment of L2 written production. The key differences lie in the fact that Spanish a richer pronoun system that includes clitics, reflexive pronouns, and gendered forms for pronouns. Additionally, pronoun placement is more flexible in Spanish which leads to a larger array of grammatically correct

combinations when compared to the stricter syntactic structure of pronouns use in English. These differences, although perhaps smaller than the conjugation systems previously mentioned, can still pose challenges and impact on the development of second-language writing skills, especially when it comes to the use of word combinations (e.g., bigram, trigrams, fixed collocations), their word order and their number and gender agreement.

Subject Pronouns. The main difference between English and Spanish use of subject pronouns is that Spanish is a pro-drop language, which means that it often omits the subject pronouns both in spoken and written mediums. Although pro-drop is presumably a typological option available to all languages, in the case of Spanish (and other indoeruopean languages) pro-drop is possible due to the rich grammatical information encoded in verbal inflection. This different from English and other non-pro-languages that cannot depend solely on context or inflection to surmise the subject of the action. The option to drop subject pronouns becomes more obviously useful in longer texts or conversations when a series of actions is attributed to the same person. Typically, in English we would have to keep repeating the subject pronoun before each verb, but that is not necessary in Spanish. This can lead to repetitiveness by English speakers learning Spanish as a result of negative transfer. For example, in Figure 2, we can see how repetitive the text results in Spanish, but how much more natural it sounds in English.

Spanish: negative transfer	English equivalent	Spanish: no negative transfer
María se despertó temprano. **Ella** hizo café y luego **ella** leyó el periódico. Después de eso, **ella**	Maria woke up early. **She** made coffee, and then **she** read the newspaper. After that, **she** went for	María se desper**tó** temprano. Hi**zo** café y luego le**yó** el periódico. Después de eso, sa**lió** a caminar.

salió a caminar. **Ella** disfrutó del aire fresco y **ella** escuchó música de camino a casa.	a walk. **She** enjoyed the fresh air, and **she** listened to music on her way home.	Disfru**tó** del aire fresco y escu**chó** música de camino a casa.

Figure 2. Examples of short text showing the use of subject pronouns in Spanish and English

Object Pronouns. The systems for object pronouns in English and Spanish have key differences in gender and number agreement, number of unique pronouns, and word order. While English has the same object pronouns for direct object (me, you, him, her, it, us, them) and indirect object pronoun (to/for me, you, him, her, it us, them), Spanish has different direct (me, te, lo/la, nos, os, los/las) and indirect (me, te, le, nos, os, les) object pronouns when it comes to the third person. Additionally, Spanish differentiates number and number (e.g., lo/los, la/las) Finally, and most impactful when analyzing the use of frequent or infrequent word combinations, is word order. English follows a Subject + Verb + Object Pronoun format, for instance, *I see her every day, He gave me a book,* and *They sent us an invitation—w*hile Spanish follows a (Subject) + Object Pronoun + Verb order with overt noun objects—e.g., *(Yo) la veo todos los días, (Él) me dio un libro,* and *(Ellos) nos enviaron una invitación* (the use of the subject pronoun here is optional and dependent on context).

Reflexive Pronouns and Clitics. However, the biggest challenge for learners of Spanish is the use of reflexive and clitic pronouns, particularly when it comes to their use and relationship with verbs. Reflexive pronouns are required in Spanish for verbs constructions in which the subject and the object of the verb phrase are the same. Typically, in English, reflexive pronouns are use for emphasis or literal reflexivity. For example, *(Yo) me lavo las manos* vs. *He loves himself.* Although he use of reflexivity in Spanish creates opportunities for English speakers to create awkward phrasings (e.g., *Me desayuno cada mañana,* meaning that I have

myself for breakfast every morning), it becomes more difficult when using more complex sentences with more verbs, on which some of these pronouns may be attached. In Spanish, these pronouns often function as clitics—unstressed elements that attach to finite or non-finite verb forms depending on syntactic context. Unlike English, which lacks pronominal clitics and relies primarily on independent subject and object pronouns, Spanish allows for the attachment of one or more clitic pronouns to verbs, particularly in infinitival, gerundive, and imperative constructions. This morphosyntactic feature poses challenges for English-speaking learners, who may struggle not only with correct placement and agreement, but also with the broader conceptual mapping between syntactic position and pragmatic salience.

The placement of clitic pronouns in Spanish is governed primarily by the finiteness of the verb. In finite verb forms, clitics precede the verb (e.g., *La vi ayer*). In non-finite forms such as infinitives, gerunds, and imperatives, they typically appear in postverbal position and are orthographically attached (e.g., *Voy a verla mañana, Estoy escribiéndola ahora, ¡Dímelo!*). In complex verb phrases that combine finite and non-finite elements, clitics may either precede the finite verb or follow the non-finite form, offering multiple positioning options (e.g., *Quiero lavarme las manos* or *Me quiero lavar las manos*). The ability to navigate these structures is often associated with higher levels of L2 proficiency, not merely because of their syntactic complexity, but because of the discourse-level demands involved in using clitics appropriately. As inherently anaphoric elements, clitics require speakers to track referents across utterances and to maintain cohesion beyond the sentence level. This ability tends to emerge at more advanced stages of development, when learners begin to manage reference in context-sensitive and interactionally appropriate ways.

Prepositions. As a unique part-of-speech, prepositions are often challenging for L2 learners because of their semantic ambiguity and the lack of direct equivalence across languages. While Spanish has fewer simple prepositions than English, with roughly 20 compared to more than 60 in English (Radden & Dirven, 2007), this does not necessarily make them easier to learn. The difficulty lies not in the size of the inventory but in the subtle distinctions prepositions encode and how they function in context. For example, *en* can translate to in, on, or at in English, depending on the situation, which creates confusion for learners who expect a one-to-one match. Spanish also does not use phrasal verbs in the way English does, so prepositions appear in different syntactic and lexical environments than learners might be used to. These differences often lead to errors in collocations or to the overgeneralization of patterns, and accurate prepositional use tends to emerge at higher levels of proficiency, when learners can more effectively manage lexical nuance and discourse-level cohesion.

Overall, the differences in lexicogrammatical features between English and Spanish creates a need to examine more closely the impact of these differences when assessing second language writing in Spanish. More specifically, it calls not only for a replication of the well-established methodology used for assessing L2 English writing, but for a closer examination of how these differences can impact the methodology itself. The literature so far has shown key differences in the complexity of inflectional systems and lexicogrammatical relationships between English and Spanish, noting that the latter presupposes a longer developmental path to achieve full mastery of such inflectional systems or lexicogrammatical relationships in the form of productive lexical proficiency measures of L2 Spanish learners.

Although some recent studies have begun to explore aspects of productive lexical proficiency in L2 Spanish (e.g., Castañeda-Jiménez & Jarvis 2013; Consolini & Kyle, 2024; Diez-Ortega & Kyle, 2024; Tracy-Ventura, 2017) and their impact on writing quality (e.g., de la Torre García et al., 2021), a multivariate approach to analyzing lexical use in graded L2 Spanish writing samples has not yet been employed. Therefore, this dissertation aims to extend previous research by using multivariate methods to model lexicogrammatical usage in L2 Spanish writing. Specifically, it will examine the relationship between indices of lexicogrammatical use and human judgments of writing quality in L2 Spanish. The focus will be on analyzing lexicogrammatical features, including lexical diversity, lexical sophistication, and the strength of bigram association measures. These features will be studied in relation to their predictive power for L2 writing quality, as assessed by human judgments. To guide this dissertation, the following research questions have been created:

1- To what extent are different operationalizations of morphological normalization when calculating lexical diversity more or less predictive of L2 Spanish writing quality in the form of human judgments?
    a. To what extent is lexical diversity predictive of L2 Spanish writing quality in the form of human judgments when tokens are morphologically normalized as raw words?
    b. To what extent is lexical diversity predictive of L2 Spanish writing quality in the form of human judgments when tokens are morphologically normalized as lemmas with parts of speech information?

c. To what extent is lexical diversity predictive of L2 Spanish writing quality in the form of human judgments when all tokens are morphologically normalized as lemmas, except for mood and tense information in verbs?

d. To what extent is lexical diversity predictive of L2 Spanish writing quality in the form of human judgments when all tokens are morphologically normalized as lemmas, except for mood, tense, person, and number information in verbs?

2- To what extent are different operationalizations of morphological normalization when calculating lexical sophistication more or less predictive of L2 Spanish writing quality in the form of human judgments?

a. To what extent is lexical sophistication predictive of L2 Spanish writing quality in the form of human judgments when tokens are morphologically normalized as raw words?

b. To what extent is lexical sophistication predictive of L2 Spanish writing quality in the form of human judgments when tokens are morphologically normalized as lemmas with parts of speech information?

c. To what extent is lexical sophistication predictive of L2 Spanish writing quality in the form of human judgments when all tokens are morphologically normalized as lemmas, except for mood and tense information in verbs?

d. To what extent is lexical sophistication predictive of L2 Spanish writing quality in the form of human judgments when all tokens are morphologically normalized as lemmas, except for mood, tense, person, and number information in verbs?

3- To what extent are features of written lexical and lexicogrammatical use in the form of lexical sophistication and lexical diversity predictive of L2 Spanish writing quality in the form of human judgments?

3 OPERATIONALIZATIONS OF MORPHOLOGICAL NORMALIZATION AND LEXICAL RICHNESS

This chapter addresses differences in the operationalization of morphological normalization (e.g., lemmas, raw words, lemmas + part of speech, etc.) and of indices involved in calculating lexical richness (i.e., lexical sophistication and lexical diversity). Furthermore, this chapter explores how considering different operationalizations is relevant when assessing the L2 lexicogrammatical features of languages other than English; and it examines two studies that offer insight for a final model that can provide predictions of L2 writing quality based on indices of lexical richness.

This chapter will be divided into three sections. The first section is subdivided into two parts: first, it reviews four ways of operationalizing morphological normalization for analyzing lexicogrammatical development in L2 Spanish writing; and secondly, it reviews the operationalization of two indices of lexical richness: lexical diversity and lexical sophistication. Additionally, the first part of this chapter examines how different operationalizations of morphological normalization may impact the results of lexical richness indices. The second and third parts review the methodology and results of two experimental studies. The first study aims to answer research questions 1a-1d which examine to what extent different operationalizations of morphological normalization are more or less predictive of L2 Spanish writing quality in the form of human judgments when calculating lexical diversity. The second one aims to answer research questions 2a-2d, which examine to what extent different operationalizations of morphological normalization are more or less predictive of L2 Spanish writing quality in the form of human judgments when calculating lexical sophistication. Finally, the chapter ends with

a discussion of the results of these studies and how the information gathered from them can inform a final multivariate model that can provide predictions of L2 writing quality based on indices of lexical richness.

Operationalizations of Morphological Normalization and Indices of Lexical Richness

As mentioned in the previous chapter, the concept of lexical richness (e.g., Yule, 1944; Laufer & Nation, 1995; Read, 2000) commonly includes three measurements of lexicogrammatical productivity: lexical diversity, sophistication, and density. However, it has been suggested (Ure, 1971; Shohamy, 1994) that density may be helpful to distinguish between different registers, but not to assess lexical productivity. For this reason, this research study focuses on the measurement of lexical diversity and lexical sophistication. In the following section, a more detailed overview is taken at the operationalizations of these indices.

Lexical Diversity

Lexical diversity can be measured in a few different ways. This research study focuses on MTLD (McCarthy & Jarvis, 2010) and MATTR (Covington & McFall, 2010). Both these indices of lexical diversity have been shown to be stable when analyzing texts of at least 50 words in length (Jarvis, 2017; Kyle et al., 2021), although smaller windows for the operationalization window of these indices (below 50) have been used successfully in studies that investigate other registers, such as spoken data (e.g., Fergadiotis, 2015; Kyle et al., 2024). Furthermore, research has shown a positive correlation ($r = .566$) between MATTR and MTLD and human judgment scores of L2 English writers (e.g., Kyle et al., 2021). Whereas both these indices measure lexical diversity, they do so slightly differently and with the potential to be complementary. For instance, MTLD reads through the text one word at a time calculating a running Type-Token-

Ratio until the ratio falls under 0.72. When this happens, all the tokens up to that point become a factor, and the running TTR starts again. This repeats until the end of the text, and then backwards to the beginning of the text. The lengths of all the factors generated from this calculation are then averaged to provide the MTLD score. This captures overall lexical diversity across longer stretches of text while being less affected by clusters of repetition since it resets once TTR drops, providing a more global view of textual diversity. This shows a clearer picture of vocabulary development over time or across proficiency. Similarly, MATTR uses a fixed window of words in the text (commonly 50 words) and then it slides that window one word at the time and calculates TTR for each slide. After, all the window-level TTRs are averaged to provide a score. This method works well for shorter text commonly found at lower proficiency levels, and it provides a more local view of diversity or repetition. While MATTR is more sensitive to the short-term reuse or repetition of words, MTLD is more sensitive to gradual variation and long-range repetition. Therefore, utilizing both when looking at textual lexical diversity can not only help cross-validate findings; it can also provide a deeper assessment of current writing development. A low MATTR score with a high MTLD score might indicate global diversity but an overuse of fixed expressions, whereas a high MATTR score with a low MTLD score might suggest a text that starts with a wider range of lexical diversity but becomes repetitive over time. Moreover, low scores on both MATTR and MTLD would suggest a mostly formulaic and repetitive text, commonly written by lower proficiency writers, while high scores on both indices could indicate global and local diversity more commonly produced by advanced proficiency writers. Recent research in English has explored the relationship between different operationalizations of lexical diversity indices (MTLD, MTLD-W and MATTR-50) and the scores of human judgments of English essays written by both native speakers and learners. This

research suggests that the best operationalization for English when calculating lexical diversity is simple lemmas with part-of-speech tags that have been manually corrected, and whenever that is not possible, using family words has been found to be an acceptable alternative (Jarvis & Hashimoto, 2021). These results might work best for English, where verbal inflection is not very complex and not many layers of grammatical information (tense, mood, person or number) are encoded in conjugated verbs. Since this is where Spanish differs from English the most, the following study will explore these ideas in more detail to examine if there are different results in L2 Spanish essays.

Lexical Sophistication

A multivariate approach to measuring lexical sophistication can provide a much richer understanding of learners' lexicogrammatical developmenttaking into consideration not only the advanced lexical items but also the use of combined lexical items and the grammatical relationship between these items. To achieve a multivariate approach that measures these characteristics, this research study focuses on the measurements of word frequency, bigram strength of association (SOA), and object, adverb, and adjective bigram dependencies SOAs. The following section explains each in detail.

Word Frequency. In order to estimate lexical sophistication measures based on the frequency of use of a given word, a reference corpus is necessary. For each word used in the learner's text, an average frequency number is calculated by noting how many times per million the same word is used in the reference corpus. This provides a more fine-grained analysis of word frequency, compared to methods that separate words into 1,000-word bands, such as the lexical frequency profile (LPF; Laufer, 1994). The use of frequency bands can make it more

difficult to observe small advances in lexical development across time or proficiency levels. In contrast, combining the average word frequency of each word used can provide a more robust analysis of lexical sophistication when analyzing individual texts. However, there are different approaches to average word frequency calculations as well. Research suggests that calculating frequency scores of content words is more predictive of writing quality than calculating frequency scores combining content and function words together (Monteiro et al., 2020). It is important to note that the frequency values used are log transformed values. Transforming the raw frequency values addressed the skewed Zipfian (Zipf, 1932) distribution of word frequencies that occur naturally in languages. Applying a logarithmic transformation to raw frequency counts can aid in normalizing the data, making it more suitable for statistical analysis. This transformation reduces the impact of extremely frequent words and spreads out the distribution of less frequent ones, which leads to improved interpretability and linearity in statistical models. For this research study, the frequency of all words and content words was calculated and then log transformed before being averaged. Content words carry semantic meaning (noun, verbs, adjectives, adverbs), while function words serve grammatical purposes (pronouns, prepositions, conjunctions, articles, determiners). When calculating all words, both content and function words were analyzed.

Bigram Frequency and SOA. Beyond word-level analysis, the frequency of use of bigrams and the strength of association between those bigrams has shown to provide insight into the lexicogrammatical sophistication of L2 writers. Bigram frequency has been shown to be more predictive of proficiency level than isolated words (Kyle & Crossley, 2015; Garner et al., 2019), and bigram strength of association measures have shown positive correlations with L2 writing development (Bestgen & Granger, 2014; Garner et al., 2019). In this research study,

bigram frequency is calculated very similarly to word frequency; however, instead of a frequency score for each word, a frequency score is created by assessing combinations of two contiguous words in an essay. These combinations of words used by the learner are then examined in the reference corpus and a frequency score is created for each pair. This provides information about the collocations that learners use in their L2 writings and how sophisticated they are in the target language by looking at how often they are commonly used in the reference corpus.

Moreover, for this research study, bigrams SOA are measured by calculating T scores, pointwise mutual information (MI) scores, and ΔP scores (i.e. Delta P; Allan, 1980). T-scores (Allan, 1980) are used to measure the SOA between bigrams or trigrams based on their frequency of co-occurrence. In linguistic analysis, a higher T-score indicates that the words that comprise the bigram have a stronger strength of association, while lower T-scores may indicate that the co-occurrence could be due to randomness rather than linguistic relationship. This measurement commonly gives more importance to frequent bigrams, making it most useful for identifying common collocations. Along with T scores, MI scores (Church & Hanks, 1989) are used to provide a more comprehensive picture of SOA measurements. This method evaluates how much more often a pair of words occur together than it would be probable to occur. Higher MI scores suggest a stronger association between the words in a given bigram, while lower MI scores closer to 0 suggest a lower association, which can indicate that the association is closer to random co-occurrence. Some research studies (e.g., Evert, 2005) have shown that MI can tend to overestimate bigrams that are rare since low-frequency pairs can get a high MI score if the words appear separately on rare occasions. As a consequence, it is common in linguistic analysis to use

T-scores and MI scores together to get a better understanding of bigrams' strength of association (e.g., Gries, 2013; Ellis, 2012).

Furthermore, Delta P calculates the probability of getting a certain outcome (for instance, word 2 in a bigram pair), given a certain cue (word 1 in the same bigram pair), minus the probability of getting the same outcome (word 2) with any other cue (word 1). Additionally, this is calculated bidirectionally. That is to say that Delta P examines the probability of the second word in a bigram pair occurring based on the word that precedes it, but it also examines the probability of the first word in the same bigram pair occurring based on the word that follows it. This is done in separate calculations, one from left to right where the first word on the bigram is the cue to calculate the probability of the second word occurring, and another calculation from right to left, where the second word is the cue. This allows to find different association patterns and their strength, since bigrams such as “upside down” or “instead of” are as strongly associated as bigrams such as “of course” or “for example”, although the predictive cue for the first ones is the first word in the bigram and for the second ones is the second word in the bigram. The directionality of the strength of association provided by Delta P is rich information, given that other measures would mark both these groups of bigrams as mutually associated. This is important from a learning perspective, since associations do not need to be mutual or equally strong in each direction (Tversky, 1977). This is very relevant at the time of assessing writers’ development in their second language, since weaker associated bigrams can be composed of very common words that can collocate with many other words, making their bond not unique and therefore less sophisticated. However, in some cases, very strongly associated bigrams (in both or either directions) tend to be more content-rich and semantically bonded expressions that carry more complex or nuanced meaning (Ellis & Ferreira, 2009; Durant & Schmitt, 2009).

Bigram Dependencies. Bigram dependencies depict grammatical relationships such as Subject-Verb, or Adjective-Noun. While calculating bigram strength of association provides an evaluation of semantic connection between contiguous words, bigram dependencies go even an extra step further and provide information about the grammatical relationship between two given words. Most importantly, when examining bigram dependencies, bigrams that are being assessed for their dependencies do not need to occur adjacently, in contrast with indices of bigram strength of association or frequency for which calculations are carried out based on contiguous words. For example, in the sentence "The tall boy quickly ran to the store," the words “boy” and “ran” are not adjacent to each other, but they are in a grammatical relationship. Research has shown that certain types of dependency relationships in different languages can be predictive of writing proficiency (Kyle & Eguchi, 2021; Paquot, 2019; Rubin et al., 2021). This research study focuses on Subject-Verb, Object-Verb, Adjective-Noun, Adverb-Verb dependencies, and the frequency of use of such dependencies.

Morphological Normalizations

Morphological normalization refers to the process of systematically standardizing the forms of words in a text to focus on their linguistic information. This leads to the ability of controlling for variation caused by inflection in order to analyze different aspects or features of lexicogrammatical use. Given the richer verbal inflectional system in Spanish compared to English or other Germanic languages, learners’ productive lexicon varies not only in the breath (how many words they know) but in how accurately they produce inflected forms that carry grammatical meaning. In this sense, different approaches to operationalizing words can allow for different assessment foci: vocabulary breadth (in the form of lemmas), grammatical control

(provided by inflectional information), and developmental stages (e.g., emergence of subjunctive mood). It is important to note that each operationalization also presumes different information about the learners' lexicon, which is also discussed. Below is a short description and rationale for the different ways in which words are operationalized in this research study.

Raw Words. Raw words are words exactly as they appear in the text. In this sense, after being tokenized, the words are not changed into any other form. This means that words include all their inflectional morphological information, which in Spanish includes number and gender inflection for nouns and adjectives, and tense, mood, person, and number for verbs. This operationalization of tokens reflects what learners are actually producing, which can be useful for in-depth error analysis and interlanguage profiling. However, due to the morphological productivity of verbs in Spanish (e.g., *hablar* has 36 unique forms while *talk* has 5 unique forms) and because every inflected form of a word counts as a different type in raw word analysis, it is possible that normalization to raw words inflates indices of lexical diversity, which, in turn, can obscure lexical patterns when assessing vocabulary size or sophistication at a large scale. Moreover, the operationalization of tokens as raw words presumes that each form is stored in the learner's lexicon separately. This means that it cannot be assumed that the learner knows the word *casas* if they only used the word *casa*. This can extend to other parts of speech such as the verbal inflectional system where the assumption could be more realistic for lower proficiency learners. For instance, knowing how to conjugate most verbs in the present tense for first person singular (e.g., *yo hablo*) does not necessarily mean that the learner can conjugate them for all other persons and their different inflections (*tú hablas, él habla, ellos hablan, nosotros hablamos*). This operationalization could potentially provide a more accurate and precise analysis of lexical use at all levels of proficiency. However, it could also obscure indices with

small inflectional variations (e.g., *casa* vs. *casas*, or *casa roja* vs. *auto rojo*). Although these small variations may be useful to examine proficiency and developmental trends in lower-proficiency learners, it may obscure or overestimate the actual richness of more advanced learners.

Lemmas with Part of Speech. Lemmas are the base or dictionary form of a word, regardless of its grammatical inflection. For instance, in Spanish, all 36 forms of the verb *hablar* would be lemmatized into the form *hablar* every time the learners used the inflected forms. In order to make the measurement of different features of lexicogrammatical use more accurate, tokens get tagged for part of speech as well. This approach of using lemmas with part of speech information for linguistic analysis reduces the noise from inflectional changes while still distinguishing grammatical categories between homographs. This combination achieves a balance that provides more accurate analysis of lexical richness by providing a more accurate type-count in lexical diversity analysis. Furthermore, the operationalization of tokens as lemmas presume that the learner knows the variations of each token and therefore the lemma form suffices to examine their lexicon breadth and depth. For example, the lemma *casa* presupposes the learner knows about the variation of *casas*. Utilizing this approach, however, all inflectional details for inflection are also presupposed. Since Spanish has such a rich inflectional system, this information can be key for tracking morphosyntactic development because we cannot presume that learners will be able to acquire and employ the whole system in lower proficiency levels. Instead, the verbal inflectional system and its complex morphosyntactic interactions are acquired in stages over time (Bowden et al., 2010; Fernández-Dobao & Herschensohn, 2019). Lemmatization may not be an issue when assessing L2 English since a lot of grammatical information is encoded more frequently in grammatical dependencies and less in grammatical

information encoded in the morphological system. For instance, *hablo* in Spanish encodes information about the person (first person singular), the tense (present) and the mood (indicative), while *talk* in English could refer to different persons (I, they, we) and only directly encodes information for tense (present).

Lemmas with Verbal Information. A step further from lemmas tagged for part-of-speech is adding mood and tense information to lemmatized verbs. In this sense, verbs continue to be lemmatized, so they do not inflate lexical diversity measures; however, mood and tense information are added to each lemmatized verb. This approach reduces lexical richness inflation due to many inflected forms, but it retains morphogrammatical markers (verbal tense and mood) that can provide insight into learners' proficiency development. Furthermore, the operationalization of tokens as lemmas with tense and mood information presume that the learners group morphological forms into lemmas but are sensitive to grammatical categories like tense and mood, grouping or organizing their mental storage both by form and function. However, assessment of person and number distinction cannot be done since this operationalization would also presume that since the tense and mood variations are known, they are known for all the persons, in both their singular and plural forms. This is particularly relevant to Spanish where the conjugation of many verb tenses and different moods might be a marker of proficiency. Nonetheless, this approach may still inflate token variation since there are more morphological categories for many same lemmas, and important learner information is still missing such as subject-verb agreement, or noun-adjectives agreement, overlooking errors such as *yo *hablas* or *autos *bonita*. It is not uncommon for English learners of Spanish to struggle with subject-verb and noun-adjective agreement (e.g., Nicol & Greth, 2003); therefore, this approach does not provide the ability to analyze agreement errors or the emergence of person-

number control in verbs. Yet another step is to tag lemmatized verbs with all the information they provide in their inflectional systems; that is, mood, tense, person and number. This approach can maximize granularity for morphological detail, providing information that may allow us to look at the developmental trajectory of verb acquisition in Spanish. For instance, the emergence of first-person plural, which tends to happen after the acquisition of first-person singular (Marrero & Aguirre, 2003), might be an indicator of proficiency development.

Study 1: Lexical Diversity

In order to answer the first set of research questions that examine the relationship between indices of lexical diversity and different operationalizations of morphological normalization, a study was conducted with a corpus comprised of L2 Spanish descriptive writings. The following sections detail the methodology and results.

Learner Corpus

The learner corpus used for this study is from a subsection of The UC Davis Corpus of Written Spanish, L2 and Heritage Speakers (COWSL2H; Sánchez-Gutiérrez, 2022), which is freely available at https://github.com/ucdaviscl/cowsl2h. This corpus is comprised of essays collected from students enrolled in university-level Spanish courses at the University of California Davis between 2017 and 2023. As of writing, the corpus includes 5,383 essays from 1,934 unique learners, and a total of 1,367,258 million tokens. The essays for this corpus were collected untimed and online through the Canvas LMS. Enrolled students were offered extra credit to complete the essays but were told that their essays would not be seen by their instructors, and that their performance would have no impact on their grades or the amount of extra credit that they would receive. Furthermore, the corpus is separated in two subsections, one

includes essays written by Spanish L2 learners, and another subsection includes essays by Spanish heritage speakers. This study will explore the subsection of Spanish L2 learners whose first language is English. This subsection includes eight descriptive prompts, of which four were selected for this study. These four prompts were selected based on their clear focus on different topics. For instance, the prompts "Write a text in Spanish about the following subject: a famous person", and "Write a text in Spanish about the following subject: a special person in your life" would most likely elicit the same type of description, that of a person. Additionally, prompts that have essays from as many different courses as possible were prioritized in order to acquire a more representative sample across proficiency levels. Table 1 shows a breakdown of the four descriptive prompts selected for this study and the number of essays for each prompt before a sample for the study was taken.

Table 1

Descriptive Statistics for Descriptive Essays on COWSL2H before sampling

Prompt	N of essays
A famous person	905
A perfect vacation	820
A terrible story	874
A beautiful story	739

The essays collected for the COWSL2H corpus are from different Spanish courses at the University of California Davis. The learners that have participated in the compiling of this corpus were enrolled in either Beginner Spanish (Spanish 1, 2, or 3), Intermediate Spanish (Spanish 21 or 22), Spanish Composition (Spanish 23 and 24) and Spanish Composition for

Heritage Speakers (Spanish 21, 32, or 33), or Advanced Spanish (upper division courses in Spanish). A copy of the UC Davis catalog of Spanish courses and their descriptions can be found in this link: https://catalog.ucdavis.edu/courses-subject-code/spa/ . For the purpose of this research study, essays from the Spanish Composition for Heritage Speakers (Spanish 21, 32, or 33) were not included.

The compilers of COWSL2H also provide a metadata file for each of the essay files. The metadata files contain data regarding the learners' background information, including the level of the Spanish course they were enrolled in at the time they submitted the essay. In order to create a balanced and stratified sample, a short Python program was created to iterate through the metadata files to provide a random sample of 100 essays per prompt, and a balanced number of essays per proficiency level. Table 2 shows a detailed view of the number of essays per prompt and per level. Additionally, to ensure that essays were long enough to obtain reliable indices scores, only essays 50 or more words in length were considered (Zenker & Kyle, 2021).

Table 2

Descriptive Statistics for Descriptive Essays on COWSL2H after sampling

Level/Prompt	Beautiful	Famous	Terrible	Vacation	Total
Beginner	33	41	45	43	162
Intermediate	15	30	28	29	102
Advanced	52	29	27	28	136
Total	100	100	100	100	400

Reference Corpus

In order to calculate average frequency for words and bigrams, a different corpus was used as a reference. The Spanish Corpus of the Web (ESCOW) is a corpus of Spanish language with over 7.3 billion tokens from the World Wide Web (Schäfer, 2015). In this study, the AX01 subset of this corpus was used. This subset included randomly sampled sentences shuffled into a series of different files, all totaling 463 million tokens. Contrasting with many other Spanish corpora that are available for reference, ESCOW is reasonably large and representative, it allows for texts to be downloaded (and subsequently processed by researchers), and is open-source (allowing products from the corpus such as frequency lists to be disseminated freely).

Human Judgments

In order to procure human judgments of writing quality for all of the sample essays, three raters were trained, and calibration sessions were carried out with examples from outside the sample utilized for analysis in this study. Essays were rated using the adapted rubric from Connor-Linton and Polio (2014), found in Appendix A. This analytic rubric provides an assessment of content, organization, vocabulary, language use, and mechanics on a scale from 0 to 20 for each assessment criterion, with a total possible score between 0 and 100. The only modifications made to the original rubric were under the Mechanics criterion where the phrasing "appropriate layout and indented paragraphs" was removed from every scale of the criterion, and "Little to no spelling or punctuation errors" was added to the highest scale of the criterion to supplement the original phrasing of "no spelling or punctuation errors". Additionally, the phrasing for assessing the number of words was removed from the "Content" criterion.

Furthermore, raters were trained by the researcher and were provided with supplemental information to guide the process of assessment, and comments in the rubric to specify details or provide contextualized examples relevant to the current study. To train the raters, several 1.5-hour long video conferencing sessions were carried out across six weeks. The first session included an introduction to the project, an overview of the rubric and what each section of it represents, and three example essays were graded by the researcher using the rubric. The essays represented a median of each of the levels represented in the corpus sample (beginner, intermediate and advanced), and raters were encouraged to ask questions during the training process. After, 12 essays were provided to the raters to rate on their own time before meeting again for a calibration session. These 12 essays included three essays from each level category and were not part of the final 400 essays used for analysis. Four calibration sessions were conducted. For each calibration, a set of 12 different essays were provided to the raters. During calibration sessions, raters would share their ratings for each essay upfront, and then the researcher and the raters would discuss in detail those essays that created disagreement in the scores. After calibration sessions were held and the essays used for training were discussed, a final batch of 31 essays were given to the raters to rate individually. Finally, a Many-Facet Rasch Model (MFRM) was run using those last 31 essays. The model included the five domains of the rubric and the three raters. The EAP reliability was 0.972, indicating the model is estimating learners' writing ability (or essay quality) very precisely, based on how the essays were rated. EAP reliability in Rasch models measures the precision of person estimates, analogous to Cronbach's alpha. Values close to 1.0 are considered highly reliable (Embretson & Reise, 2000). Moreover, mean-square fit statistics were acceptable, with an average outfit of 1.257 (SD = 0.912) and infit of 1.219 (SD = 0.789), both within the productive measurement range of 0.7–1.3

(Linacre, 2002; Bond & Fox, 2015). Lastly, analysis of rater severity revealed small differences: Rater 1 was slightly more severe (-0.049 logits), Rater 2 slightly more lenient (0.051 logits), and Rater 3 approximately neutral (-0.002 logits). Differences of less than 0.5 logits between raters are generally considered minimal and unlikely to introduce significant bias into measurement (Linacre, 2002; Engelhard, 2013). A summary of this model can be found in Table 3.

Table 3

Summary of Many-Facet Rasch Model for Calibration Data

Parameter	**Value**
AIC	2189
BIC	2228
EAP Reliability	0.972
Mean Infit MSQ	1.219 (SD = 0.789)
Mean Outfit MSQ	1.257 (SD = 0.912)
Rater 1 Severity	-0.049
Rater 2 Severity	0.051
Rater 3 Severity	-0.002

Finally, the 400 essays for research analysis, after calibration, were randomly assigned to each rater with a short Python program that ensured that each essay had two different raters assigned and that each rater had a similar number of essays to analyze. Following this design ensured scoring reliability since each essay was independently scored by two different rates,

while the third rater could intervene as an adjudicator to provide a third score in cases of substantial scoring differences. To initialize adjudication, a difference threshold of 20% between scores was established, in line with best practices in performance assessment (e.g., Johnson et al., 2000; ETS, 2012). This design is utilized to minimize rater disagreement without unnecessarily increasing the scoring burden. In the present study, none of the essays exhibited differences exceeding the 20% threshold, and therefore no adjudication was required. This protocol ensured both the validity and efficiency of the scoring process while maintaining high interrater reliability. Table 4 shows the distribution of essays per rater.

Table 4

Distribution of essays per rater

Rater	Number of essays
Rater 1	281
Rater 2	257
Rater 3	262

Linguistic Analysis

Prior to analyzing the COWSL2H L2 Spanish sub-corpus, a phase when data was pre-processed was necessary to take into account tokens that might not be appropriate to consider when analyzing L2 Spanish lexical proficiency. Firstly, tokens were removed from COWSL2H that had a direct impact on lexicogrammatical scores, but that did not directly represent lexical proficiency in Spanish, such as common English words, proper names in English, and numbers.

This was an important step, since some of these tokens were found in the Spanish reference corpus (ESCOW), but since they appeared sporadically, they turned out with very low frequency scores (e.g. massage [n = 40], occupational [n = 51]). These instances sometimes significantly lowered the score for word frequency in Spanish essays that had no low-frequency words in Spanish. In order to account for these tokens, all the tokens in COWSL2H that appeared in the list of 10,000 most frequent words in the written section of the Corpus of Contemporary American English (COCA; Davies 2010) were filtered and subsequently removed and saved into a list. The list was later revised to ensure all words in Spanish and all homographs between English-Spanish (e.g. alcohol, festival, local) were still included in the final analysis. These steps were key to ensure that the analysis produced scores that accurately represented the Spanish lexicogrammatical features used by L2 writers.

Additionally, most of the essays in COWSL2H had been processed to edit out potentially private or identifying information. Therefore, names of people, cities, states, countries, and dates were removed from the essays and tags with a corresponding tag, such as *CITY*, or *NAME*, before being shared publicly as part of the corpus. In order to account for this, and with the goal of ensuring that these tag words were not counted as part of the essay, a short python program was created to replace each of the tags with a random related token. For instance, the tag *CITY* was replaced with a random city from a list of American cities. Finally, since the essays and the metadata from the essays were in separate files, a python script was created to extract the course level from the metadata file and pair it with the essay file. This is useful as an extra measure that informs proficiency along the human judgment scores.

Lexical Diversity. In order to explore and answer the set of research questions for study one, the MTLD and MATTR indices were calculated four different times with the same 400 sample essays. The first time, the words were operationalized as raw words, where all forms of the words used by the learners are kept as they are without any modifications. For example, the words in the sample sentence *En mi opinión las vacaciones perfectas deberían de ser en *un parte que uno nunca ha visitado* ('In my opinion the perfect vacations should be at a place one has never visited') would be tokenized as they appear in the essay, and not modified, for example, into their base or dictionary forms. The second time, words were lemmatized and tagged with part-of-speech information. While all words were reduced to their base or dictionary form, they retained information about their part of speech. In the same sample sentence, the content words are reduced to their dictionary form while the function words are kept as they are. Table 5 provides a clearer understanding of what is missing in comparison.

Table 5

Lemmatization of sample sentence

Original Word	**Part of Speech**	**Lemma**
en	Preposition	en
mi	Possessive adjective	mi
opinión	Noun	opinión
las	Article	el
vacaciones	Noun	vacación
perfectas	Adjective	perfecto
deberían	Verb	deber

de	Preposition	de
ser	Verb	ser
en	Preposition	en
un	Article	un
parte	Noun	parte
que	Relative pronoun	que
uno	Pronoun	uno
nunca	Adverb	nunca
ha	Auxiliary verb	haber
visitado	Verb	visitar

The third time, the essays were lemmatized with part of speech, except that verbs were also tagged for tense and mood. And finally, the fourth time, the essays were lemmatized with part of speech, except that verbs were also tagged for tense, mood, person and number.

Statistical analysis

In order to address the first set of research questions, first an inter-rater reliability and rater severity analysis was conducted. Although the training data was a good fit for a Multi-Facet Rasch Model (all raters in training scored all essays), the data used for the study's analysis is different. In this case, only two of the three available raters provided scores for each essay, allowing the third rater in each pair to be an adjudicator in case of scoring differences greater than 20%. The first model with the calibration data was a fully crossed design (3 raters, 3 scores per essay). However, with the design chosen for the study's 400 essays, the data becomes

incompletely crossed (2 raters randomly assigned from a pool of 3 raters) to allow for adjudication and reduce scoring burden on resources. Not fully crossed designs (i.e., when not all raters rate all essays) can result in unstable or non-identifiable solutions (Engelhard, 2013; Linacre, 1994). This is due to the fact that adequate crossing among observations is crucial for identifying latent trait parameters in Many-Facet Rasch Models. Since each essay was scored by two randomly assigned raters and the essay ratings were continuous numeric scores, interrater reliability was assessed using the intraclass correlation coefficient (ICC(2,1)), which accounts for random assignment of raters and focuses on absolute agreement (Shrout & Fleiss, 1979).

The results of ICC interrater reliability indicated moderate to good reliability across scoring categories. Tables 6 and 7 have a detailed summary of these results. Of note are the reliability of scores for the categories of Language Use (ICC = 0.73) and Vocabulary (ICC = 0.72), which are considered good reliability scores (Shrout & Fleiss, 1979). Furthermore, to assess rater severity, the mean score, standard deviation, and number of ratings provided by each rater across all categories were calculated. Rater 1 showed the highest mean score (M = 13.79, SD = 3.63, n = 1285), suggesting relative leniency, while Rater 2 had the lowest mean score (M = 12.59, SD = 3.23, n = 1405), suggesting greater severity. Finally, the mean scores from Rater 3 (M = 13.40, SD = 3.07, n = 1310) are between the other two raters. Overall, variability in scoring is comparable across raters, although minor differences in central tendency can be seen. Tables 6 and 7 summarizes this information in detail, and Figure 3 provides a visualization of rater severity.

Table 6

Inter-reliability results from Intraclass correlation coefficients

Category	ICC
Content	0.57
Language Use	0.73
Mechanics	0.65
Organization	0.59
Vocabulary	0.72

Table 7

Rater severity scores from ICC

Rater	Mean Score	Standard Deviation	Number of Ratings
1	13.79	3.63	1285
2	12.59	3.23	1405
3	13.4	3.07	1310

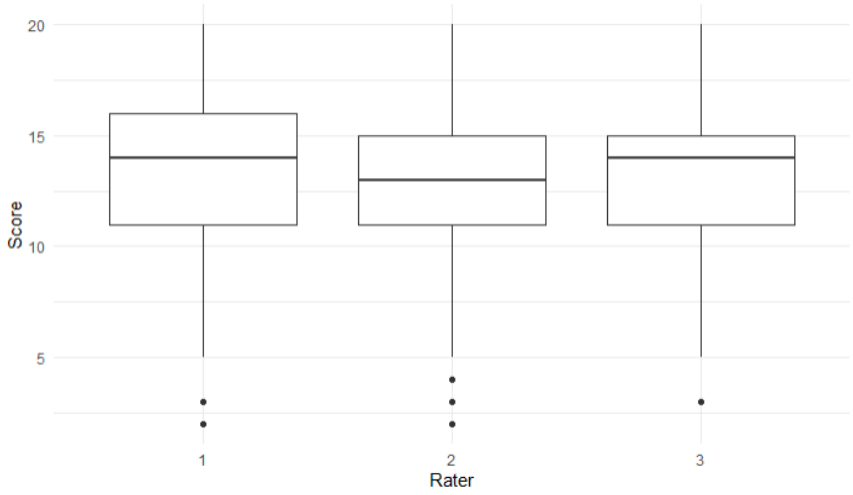


Figure 3. Rater severity from ICC

After inter-rater reliability and rater severity scores were finalized, and with the understanding that scores did not different between raters more than 20%, scores per category and total scores were averaged for each essay. Therefore, one final score per category and as total score was used for each essay. The following steps were to run regression analysis for each index and each operationalization within each index. These analyses were then compared to identify which operationalization provided better predictions of human judgment scores. To answer the first set of research questions, lexical diversity indices were calculated first.

Firstly, MTLD and MATTR scores were visualized using R (R Core Team, 2016). Histograms were created for each variable to account for a normal distribution of the data. This included not only the MTLD and MATTR scores for each operationalization, but also the scores from the human judgments for each of their categories and the final scores. Secondly, scatter plots were used to visualize the linearity of each variable in reference to the human judgment scores. Moreover, the cor() function (R Core Team, 2016) was used to examine the size of the correlations between the human judgments scores and MTLD and MATTR scores, separately.

After these steps were completed, multiple regressions were conducted to investigate the predictive ability of each operationalization of MTLD and MATTR. For each index, the predictive power of each operationalization was compared by contrasting the models on R^2 and Akaike's Information Criterion (AIC). Higher R^2 and lower AIC values were interpreted as indicators of better predictive performance (Burnham & Anderson, 2002; Field, 2013). For this study, AIC differences greater than 2 points are considered indicative of meaningful differences in model quality (Burnham & Anderson, 2002).

Results

The initial analysis of distribution of the human judgment and lexical diversity scores showed that the data did not strongly deviate from a normal distribution. Additionally, all operationalizations of the indices showed a relative linear relationship with the human judgment scores. To examine the extent to which lexical diversity predicts essay quality, linear regressions were conducted with MTLD and MATTR scores as predictors and human judgment scores as outcomes. All statistical files can be found in the following link: https://osf.io/3wdes/?view_only=1071a0474bf546f6a740016e9de0d006

Moreover, models were tested across four morphological operationalizations: raw words, lemmas, lemmas tagged with tense and mood (TM), and lemmas tagged with tense, mood, person, and number (TMPN). In this section, human judgment scores for the categories of Vocabulary and Language Use are reported along with the Total score, while scores for Mechanics, Content, and Organization are not reported since these are not constructs evaluated by indices of lexical diversity. Moreover, to evaluate the relative fit of models using different operationalizations, Akaike Information Criterion (AIC) and R^2 values were compared between

models. While AIC does not provide a statistical significance test, differences of 2 or more points between models are commonly interpreted as meaningful, and differences of 10 or more points as strong indications of improved model fit (Burnham & Anderson, 2002). Similarly, R^2 values that cross benchmark thresholds (from small = .01, to medium = .09, to large = .25) represent improvements in model fit (Cohen, 1988).

Based on comparisons of AIC values and R^2, the TMPN operationalization consistently showed the best fit across all evaluated categories of measurement (Vocabulary, Language Use, Total) for both MTLD and MATTR. Tables 8-10 below show consistently higher R^2 and lower AIC for all categories with the operationalization that includes lemmas and verbal information for tense, mood, person and number for MTLD scores.

Table 8

MTLD regression models comparison for total scores

Category	Operationalization	β (Beta)	R^2	AIC
Total	Raw	0.253	0.172	3001.11
Total	Lemma	0.432	0.178	2998.04
Total	Lemma + Tense/Mood	0.357	0.173	3000.67
Total	Lemma + TMPN	0.301	0.200	2987.22

Table 9

MTLD Regression models comparison for vocabulary

Category	Operationalization	β (Beta)	R^2	AIC
Vocabulary	Raw	0.059	0.183	1802.31
Vocabulary	Lemma	0.101	0.195	1796.61
Vocabulary	Lemma + Tense/Mood	0.086	0.196	1795.85
Vocabulary	Lemma + TMPN	0.070	0.215	1786.53

Table 10

MTLD Regression models comparison for language use

Category	Operationalization	β (Beta)	R^2	AIC
Language Use	Raw	0.065	0.225	1781.91
Language Use	Lemma	0.109	0.223	1782.55
Language Use	Lemma + Tense/Mood	0.092	0.228	1780.35
Language Use	Lemma + TMPN	0.076	0.250	1768.41

These summarized tables also show higher β (Beta) scores for simple lemmatization without verbal information, indicating only that the variation in the slope of the predictor is bigger since there is less morphological precision or detailed information for each token.

Consequently, the β (Beta) score for lemmatized token with tense, mood, person and number information is slightly lower since more fine-grained morphological information creates a smaller slope for the predictor. Finally, for the most completely operationalization (Lemma + TMPN) the R^2 scores are higher than the R2 scores of other operationalization by between 0.19 and 0.32; and the AIC scores are lower by between 10 and 16 points. Similarly, results for MATTR consistently showed higher R^2 (between .11 and .23) and lower AIC scores (between 5 and 10 points) for all categories with the operationalization that includes lemmas and verbal information for tense, mood, person and number. Tables 11-13 below summarize this information clearly.

Table 11

MATTR regression models comparison for total scores

Category	Operationalization	β (Beta)	R^2	AIC
Total	Raw	100.43	0.223	2975.84
Total	Lemma	100.41	0.212	2981.30
Total	Lemma + Tense/Mood	95.62	0.211	2981.83
Total	Lemma + TMPN	101.02	0.234	2970.19

Table 12

MATTR Regression models comparison for Vocabulary

Category	Operationalization	β (Beta)	R^2	AIC
Vocabulary	Raw	23.08	0.232	1777.41
Vocabulary	Lemma	23.38	0.227	1780.19
Vocabulary	Lemma + Tense/Mood	22.44	0.230	1778.90
Vocabulary	Lemma + TMPN	23.38	0.247	1769.67

Table 13

MATTR Regression models comparison for Language Use

Category	Operationalization	β (Beta)	R^2	AIC
Language Use	Raw	24.75	0.267	1759.51
Language Use	Lemma	25.37	0.267	1759.39
Language Use	Lemma + Tense/Mood	24.44	0.272	1756.66
Language Use	Lemma + TMPN	25.11	0.285	1749.56

Finally, a comparison of MATTR and MTLD models was carried out. Table 14 below shows consistently higher R^2 and lower AIC scores for MATTR across categories. This indicates that calculating lexical diversity with MATTR and operationalizing tokens as lemmas with verbal information (mood, tense, person, and number) provides a stronger prediction of writing

quality for the 400 essays from the COWSL2H corpus analyzed in this study, compared to other forms of operationalizations.

Table 14

Lemma + TMPN comparison between MTLD and MATTR

Category	R^2 MTLD TMPN	AIC MTLD TMPN	R^2 MATTR TMPN	AIC MATTR TMPN
Content	0.109	1949.48	0.126	1941.80
Language Use	0.250	1768.41	0.285	1749.56
Mechanics	0.075	2042.217	0.091	2035.29
Organization	0.099	1797.474	0.119	1788.54
Vocabulary	0.215	1786.53	0.247	1769.67
Total	0.200	2987.22	0.234	2970.19

Discussion

This first study investigated the relationship between different operationalizations of morphological normalization when calculating indices of lexical diversity for 400 L2 Spanish essays, and their relationship with human judgment scores. To explore the relationship between human judgement scores and the different indices of lexical diversity and their operationalization, MTLD and MATTR scores were calculated for all 400 essays with four different operationalizations: raw words, lemmas tagged with part-of-speech, lemmas tagged with part-of-speech and mood and tense verbal information, and lemmas tagged with part-of-

speech and mood, tense, person, and number verbal information. After assessing normality and linearity, linear models were created to analyze the relationship between each of these indices' operationalizations and human judgment scores of writing quality that assess the categories of Language Use, Vocabulary, and total score. The results, discussed in more detail below, have implications for the operationalization of lexical diversity in multivariate models to predict L2 writing quality in Spanish.

The results of the regression models indicated that MTLD with the operationalization of lemmas tagged with part-of-speech and mood, tense, person, and number verbal information accounted for 20% of the variance in total scores of human judgments, while accounting for 21% and 25% of the variance in Vocabulary and Language Use scores of human judgments, respectively. Other operationalizations of MTLD accounted for 17% (tagged lemmas plus tense and mood), 18% (tagged lemmas), and 17% (raw words) of total scores of human judgements. Similarly, these operationalizations accounted for 20% of the variance for Vocabulary scores, and 22% of the variance for Language Use scores. The difference between operationalizations, as assessed by Akaike's Information Criterion (AIC), was most notable between lemmas with full verbal information and all others. AIC differences greater than 2 points are considered indicative of meaningful differences in model quality (Burnham & Anderson, 2002). The AIC scores for MTLD with lemmas and full verbal information for total human judgment scores was 2987, while the AIC scores for the other operationalizations were 3000 (lemmas with partial verbal information), 2998 (simple lemmas), and 3001 (raw words). Overall, across categories, the AIC differences between lemmas with full verbal information and all other operationalizations were between 9 and 15 points in favor of lemmas with full verbal information. These AIC differences greater than 5 points indicate a very strong difference in fit

with preference for the use of lemmas with full verbal information. These results indicate that more detailed information in the use of verbs is a stronger predictor of writing quality when assessing lexical diversity for L2 Spanish essays.

Regarding the lexical diversity index as calculated by MATTR, the results of the regression models indicated that the operationalization of lemmas tagged with part-of-speech and mood, tense, person, and number verbal information accounted for 23% of the variance in total scores of human judgments, while accounting for 25% and 29% of the variance in Vocabulary and Language Use scores, respectively. This shows an increase of 3 percentage points in explained variance for the total scores compared to MTLD, and 4 points for Vocabulary and Language Use scores compared to MTLD. Differences of 3-4 in the percentage of variance the models can explain (R^2) can be interpreted as moderately meaningful improvements in model fit (Cohen, 1988). Overall, MATTR scores with different operationalizations have small but meaningful differences in model performance. Total scores variance was accounted 21% with lemmas with partial verbal information, 21% with simple lemmas, and 22% with raw words. Similarly, Vocabulary scores variance was accounted 23% in all other instances that were not lemma with full verbal information; and Language Use scores variance was accounted 27% in all other instances that were not lemma with full verbal information as well. Finally, AIC scores for lemmas with full verbal information were lower in most categories when calculating MATTR. For total scores, lemmas with full verbal information scored lower (AIC = 2970) than lemmas with partial verbal information (AIC = 2981), simple lemmas (AIC = 2981), and raw words (AIC = 2975). For the Vocabulary category the differences ranged between 7 and 10 points, with lemmas with full verbal information (AIC = 1769) scoring lower than lemmas with partial verbal information (AIC = 1778), simple lemmas (AIC = 1780), and raw words (AIC = 1777).

Similarly, the Language Use category had AIC scores differences between 7 and 9 points, with lemmas with full verbal information (AIC = 1749) scoring lower than lemmas with partial verbal information (AIC = 1756), simple lemmas (AIC = 1759), and raw words (AIC = 1759). Finally, the AIC scores for these models indicate that a model with morphological operationalizations in the form of lemmas with full verbal information (tense, mood, person and number) is a better fit to predict writing quality in the form of human judgments when assessing lexical diversity for L2 Spanish essays.

Lastly, when comparing MTLD and MATTR R and AIC scores, there was a notable difference between models, particularly when assessing model fit differences with the operationalization of tokens as lemmas with full verbal information. For total scores, the AIC difference between MTLD (AIC=2987) and MATTR (AIC=2970) is of 17 points, which indicates a stronger fit in favor of MATTR. For the category of Vocabulary, the difference between MTLD (AIC=1786) and MATTR (AIC=1769) is also of 17 points, while the difference for the Language Use category is of 19 points between MTLD (AIC= 1768) and MATTR (AIC= 1749). These results indicate that a model with MATTR and lemmas that have full verbal information (mood, tense, person, and number) is a better fit to account for the variance in scores across all categories of human judgment scores, particularly Vocabulary and Language, which are closely related to the assessment of lexical diversity. This is in contrast to models that use MTLD as a measure of lexical diversity, and in contrast to operationalizations that are not lemmatized, or are lemmatized but do not have the full verbal information representative of the rich Spanish verbal inflectional system. Moreover, these results add to previous research in English language that assessed different operationalizations when calculating lexical diversity indices (e.g., Jarvis & Hashimoto, 2021). In contrast with these current results, previous research

in English showed lexical diversity indices were not significantly impacted by the type of operationalization as shown in this research study. This is most likely due to the difference in morphological richness that encode significant information in Spanish verbs compared to the more straightforward inflectional system for English verbs.

The implications of these results can inform the creation of multivariate models that attempt to predict writing quality scores or account for their variance in studies of L2 Spanish writing development and proficiency.

Study 2: Lexical Sophistication

In order to answer the second set of research questions, which examines the relationship between indices of lexical sophistication and different operationalizations of morphological normalization, a second study was conducted with the same learner corpus, the same reference corpus, and the same rater-scored essays as in the previous study, Study one. Similarly, the same linguistic analysis was carried out before calculating scores for measures of lexical sophistication. The following sections detail the methodology for calculating measures of lexical sophistication with different operationalizations of words, and the results.

Lexical sophistication

For study two, measures of word frequency, bigram strength of association (SOA), and bigram dependencies SOA measures were calculated four times with the same 400 sample essays and human judgment scores as study one. Word frequency was calculated for all words (content and function words), and for content words alone. This was done first with all raw words, then lemmas with part-of-speech information, then lemmas with half of verbal information, and finally lemmas with all verbal information. Furthermore, Bigram SOA scores were calculated in

the form of T-scores, MI scores, and ΔP scores (i.e. Delta P; Allan, 1980), with each of the four operationalizations mentioned before. Finally, bigram dependencies SOA scores were focused on Object-Verb, Adjective-Noun, Adverb-Verb dependencies. In order to analyze bigrams' syntactic dependencies, the essays were tokenized, and each token was tagged with their part-of-speech (POS) label. Following this step, the dependency parser SpaCy (Honnibal et al., 2020) was used to identify the syntactic dependencies in each bigram. For example, in "ella va a la escuela," "va" (verb) is the head of "ella" (subject), "a" is a preposition dependent on "va," and "escuela" is the object of the preposition. After this parsing, dependency-based bigrams were extracted (e.g., Verb-Preposition ("va", "a"); Preposition-Noun ("a", "escuela"). All four different operationalizations were also used for this measure, since bigram dependencies are still concerned with collocation as much as grammatical relationship between tokens.

Statistical analysis

In order to address the second set of research questions, the human judgment scores for the previous study that have been already assessed for inter-rater reliability and rater severity were used in this study as well. After securing these, the first step was to visualize frequency, strength of association, and dependency scores using R (R Core Team, 2016). Histograms were created for each variable to account for a normal distribution of the data. Secondly, scatter plots were used to visualize the linearity of each variable in reference to the human judgment scores. These two steps included only the lexical sophistication indices since the human judgement scores were previously checked for normality and linearity in the previous study. Finally, the cor() function (R Core Team, 2016) was used to examine the size of the correlations between the human judgments scores and different indices of lexical sophistication, separately.

After these steps were completed, multiple regressions were conducted to investigate the predictive ability of each operationalization of word frequency, bigram strength of association (SOA), and bigram dependencies SOA measures. For word frequency, there is a score for all words, and a score for only content words. For bigram SOA there are T-scores, MI scores, and Delta P scores. And for bigram dependencies SOA there are also T-scores, MI scores, and Delta P scores for Object dependencies, Adjective dependencies, and Adverb dependencies. For each index, the predictive power of each operationalization was compared by contrasting the models on R^2 and Akaike's Information Criterion (AIC). Higher R^2 and lower AIC values were interpreted as indicators of better model fit (Burnham & Anderson, 2002; Field, 2013). For this study, AIC differences greater than 2 points are considered indicative of meaningful differences in model quality (Burnham & Anderson, 2002).

Results

The initial analysis of distribution of human judgment and lexical sophistication scores showed that the data did not strongly deviate from normal distribution. Additionally, all operationalizations of the indices showed a relative linear relationship with the human judgment scores. To examine the extent to which lexical sophistication predicts essay quality, linear regressions were conducted with word frequency, bigram SOA, and bigram dependency SOA scores as predictors, and human judgment scores as outcomes. Models were tested across four morphological operationalizations: raw words, lemmas, lemmas tagged with tense and mood (TM), and lemmas tagged with tense, mood, person, and number (TMPN). In this section, human judgment scores for the categories of Vocabulary and Language Use are reported along with the Total score; while scores for Mechanics, Content, and Organization are not reported since these

are not constructs evaluated by indices of lexical sophistication. Moreover, to evaluate the relative fit of models using different operationalizations, Akaike Information Criterion (AIC) and R^2 values were compared between models. While AIC does not provide a statistical significance test, differences of 2 or more points between models are commonly interpreted as meaningful, and differences of 10 or more points as strong indications of improved model fit (Burnham & Anderson, 2002). Similarly, R^2 values that cross benchmark thresholds (from small = .01, to medium = .09, to large = .25) represent improvements in model fit (Cohen, 1988). The following sections provide these results by each lexical sophistication category.

Word Frequency. To examine the extent to which word frequency can predict essay quality, a series of linear regressions were conducted with the algorithmic frequency scores of all words and only content words, separately, as predictor variables and human judgment scores for the categories of Language Use, Vocabulary, and Total scores as outcomes. Models were tested across four morphological operationalizations: raw words, lemmas, lemmas tagged with tense and mood (TM), and lemmas tagged with tense, mood, person, and number (TMPN). Based on comparisons of AIC values and R^2, the lemma operationalization with part of speech tags but no verbal information consistently showed the best fit across all categories of measurement (Vocabulary, Language Use, Total scores) for content words, and a mix of simple lemmas and lemmas with partial verb information for all words. Tables 15-17 below show consistently higher R^2 and lower AIC for all categories with the simple lemma operationalization for all words frequency scores, except for the category of Language Use that scored a higher R2 and lower AIC for the lemma operationalization with partial verbal information.

Table 15

All words frequency regression models comparison for total scores

Category	Operationalization	β (Beta)	R^2	AIC
Total	Raw	-0.1430	0.000017	3076.58
Total	Lemma	3.2660	0.006580	3073.95
Total	Lemma + Tense/Mood	-0.4570	0.000170	3076.52
Total	Lemma + TMPN	1.2480	0.002200	3075.71

Table 16

All words frequency regression models comparison for Vocabulary

Category	Operationalization	β (Beta)	R^2	AIC
Vocabulary	Raw	-0.0150	0.0000035	1883.19
Vocabulary	Lemma	0.5920	0.004280	1881.48
Vocabulary	Lemma + Tense/Mood	-0.0350	0.000019	1883.18
Vocabulary	Lemma + TMPN	0.2400	0.001610	1882.55

Table 17

All words frequency regression models comparison for Language Use

Category	Operationalization	β (Beta)	R^2	AIC
Language Use	Raw	-0.2750	0.00126	1883.13
Language Use	Lemma	0.3320	0.00134	1883.10
Language Use	Lemma + Tense/Mood	-0.4700	0.00348	1882.24
Language Use	Lemma + TMPN	-0.0730	0.00015	1883.57

These summarized tables also show different β (Beta) scores, including some of them in the negative range. Since the β (Beta) scores are small and the R^2 scores are very small, these numbers are not meaningful in this context, but they are reported for data transparency. Similarly, as R^2 scores are very small, AIC scores are a good guide to assess which model is a better fit to account for the variance in human judgment scores in relationship with word frequency scores. In this case, all categories show a slightly lower AIC score (between 0.5 to 2.6) for simple lemmas with POS tags when contrasted with all other operationalizations, except for the Language Use category that has a lower AIC ($\Delta AIC = 0.8$) for the lemma operationalization with partial verbal information.

Furthermore, results for content words consistently showed higher R^2 and lower AIC for all categories with the operationalization that includes simple lemmas with POS tags. However, in this case, the difference in R^2 and AIC scores between the models with simple lemmas and all

other operationalization were notably bigger. Tables 18-20 below summarize this information clearly.

Table 18

Content words frequency regression models comparison for total scores

Category	Operationalization	β (Beta)	R^2	AIC
Total	Raw	0.981	0.001340	3076.06
Total	Lemma	5.1250	0.032500	3063.40
Total	Lemma + Tense/Mood	1.8680	0.004840	3074.65
Total	Lemma + TMPN	1.1000	0.006760	3073.88

Table 19

Content words frequency regression models comparison for Vocabulary

Category	Operationalization	β (Beta)	R^2	AIC
Vocabulary	Raw	0.472	0.006100	1880.74
Vocabulary	Lemma	1.1570	0.032700	1869.89
Vocabulary	Lemma + Tense/Mood	0.6400	0.011220	1878.67
Vocabulary	Lemma + TMPN	0.3270	0.011790	1878.44

Table 20

Content words frequency regression models comparison for Language Use

Category	Operationalization	β (Beta)	R^2	AIC
Language Use	Raw	0.049	0.00007	1883.61
Language Use	Lemma	0.9230	0.02080	1875.24
Language Use	Lemma + Tense/Mood	0.1910	0.00100	1883.23
Language Use	Lemma + TMPN	0.2810	0.00872	1880.13

Lastly, a comparison of frequency scores for all words and content words models was carried out. Table 21 below shows consistently higher R^2 and lower AIC scores for content words across categories.

Table 21

Lemma Operationalization comparison between All Words and Content Words

Category	R^2 All Words Lemma	AIC All Words Lemma	R^2 Content Words Lemma	AIC Content Words Lemma
Language Use	0.00100	1883.10	0.0208	1875.24
Vocabulary	0.00428	1881.48	0.0327	1869.89
Total	0.00700	3073.95	0.0325	3063.40

Bigram SOA. To examine the extent to which bigram SOA measurements can predict essay quality, a series of linear regressions were conducted with T-scores, MI scores, and Delta P scores as predictor variables and human judgment scores for the categories of Vocabulary, Language Use, Total scores as outcomes. Models were tested across four morphological operationalizations: raw words, lemmas, lemmas tagged with tense and mood (TM), and lemmas tagged with tense, mood, person, and number (TMPN). Based on comparisons of AIC values and R^2, the raw word operationalization consistently showed the best fit across all categories of measurement (Vocabulary, Language Use, Total scores) for T-scores, MI scores, and Delta P scores, with some small exceptions for the right to left direction of Delta P when calculating SOA using the second word of the bigram as a predictor of the preceding word. Tables 22-33 below show consistently higher R^2 and lower AIC for all categories with the raw word operationalization for all SOA scores.

Table 22

T'-Score Regression models comparison for total scores

Category	Operationalization	β (Beta)	R^2	AIC
Total	Raw	0.0422	0.0737	3046
Total	Lemma	0.0253	0.0337	3062.9
Total	Lemma + Tense/Mood	0.026	0.0383	3061
Total	Lemma + TMPN	0.00462	0.0463	3057.6

Table 23

T'-Score Regression models comparison for Vocabulary

Category	Operationalization	β (Beta)	R^2	AIC
Vocabulary	Raw	0.00723	0.0427	1865.7
Vocabulary	Lemma	0.00419	0.0183	1875.8
Vocabulary	Lemma + Tense/Mood	0.00427	0.0204	1874.9
Vocabulary	Lemma + TMPN	0.00081	0.0282	1871.8

Table 24

T'-Score Regression models comparison for Language Use

Category	Operationalization	β (Beta)	R^2	AIC
Language Use	Raw	0.00836	0.057	1860.2
Language Use	Lemma	0.00521	0.0283	1872.2
Language Use	Lemma + Tense/Mood	0.00536	0.032	1870.6
Language Use	Lemma + TMPN	0.00092	0.0363	1868.8

Table 25

MI Regression models comparison for total scores

Category	Operationalization	β (Beta)	R^2	AIC
Total	Raw	6.072	0.045	3058.2
Total	Lemma	4.816	0.0147	3070.7
Total	Lemma + Tense/Mood	6.293	0.0302	3064.3
Total	Lemma + TMPN	2.565	0.0373	3061.4

Table 26

MI Regression models comparison for total Vocabulary

Category	Operationalization	β (Beta)	R^2	AIC
Vocabulary	Raw	1.359	0.0446	1864.9
Vocabulary	Lemma	1.239	0.0192	1875.5
Vocabulary	Lemma + Tense/Mood	1.521	0.0349	1869
Vocabulary	Lemma + TMPN	0.569	0.0362	1868.4

Table 27

MI Regression models comparison for total Language Use

Category	Operationalization	β (Beta)	R^2	AIC
Language Use	Raw	1.606	0.0622	1857.9
Language Use	Lemma	1.384	0.0239	1874
Language Use	Lemma + Tense/Mood	1.865	0.0524	1862.1
Language Use	Lemma + TMPN	0.637	0.0454	1865.1

The calculation of Delta P from left to right, utilizing the first word to predict the probability of the second word in the bigram occurring, showed consistently low AIC scores across all categories for the raw words operationalization.

Table 28

Delta P1 Regression models comparison for total scores

Category	Operationalization	β (Beta)	R^2	AIC
Total	Raw	170.8	0.0313	3063.9
Total	Lemma	123	0.0204	3068.4
Total	Lemma + Tense/Mood	120.1	0.0217	3067.8
Total	Lemma + TMPN	18.05	0.0079	3073.4

Table 29

Delta P1 Regression models comparison for Vocabulary

Category	Operationalization	β (Beta)	R^2	AIC
Vocabulary	Raw	37.63	0.0301	1871
Vocabulary	Lemma	25.72	0.0176	1876.1
Vocabulary	Lemma + Tense/Mood	24.56	0.0179	1876
Vocabulary	Lemma + TMPN	3.16	0.0048	1881.3

Table 30

Delta P1 Regression models comparison for Language Use

Category	Operationalization	β (Beta)	R^2	AIC
Language Use	Raw	38.37	0.0312	1870.9
Language Use	Lemma	28.38	0.0214	1875
Language Use	Lemma + Tense/Mood	28.14	0.0235	1874.1
Language Use	Lemma + TMPN	3.41	0.0056	1881.4

However, for the calculation of Delta P from right to left, utilizing the second word to predict the probability of the first word in the bigram occurring, showed the lowest AIC for simple lemmas for Total scores and Vocabulary. Language Use showed very similar AIC scores across all operationalizations.

Table 31

Delta P2 Regression models comparison for total scores

Category	Operationalization	β (Beta)	R^2	AIC
Total	Raw	6.98	0.0001	3076.5
Total	Lemma	-51.53	0.0052	3074.5
Total	Lemma + Tense/Mood	-13.64	0.0004	3076.4
Total	Lemma + TMPN	6.89	0.0017	3075.9

Table 32

Delta P2 Regression models comparison for Vocabulary

Category	Operationalization	β (Beta)	R^2	AIC
Vocabulary	Raw	-0.64	2E-05	1883.2
Vocabulary	Lemma	-8.32	0.0027	1882.1
Vocabulary	Lemma + Tense/Mood	-0.5	1E-05	1883.2
Vocabulary	Lemma + TMPN	1.39	0.0014	1882.6

Table 33

Delta P2 Regression models comparison for Language Use

Category	Operationalization	β (Beta)	R^2	AIC
Language Use	Raw	2.73	0.0004	1883.5
Language Use	Lemma	-10.05	0.0039	1882.1
Language Use	Lemma + Tense/Mood	0.73	2E-05	1883.6
Language Use	Lemma + TMPN	2.78	0.0054	1881.5

Finally, a comparison of all SOA indices was carried out. Table 34 below shows consistently higher R^2 and lower AIC scores for MI for the Vocabulary and Language Use.

Table 34

Raw word Operationalization comparison between all SOAs

Category	R^2 T-Score Raw Words	AIC T-Score Raw Words	R^2 MI Raw Words	AIC MI Raw Words	R^2 DeltaP1 Raw Words	AIC DeltaP1 Raw Words	R^2 DeltaP2 Raw Words	AIC DeltaP2 Raw Words
Language Use	0.057	1860.17	0.062	1857.94	0.031	1870.94	0.01	1877.73
Vocabulary	0.043	1865.74	0.045	1864.94	0.030	1870.98	0.03	1872.23
Total	0.074	3045.97	0.045	3058.16	0.031	3063.85	0.03	3064.98

Bigram dependencies. To examine the extent to which bigram dependencies SOA measurements can predict essay quality, a series of linear regressions were conducted with T-scores, MI scores, and Delta P scores as predictor variables and human judgment scores with the categories of Vocabulary, Language Use, Total scores as outcomes, for object dependencies, adverb dependencies, and adjective dependencies. All models were tested across four morphological operationalizations: raw words, lemmas, lemmas tagged with tense and mood (TM), and lemmas tagged with tense, mood, person, and number (TMPN).

Regarding bigram Object dependencies, based on comparisons of AIC scores and R^2 values, the raw operationalization showed the best fit for the Vocabulary and Language Use categories for T-scores, while the lemma operationalization with partial verbal information showed the best AIC scores and R^2 values for MI scores. For Delta P scores, the simple lemma operationalization was a better fit for the Language Use category, and the lemma operationalization with full verbal information showed a better fit for the Vocabulary category when calculated from left to right, utilizing the first word to predict the probability of the second word in the bigram occurring. While for Delta P scores calculated from right to left, the simple lemma operationalization showed lower AIC scores across all categories. Tables 35-38 below show R^2 values and AIC scores for all categories and their operationalization for all SOA bigram object dependencies scores.

For adverb dependencies, the raw words and lemma with partial verbal information operationalizations had negligent differences and both had the lower AIC scores and higher R^2 scores in the Vocabulary and Language Use categories when the model used T-score. For MI scores, lemmas for the Vocabulary category and lemmas with partial verbal information for the

Language Use category had the lowest AIC, highest R^2, and showed very similar AIC scores with minimal and not meaningful differences. Furthermore, Delta P scores from left to right smaller AIC and higher R^2 for the model with lemmas with partial verbal information for Language Use, and smaller AIC and higher R^2 values for the model with raw word for the Vocabulary category. For Delta P calculated from right to left, the simple lemma operationalization has lower AIC scores and higher R2 values across categories. Details for adverb dependencies can be seen in Tables 39-42 below

Table 35

T-scores Object dependencies summary

Category	β Raw	R^2 Raw	AIC Raw	β Lemma	R^2 Lemma	AIC Lemma	β Lemma +TM	R^2 Lemma+ TM	AIC Lemma +TM	β Lemma+ TMPN	R^2 Lemma +TMPN	AIC Lemma +TMPN
Language Use	-0.049	0.028	1872.42	-0.017	0.006	1881.27	-0.039	0.024	1873.83	-0.033	0.010	1879.65
Vocabulary	-0.051	0.031	1870.70	-0.027	0.015	1877.25	-0.044	0.031	1870.75	-0.025	0.006	1880.84
Total	-0.281	0.047	3057.42	-0.140	0.020	3068.53	-0.243	0.048	3056.83	-0.106	0.005	3074.51

Table 36

MI scores Object dependencies summary

Category	β Raw	R^2 Raw	AIC Raw	β Lemma	R^2 Lemma	AIC Lemma	β Lemma +TM	R^2 Lemma+ TM	AIC Lemma +TM	β Lemma+ TMPN	R^2 Lemma +TMPN	AIC Lemma +TMPN
Language Use	0.213	0.007	1880.72	0.178	0.003	1882.44	0.367	0.016	1877.36	-0.014	0.000	1883.62
Vocabulary	0.064	0.001	1882.93	0.053	0.000	1883.08	0.192	0.004	1881.49	0.114	0.001	1882.74
Total	0.448	0.002	3075.94	0.260	0.000	3076.46	1.162	0.008	3073.41	0.243	0.000	3076.49

Table 37

DeltaP1 Object dependencies summary

Category	β Raw	R^2 Raw	AIC Raw	β Lemma	R^2 Lemma	AIC Lemma	β Lemma +TM	R^2 Lemma+ TM	AIC Lemma +TM	β Lemma+ TMPN	R^2 Lemma +TMPN	AIC Lemma +TMPN
Language Use	0.811	0.000	1883.61	3.358	0.002	1882.83	0.073	0.000	1883.63	1.710	0.000	1883.61
Vocabulary	3.621	0.001	1882.70	2.178	0.001	1882.85	1.930	0.000	1883.06	18.996	0.006	1880.65
Total	-2.901	0.000	3076.57	3.931	0.000	3076.53	-12.455	0.001	3076.32	44.219	0.002	3075.89

Table 38

DeltaP2 Object dependencies summary

Category	β Raw	R^2 Raw	AIC Raw	β Lemma	R^2 Lemma	AIC Lemma	β Lemma +TM	R^2 Lemma+ TM	AIC Lemma +TM	β Lemma+ TMPN	R^2 Lemma +TMPN	AIC Lemma +TMPN
Language Use	-7.816	0.015	1877.73	-13.539	0.044	1865.84	-8.159	0.024	1873.91	-18.079	0.005	1881.77
Vocabulary	-10.610	0.027	1872.23	-15.410	0.056	1859.97	-10.109	0.037	1868.15	-13.141	0.003	1882.21
Total	-48.502	0.029	3064.98	-67.850	0.055	3053.81	-44.391	0.036	3061.92	-63.815	0.003	3075.41

Table 39

T'-scores Adv dependencies summary

Category	β Raw	R^2 Raw	AIC Raw	β Lemma	R^2 Lemma	AIC Lemma	β Lemma +TM	R^2 Lemma+ TM	AIC Lemma +TM	β Lemma+ TMPN	R^2 Lemma +TMPN	AIC Lemma +TMPN
Language Use	-0.055	0.039	1867.58	-0.031	0.030	1871.61	-0.044	0.039	1867.62	-0.005	0.004	1881.95
Vocabulary	-0.054	0.038	1867.72	-0.027	0.023	1874.09	-0.040	0.033	1869.66	-0.002	0.001	1882.82
Total	-0.215	0.031	3063.94	-0.130	0.027	3065.52	-0.175	0.031	3063.80	-0.012	0.001	3076.07

Table 40

MI scores Adv dependencies summary

Category	β Raw	R^2 Raw	AIC Raw	β Lemma	R^2 Lemma	AIC Lemma	β Lemma +TM	R^2 Lemma+ TM	AIC Lemma +TM	β Lemma+ TMPN	R^2 Lemma +TMPN	AIC Lemma +TMPN
Language Use	0.062	0.000	1883.52	-0.079	0.000	1883.49	0.126	0.001	1883.24	0.038	0.000	1883.44
Vocabulary	-0.052	0.000	1883.11	-0.173	0.002	1882.52	-0.055	0.000	1883.12	0.044	0.001	1882.93
Total	0.248	0.000	3076.50	-0.686	0.001	3076.05	0.225	0.000	3076.53	0.252	0.001	3076.15

Table 41

DeltaP1 scores Adv dependencies summary

Category	β Raw	R^2 Raw	AIC Raw	β Lemma	R^2 Lemma	AIC Lemma	β Lemma +TM	R^2 Lemma+ TM	AIC Lemma +TM	β Lemma+ TMPN	R^2 Lemma +TMPN	AIC Lemma +TMPN
Language Use	-72.964	0.008	1880.54	-3.773	0.000	1883.56	-67.491	0.009	1879.98	0.244	0.000	1883.63
Vocabulary	-41.562	0.003	1882.19	9.751	0.001	1882.72	-34.570	0.002	1882.23	5.259	0.002	1882.47
Total	-226.50	0.004	3075.08	15.201	0.000	3076.53	-231.99	0.005	3074.41	-3.860	0.000	3076.57

Table 42

DeltaP2 scores Adv dependencies summary

Category	β Raw	R^2 Raw	AIC Raw	β Lemma	R^2 Lemma	AIC Lemma	β Lemma +TM	R^2 Lemma+ TM	AIC Lemma +TM	β Lemma+ TMPN	R^2 Lemma +TMPN	AIC Lemma +TMPN
Language Use	-4.599	0.010	1879.42	-7.777	0.023	1874.21	-6.441	0.017	1876.88	-0.646	0.002	1883.03
Vocabulary	-5.423	0.015	1877.32	-8.122	0.025	1872.89	-7.783	0.024	1873.28	-0.272	0.000	1883.08
Total	-22.552	0.013	3071.45	-42.028	0.034	3062.56	-35.710	0.026	3066.02	-1.501	0.000	3076.42

Finally, for adjective dependencies only scores for the first three operationalization are available since the last operationalization (lemmas with full verb information) returned N/A for Beta and R2 values due to singularities in the calculations. This is most likely due to the predictor variables for these calculations being constant and not possible to be estimated. However, all other calculations were possible. The differences between AIC and R^2 scores for bigram dependencies SOAs for adjective dependencies showed to be very small across more Vocabulary and Language Use categories. Overall, they all point to raw words as the better fit for explaining variance in human judgment scores when it comes to T-scores and MI scores. For Delta P, models with raw words provided the highest R2 values and lower AIC scores for Delta P from left to right for both Vocabulary and Language Use. However, for Delta P from right to left, the model with raw words showed higher R2 values and lower AIC only for Vocabulary, while for Language Use the lemma operationalization with partial verbal information showed a higher R2 and a lower AIC. Tables 43-46 show this in more detail.

Table 43

T'-scores Adj dependencies summary

Category	β Raw	R^2 Raw	AIC Raw	β Lemma	R^2 Lemma	AIC Lemma	β Lemma +TM	R^2 Lemma +TM	AIC Lemma +TM	β Lemma+ TMPN	R^2 Lemma+ TMPN	AIC Lemma +TMPN
Language Use	0.026	0.016	1877.29	0.015	0.005	1881.57	0.015	0.005	1881.58	NA	NA	NA
Vocabulary	0.020	0.009	1879.50	0.014	0.005	1881.26	0.014	0.005	1881.25	NA	NA	NA
Total	0.067	0.013	3071.60	0.066	0.005	3074.47	0.066	0.005	3074.47	NA	NA	NA

Table 44

MI scores Adj dependencies summary

Category	β Raw	R^2 Raw	AIC Raw	β Lemma	R^2 Lemma	AIC Lemma	β Lemma +TM	R^2 Lemma +TM	AIC Lemma +TM	β Lemma+ TMPN	R^2 Lemma+ TMPN	AIC Lemma +TMPN
Language Use	0.130	0.008	1880.45	-0.042	0.001	1883.41	-0.041	0.001	1883.42	NA	NA	NA
Vocabulary	0.105	0.005	1881.11	-0.054	0.001	1882.82	-0.053	0.001	1882.84	NA	NA	NA
Total	0.389	0.004	3075.15	-0.392	0.002	3075.61	-0.385	0.002	3075.64	NA	NA	NA

Table 45

DeltaP1 scores Adj dependencies summary

Category	β Raw	R^2 Raw	AIC Raw	β Lemma	R^2 Lemma	AIC Lemma	β Lemma +TM	R^2 Lemma +TM	AIC Lemma +TM	β Lemma+ TMPN	R^2 Lemma+ TMPN	AIC Lemma +TMPN
Language Use	2.821	0.003	1882.89	3.820	0.003	1882.60	3.823	0.003	1882.60	NA	NA	NA
Vocabulary	1.751	0.001	1882.71	2.282	0.001	1882.82	2.298	0.001	1882.82	NA	NA	NA
Total	6.577	0.001	3076.25	13.027	0.002	3075.98	13.052	0.002	3075.98	NA	NA	NA

Table 46

DeltaP2 scores Adj dependencies summary

Category	β Raw	R^2 Raw	AIC Raw	β Lemma	R^2 Lemma	AIC Lemma	β Lemma +TM	R^2 Lemma +TM	AIC Lemma +TM	β Lemma+ TMPN	R^2 Lemma+ TMPN	AIC Lemma +TMPN
Language Use	6.242	0.010	1879.58	6.242	0.010	1879.42	6.262	0.011	1879.42	NA	NA	NA
Vocabulary	5.014	0.009	1879.42	2.985	0.002	1882.27	3.045	0.003	1882.20	NA	NA	NA
Total	6.577	0.001	3073.05	16.382	0.004	3075.18	16.576	0.004	3075.18	NA	NA	NA

Discussion

This second study investigated the relationship between different operationalizations of morphological normalization when calculating indices of lexical sophistication for 400 L2 Spanish essays, and their relationship with human judgment scores. To explore the relationship between human judgement scores and the different indices of lexical sophistication and their operationalization, word frequency, bigram strength of association (SOA) and bigram dependencies' SOAs scores were calculated for all 400 essays with four different operationalizations: raw words, lemmas tagged with part-of-speech, lemmas tagged with part-of-speech and mood and tense verbal information, and lemmas tagged with part-of-speech and mood, tense, person, and number verbal information. After assessing normality and linearity, linear models were created to analyze the relationship between each of these indices' operationalizations and human judgment scores of writing quality that assess five categories: content, organization, language use, vocabulary and mechanics. The results, discussed in more detail below, have implications for the operationalization of lexical sophistication in multivariate models to predict L2 writing quality in Spanish.

Word Frequency. The results of the regression models indicated that the algorithmic frequency of all words (including content and function words) with the operationalization of lemmas tagged with part-of-speech accounted for 0.6% of the variance in total scores of human judgments, while accounting for 0.4% and 0.1% of the variance in Vocabulary and Language Use scores, respectively. The difference between operationalizations, as assessed by Akaike's Information Criterion (AIC), was very small across operationalizations and categories with differences in AIC scores between 1 and 2.5 points. AIC differences greater than 2 points are considered indicative of meaningful differences in model quality (Burnham & Anderson, 2002).

These differences are not substantial between operationalizations, which indicates that even though simple lemmas may be the operationalization for word frequency scores for all words that explain the most variance of human judgment scores, any of the operationalizations explain roughly the same amount of variance.

Regarding the words frequency calculated only with content words, the results of the regression models indicated that the operationalization of lemmas tagged with part-of-speech accounted for 3% of the variance in total scores of human judgments, while accounting for 3% and 2% of the variance in Vocabulary and Language Use scores, respectively. This is an increase of 2.4% for the total scores compared to the calculation with all words, 2.6% and 2.9% for Vocabulary and Language Use scores, respectively, compared to all words calculation. The differences (lower than 3%) in the percentage of variance the models can explain is a practically meaningful difference. Finally, AIC scores for simple lemmas were lower when calculating word frequency only with content words. For total scores, simple lemmas scored lower (AIC = 3076) than lemmas with all verbal information (AIC = 3073), lemmas with partial verbal information (AIC = 3074), and raw words (AIC = 3063). For the Vocabulary category the differences ranged between 9 and 11 points, with simple lemmas (AIC = 1869) scoring the lower than lemmas with full verb information (AIC=1878), lemmas with partial verbal information (AIC = 1878), and raw words (AIC = 1880). Similarly, the Language Use category had AIC scores differences between 5 and 8 points, with simple lemmas (AIC = 1875) scoring the lower than lemmas with full verb information (AIC=1880), lemmas with partial verbal information (AIC = 1883), and raw words (AIC = 1883). These results show an AIC difference between 10 and 12 points for the total score category, a difference between 8 and 10 points for the Vocabulary category, and a difference between 5 and 8 for the Language Use category, all favoring simple lemmas over

other operationalizations. Similarly, the models with simple lemmas explain between 1.2-2.8% more variance than all other operationalizations according to R2 values. These differences cross Cohen's (1988) benchmarks for small to medium effects, making them practically meaningful. This indicates that the simple lemmas operationalization when calculating word frequency score for content words explains more variance in human judgment scores than the other operationalizations: raw words, and lemmas with full or partial verbal information.

Lastly, when comparing all words and content words R2 and AIC scores, there was a difference between models, particularly when assessing models fit differences with the operationalization of tokens as simple lemmas. For total scores, the AIC difference between all words (AIC=3073) and content words (AIC=3063) is 10 points, which indicates a stronger fit in favor of content words. For the category of Vocabulary, the difference between all words (AIC=1881) and content words (AIC=1869) is 12 points, while the difference for the Language Use category is also 12 points between all words (AIC= 1883) and content words (AIC= 1875). Furthermore, the model for content words explains between 2-2.8% more variance than the model for all words, which includes content and function words. These are moderately meaningful improvements in model fit, suggesting that removing function words enhances the predictive power of frequency scores (Cohen, 1988). This indicates that calculating word frequency scores with content words only and operationalizing tokens as simple lemmas with no verbal information provides a stronger prediction of writing quality for the 400 essays from the COWSL2H corpus analyzed in this study. Moreover, these results are in line with previous studies of L2 English that showed content words to better predictor of writing quality, commonly lemmatized (e.g., Bestgen & Granger, 2014; Kim et al., 2018).

Bigram SOA. The results of the regression models indicated that all T-score measurements with the raw words operationalization for bigrams accounted for 7% of the variance in total scores of human judgments, while accounting for 4% and 5% of the variance in Vocabulary and Language Use scores, respectively. Other operationalizations of T-score accounted for less than 4% (lemmas with full verbal information) or 3% (simple lemmas and lemmas with partial verbal information) of total scores of human judgements. The difference between operationalizations, as assessed by Akaike's Information Criterion (AIC), was from moderate to high across operationalizations and categories. For instance, for AIC scores for Total scores were between 11 and 16 points lower for raw words (AIC=3046) than simple lemmas (AIC=3062), or lemmas with partial (AIC=3061) and full verbal information (AIC=3057). For the Vocabulary category the differences ranged between 6 and 10 points, with raw words (AIC = 1865) scoring lower than simple lemmas (AIC = 1875), lemmas with partial verbal information (AIC = 1874), and lemmas with full verbal information (AIC = 1871). Similarly, the Language Use category had AIC scores differences between 8 and 12 points, with raw words (AIC = 1860) scoring lower than simple lemmas (AIC = 1872), lemmas with partial verbal information (AIC = 1870), and lemmas with full verbal information (AIC = 1868). These results show an AIC difference between 11 and 16 points for the total score category, a difference between 6 and 10 points for the Vocabulary category, and a difference between 8 and 12 for the Language Use category, all favoring raw words over other operationalizations. Similarly, the models with raw words explain between 1.4-4% more variance than all other operationalizations according to R2 values. These differences cross Cohen's (1988) benchmarks for small to medium effects, making them practically meaningful. This indicates that the raw word operationalization when

calculating T-scores for bigrams SOA explains more variance in human judgment scores than the other operationalizations: simple lemmas, and lemmas with full or partial verbal information.

Regarding MI scores, the results of the regression models indicated that the raw word operationalization for bigrams accounted for 4.5% of the variance in total scores of human judgments, while accounting for 4.5% and 6% of the variance in Vocabulary and Language Use scores of human judgments, respectively. This is a decrease of 2.5% for the total scores compared to the model using T-scores, and a 0.5% and 1% increase for Vocabulary and Language Use scores, respectively. Other operationalizations of MI accounted for 3.7% (lemmas with full verbal information), 3% (lemmas with partial verbal information), and 1.5% (simple lemmas) of total scores of human judgements. The difference between operationalizations, as assessed by Akaike's Information Criterion (AIC), was from moderate to high across operationalizations and categories. AIC scores for Total scores were between 2 and 12 points lower for raw words (AIC=3058) than simple lemmas (AIC=3070), or lemmas with partial (AIC=3064) and full verbal information (AIC=3061). For the Vocabulary category the differences ranged between 4 and 11 points, with raw words (AIC = 1864) scoring lower than simple lemmas (AIC = 1875), lemmas with partial verbal information (AIC = 1869), and lemmas with full verbal information (AIC = 1868). Similarly, the Language Use category had AIC scores differences between 5 and 18 points, with raw words (AIC = 1857) scoring lower than simple lemmas (AIC = 1874), lemmas with partial verbal information (AIC = 1862), and lemmas with full verbal information (AIC = 1865). these results show an AIC difference between 3 and 12 points for the total score category, a difference between 3 and 10 points for the Vocabulary category, and a difference between 7 and 16 for the Language Use category, all favoring raw words over other operationalizations. Similarly, the models with raw words explain between 0.7-

3% more variance than all other operationalizations according to R2 values. Some of these differences cross Cohen's (1988) benchmarks for small to medium effects, making them practically meaningful. The operationalization with closest AIC scores and R2 values is the lemma operationalization with full verbal information. This indicates that the raw word operationalization when calculating MI scores for bigrams SOA explains more variance in human judgment scores than the other operationalizations: simple lemmas, and lemmas with full or partial verbal information. However, the differences in less than .01 on R2 values and less than 4 points for AIC scores, indicates that the lemma operationalization with full verbal information may explain a similar amount of variance when calculating MI scores.

Regarding Delta P scores calculated from left to right, with the first item of the bigram as a cue, the results of the regression models indicated that the raw word operationalization for bigrams accounted for 3% of the variance in total scores, Vocabulary, and Language Use scores of human judgments. This is a decrease of 4% and 1.5% for the total scores compared to the models using T-scores and MI, respectively. Additionally, it is a decrease of between 1-1.5% and 2-6% for Vocabulary and Language Use scores. For all other categories, the model accounted for less than 2% of the variance. Other operationalizations of Delta P from left to right accounted for 0.7% (lemmas with full verbal information), 2% (lemmas with partial verbal information), and 2% (simple lemmas) of total scores of human judgements. The difference between operationalizations, as assessed by Akaike's Information Criterion (AIC), was from moderate to high across operationalizations and categories. AIC scores for Total scores were between 4 and 10 points lower for raw words (AIC=3063) than simple lemmas (AIC=3068), or lemmas with partial (AIC=3067) and full verbal information (AIC=3073). For the Vocabulary category the differences ranged between 5 and 10 points, with raw words (AIC = 1871) scoring

lower than simple lemmas (AIC = 1876), lemmas with partial verbal information (AIC = 1876), and lemmas with full verbal information (AIC = 1881). Similarly, the Language Use category had AIC scores differences between 4 and 11 points, with raw words (AIC = 1870) scoring lower than simple lemmas (AIC = 1875), lemmas with partial verbal information (AIC = 1874), and lemmas with full verbal information (AIC = 1881). These results show an AIC difference between 4 and 10 points for the total score category, a difference between 5 and 10 points for the Vocabulary category, and a difference between 3 and 10 for the Language Use category, all favoring raw words over other operationalizations. Similarly, the models with raw words explain between 1-2.5% more variance than all other operationalizations according to R2 values. Some of these differences cross Cohen's (1988) benchmarks for small to medium effects, making them practically meaningful. This indicates that the raw words operationalization when calculating Delta P score from left to right explains more variance in human judgment scores than the other operationalizations: simple lemmas, and lemmas with full or partial verbal information.

Regarding Delta P scores calculated from right to left, with the second item of the bigram as a cue, the results of the regression models indicated that the simple lemma operationalization for bigrams accounted for 0.5% of the variance in total scores of human judgments, while accounting for 0.3% for Vocabulary. However, for Language Use scores the lemma operationalization with full verbal information was higher than the simple lemma, accounting for 0.5% in contrast to 0.4%, indicating that either operationalization would fit the model well. Overall, this model accounts for very little compared to the models for Delta P from left to right (3%), MI (4.5%), and T-scores (7%). Other operationalizations of Delta P from right to left accounted for 0.2% (lemmas with full verbal information), 0.04% (lemmas with partial verbal information), and 0.001% (raw words) of total scores of human judgements. The difference

between operationalizations, as assessed by Akaike's Information Criterion (AIC), was small across operationalizations and categories. AIC scores for Total scores were between 1 and 2 points lower for simple lemmas (AIC=3074) than raw words (AIC=3076), or lemmas with partial (AIC=3076) and full verbal information (AIC=3075). For the Vocabulary category the differences ranged between 1 or less points, with simple lemmas (AIC = 1882.1) scoring slightly lower than raw words (AIC = 1883.2), lemmas with partial verbal information (AIC = 1883.2), and lemmas with full verbal information (AIC = 1882.6). Similarly, the Language Use category had AIC scores differences between 1 or less points, with simple lemmas (AIC = 1882.1) scoring slightly lower than raw words (AIC = 1883.5), lemmas with partial verbal information (AIC = 1883.6), and lemmas with full verbal information (AIC = 1881.5). results show the simple lemma operationalization to be a better fit when predicting variance of total scores and Vocabulary scores, but the lemma with full verbal information was a better fit to predict variance in the Language Use scores. Looking more closely, these results show an AIC difference between 1.4 and 2 points for the total score category, and a difference between 0.5 and 1 points for the Vocabulary category, both in favor of the simple lemma operationalization. However, results show a difference between 0.6 and 2 points for the Language Use category, in favor of the lemma operationalization with full verbal information. Similarly, the models with simple lemmas explain between 0.3-0.5% more variance than all other operationalizations according to R2 values for total scores. When looking at the Vocabulary and Language Use categories, the percentage of variance explained is very small (R2 between .002 and .003), and the differences between operationalizations are even smaller. These differences do not cross Cohen's (1988) benchmarks for small to medium-sized effects, making them practically not meaningful. This indicates that the lemma operationalization when calculating Delta P score from right to left may

explain the most variance, but the variance explained is very little, and the differences with other operationalizations are not practically meaningful.

When comparing different indices of strength of association in bigrams, the model comparison shows that, overall, raw words are a better operationalization for all indices, except when calculating Delta P from left to right, which model showed results in favor of using simple lemmas. However, even when using a different operationalization, Delta P scores did not account in any model for more than 3% of the variance in human judgment scores, which is less than the other indices. When comparing indices across the raw word operationalization, MI scores account for 4.5% of the variance in the essay's total scores, while the model using T-score can account for 7.3% of the variance. In contrast, Delta P can account only for 3% of the variance. For the Vocabulary category, the model with MI scores can account for 4.5% of the variance in human scores, and the model with T-score can account for 4.2%, while the model with Delta P can account for 3%. For the Language Use category, MI scores can account for 6.2% of the variance, T-score for 5.7%, and Delta P 3%. These results indicate that a model that calculates SOA bigram scores with MI using raw words as morphological operationalization is overall a better fit to account for the variance in scores across all categories of human judgment scores. However, because the differences in explained variance between MI and T-scores were small (0.2-0.3%), and much larger between MI and both Delta P calculations (1.5-5%), this indicates that calculating SOA for bigrams with MI or T-scores and operationalizing tokens as raw words provides a stronger prediction of writing quality for the 400 essays from the COWSL2H corpus analyzed in this study.

Bigram dependencies. The results of the regression models to calculate SOA scores for bigram Object dependencies indicated that most T-score measurements with the raw words

operationalization accounted for 4.7% of the variance in total scores of human judgments, while accounting for 3% and 2.8% of the variance in Vocabulary and Language Use scores, respectively. The lemma operationalization with partial verbal information accounted for a very similar amount of the variance, with AIC scores that differ between 0.5 and 1.5 points, which is a small and not practically meaningful difference. Other operationalizations of T-score accounted for less than 1% (simple lemmas) or 0.5% (lemmas with full verbal information) of total scores of human judgements. The difference between operationalizations, as assessed by Akaike's Information Criterion (AIC), was very small between raw words and lemmas with partial information, but moderate to high between those operationalization and simple lemmas and lemmas with full verbal information. For instance, for AIC scores for Total scores were between 1 and 17 points lower for raw words (AIC= 3057.42) than simple lemmas (AIC= 3068), or lemmas with partial (AIC= 3056.83) and full verbal information (AIC= 3074). For the Vocabulary category the differences ranged between less that 1and 10 points, with raw words (AIC = 1870.69) scoring lower than simple lemmas (AIC = 1877), lemmas with partial verbal information (AIC = 1870.74), and lemmas with full verbal information (AIC = 1880). Similarly, the Language Use category had AIC scores differences between 1 and 9 points, with raw words (AIC = 1872.42) scoring lower than simple lemmas (AIC = 1881), lemmas with partial verbal information (AIC = 1873.82), and lemmas with full verbal information (AIC = 1879). All other categories had between 8 and 14 points of difference for AIC scores between operationalization, always in favor of the use of raw words or lemmas with partial verbal information. This indicates that raw words or lemmas with partial verbal information are a good fit to explain the variance in human judgment scores when calculating T-score for object dependency bigrams' SOA.

Regarding MI scores, the model for Object dependencies indicated that MI scores with lemmas operationalization with partial verbal information accounted for 0.79% of the variance in total scores of human judgments, while accounting for 0.43% and 1.6% of the variance in Vocabulary and Language Use scores, respectively. For all other categories, the model accounted for 0.3-0.4% of the variance. Other operationalizations for total scores, Vocabulary, and Language Use accounted for less than 0.03% of the variance. The difference between operationalizations, as assessed by Akaike's Information Criterion (AIC), was small for all operationalizations for total scores, showing a lower AIC score for lemmas with particular information (AIC=) between 2.5 and 3 points lower than all other operationalization. However, for all other categories, the AIC difference between lemmas with particular verbal information and other operationalizations was between 1.4-1.6 points. The only exception is Language Use for which the AIC was between 3 and 6 points lower than all other operationalizations. This indicates that lemmas with partial verbal information are the best fit to explain the variance in human judgment scores when calculating MI scores for object dependency bigrams' SOA.

Finally, regarding models of Delta P scores for Object dependencies, models indicated that the lemma operationalization with full verbal information and the simple lemmas operationalization accounted for 0.17% (Delta P left-right) and 5.5% (Delta P right-left) of the variance for total scores of human judgments. The model for Delta P from left to right showed almost no difference between operationalizations across categories (AIC lower than 1), except for Vocabulary for which there are small but meaningful differences between operationalizations, favoring the lemma operationalization with full verbal information with an AIC score 2 points lower than all other operationalizations. The model for Delta P from right to left showed higher differences with all other operationalizations across all categories, with AIC

scores between 8 and 20 points lower for total scores, Vocabulary and Language Use when compared to all other operationalizations. These results indicate that lemmas or lemmas with full verb information are better fits for Delta P calculations, depending on their directionality.

Furthermore, the results of the regression models to calculate SOA scores for bigram Adverb dependencies indicated that T-scores with both raw words and lemma with partial verb operationalizations accounted for 3.1% of the variance in total scores of human judgments, while accounting for 3.7% and 3.3% of the variance in Vocabulary scores, respectively, and 3.9% of the variance in Language Use scores. The difference for Vocabulary and Language Use scores are moderate when compared to other operationalization, since both raw words and lemmas with partial verbal information have AIC scores 4 to 15 points lower than all other operationalization, while the AIC score between them is consistently less than 1.9 points. This indicates that when calculating T-scores for adverb bigram SOAs to account for the variance in human judgment scores, both raw words and lemmas with partial verb operationalizations are a good fit.

For MI scores, models indicated that even though the simple lemma and the lemma with partial verbal information showed higher R^2 and lower AIC scores, the differences between all operationalizations across all categories is less than 2 points for AIC scores. Meanwhile, the simple lemmas operationalization explains 0.13% of the variance for total scores of human judgements, 0.16% of the variance for Vocabulary Scores, and 0.03% of the variance for Language Use, for which the lemmas with partial verb information explains 0.09% of the variance. These results indicate that when calculating MI for adverb bigram SOA, lemmas or lemmas with partial verb information are a better fit to explain variance in human judgment scores.

Finally, for Delta P scores, models indicated that the lemma with partial verbal information explained 0.5% of the variance in total scores of human judgments, 0.2% of the variance in Vocabulary scores, and 0.9% of the variance in Language Use, when calculated from left to right. There are no practically meaningful differences across categories with any other operationalizations, except with simple lemmas, for which AIC are lower 3 points in favor of lemmas with partial verbal information in Language Use scores, and 2 points lower for total scores. For Delta P calculated from right to left, the simple lemma operationalization explained 1.3% of the variance in total scores of human judgments, 1.5% of the variance in Vocabulary scores, and 1% of the variance in Language Use scores. When compared to all other operationalizations, simple lemmas showed between 3- and 14-points lower AIC scores across all categories, except for Vocabulary with lemmas with partial verb information for which the AIC score was not meaningfully different (lower than 1 point). This indicates that simple lemmas and lemmas with particular verbal information are goof fits when calculating Delta P for adverb dependency bigrams to explain variance in human judgment scores, depending on the direction of the calculation.

Lastly, the results of the regression models to calculate SOA scores for bigram Adjective dependencies indicated that across all operationalization and across all categories the differences between operationalization were very small. Overall, the best fit for T-score and MI is raw words, with small differences for the Language Use category (between 3 and 4 points). For Delta P, left to right calculations showed almost identical AIC scores for all operationalizations across all categories, while Delta P calculated from right to left, showed 2-points lower AIC scores for raw words when looking at total scores and Vocabulary scores, and almost identical AIC scores when looking at Language Use scores. Overall, the SOA scores for Adjective bigram

dependencies only account for a very small percentage of variation of human judgment scores. T-scores account for 1% of variation in total scores, MI 0.3%, and both Delta P scores account for 0.09% of variation in total scores. These results indicate that when calculating SOA scores for bigram Adjective dependencies, the raw words operationalization is the best fit for all indices, and that T-score is the index that can explain the most variance in human judgment scores.

Overall, the amount of variance explained by SOA measures of bigram dependencies are not very large, and the differences between operationalizations tends to be very small. When looking at recommendations for multivariate approaches to modeling lexicogrammatical indices of lexical sophistication to predict human judgment scores, variance explanations were most commonly higher from T-scores, however, Delta P measures can capture directional dependency strength, which reflects the complex nature of natural languages better than T-scores, for example, which focuses on frequency of co-occurrence. Finally, regarding operationalization, the results showed R^2 and AIC scores that favored the use of raw words and simple lemmas the most. In this case the recommendation would be to use lemmas which have been more consistently strong for Delta P calculations, and not worse in performance for T-scores or MI.

4 LEXICAL RICHNESS AS PREDICTOR OF L2 SPANISH WRITING QUALITY: A MULTIVARIATE APPROACH

Multivariate approaches to exploring the relationship between the use of lexicogrammatical features and levels of L2 language proficiency or L2 writing quality have increased in the last 10 or 15 years, given the advancements in computational methods and growing availability of learner corpora. While early studies such as Biber's Multidimensional Analysis (1988) laid the foundation by incorporating co-occurring linguistic features into assessments of language use, the application of multivariate modeling specifically aimed at L2 writing development or the prediction of proficiency has been evolving much more rapidly since the 2000's. This is due to a combination of more access to large-scale learner corpora, not only in English but in other L2s as well (e.g., Granger et al., 2009; Lüdeling et al, 2008; Callies & Zaytseva, 2013), and an increased used of regression modeling and factor analysis that allowed for an easier comparison of models and predictors.

Regarding applied research, multivariate approaches to modeling the relationship of lexicogrammatical features and proficiency or writing quality have been most notably pursued in English as a second or foreign language (ESL/EFL). This research has shown that, on average, more proficient learners tend to use more sophisticated vocabulary, more commonly occurring and more strongly associated collocations in the form of bigrams and trigrams, and tend to use a larger vocabulary repertoire avoiding repetitiveness (Bestgen & Granger, 2014; Garner et al., 2019; Guo et al., 2013; Kim et al., 2018; Kyle et al., 2018; Kyle & Crossley, 2015; inter alia). However, much of this research has not been duplicated with other languages. This is an important step in the development of second language assessment tools, since different languages may have lexicogrammatical features that work differently than the ones in English,

prompting a revision of certain aspects of modeling or operationalization in order to better assess the markers of proficiency of different target languages.

Specifically for Spanish as a second language, research has focused, until very recently, on exploring isolated writing features as opposed to more comprehensive multivariate approaches (e.g., García Salido & Garcia, 2018; Tracy-Ventura, 2017; Tracy-Ventura et al., 2021, Schnur & Rubio, 2021). This research has contributed greatly to the understanding of second language Spanish development and has provided informative data about lexical sophistication, lexical diversity, lexical complexity and their individual relationships with proficiency and writing quality. However, to achieve a more comprehensive understanding of the lexicogrammatical features that drive L2 Spanish development and how they relate to proficiency levels or human judgments of writing quality, more research that utilizes multivariate approaches are necessary, not only with different target languages, but with different corpora for those target languages.

Recently, multivariate approaches to modeling the relationship between features of lexicogrammatical use and proficiency in L2 Spanish have emerged. For instance, Diez-Ortega and Kyle (2024), looked at the L2 Spanish development of lexical richness over time with a multivariate approach that included measures of lexical diversity, word frequency, bigram strength of association and psycholinguistic properties. This study did not look at the relationship with a particular proficiency measure or with human judgment scores of written essays, since the corpus the authors studied was LANGSNAP, which is comprised of oral interviews of Spanish learners over a period of 21 months. Nevertheless, the authors found the use of a multidimensional approach important when assessing Spanish L2 lexical development, highlighting that different lexical features may develop at different rates over time. Another

Spanish study with a multivariate approach (Consolini & Kyle, 2024) showed that almost 50% of variance in proficiency scores as operationalized by a university language placement test score could be explained by a multidimensional model of lexical richness that included lexical diversity measures, bigram strength of association scores, and verb frequency of use scores.

However, these studies have not examined the relationship between indices of lexical richness and human judgment scores of writing quality, and they have also not examined the potential differences between models when utilizing different operationalizations for morphological normalization. These two points are relevant to current studies of writing proficiency and development in L2 Spanish. Firstly, being able to model lexical richness indices to predict human judgments of writing quality could provide the initial stages for the development of automatic writing assessment tools in L2 Spanish at a similar scale of those already existing in English (e.g., TOEFL). Secondly, different operationalizations of tokens into lemmas, raw words, and lemmas with partial or full verbal information may provide relevant language-specific information, which can in turn create models that better fit the target language data. As discussed in previous chapter and as evidenced by the studies in chapter three, Spanish has a very rich inflectional verb system compared to English, and the operationalization of tokens into lemmas with full verbal information can provide better model predictions for lexical diversity indices.

This chapter examines the results of a multivariate approach to modeling lexical richness indices to predict human judgment scores of L2 Spanish writing quality. This multivariate approach is informed by the models of the two studies described in chapter 3. These studies examine which operationalizations of morphological normalization produced models that explained the most variance across human judgments scores of L2 Spanish written quality in the

categories of Vocabulary and Language Use when assessing 400 essays from the COWSL25H corpus. Additionally, this chapter is guided by the third research question, "To what extent are features of written lexical and lexicogrammatical use in the form of lexical sophistication and lexical diversity predictive of L2 Spanish writing quality in the form of human judgments?". Finally, the chapter is divided into three sections: the methods for the assembly of the final model to describe the relationship between indices of lexical richness and human judgment scores of writing quality, the results provided by the model, and a discussion of the results and their implications for future research.

Methods

Final Models

Two models were created and evaluated to answer the last research question, one to evaluate variance of human judgment scores for the Vocabulary category, and one to evaluate variance for the Language Use category. This decision was made based on the results from the models evaluated in chapter three and the criteria followed by raters when assessing the essays with the rubric separated into 5 categories: Content, Organization, Mechanics, Vocabulary and Language Use. As part of the raters' training, emphasis was placed on the difference between the Vocabulary and Language Use categories. As seen in the rubric in Appendix A, both categories share features that are evaluated by indices of lexical sophistication and diversity, but with different emphases. For the Vocabulary category, the focus is on a diverse use of lexical choices, advanced or sophisticated vocabulary, and idiomatic use of lexical items. These are best reflected in measures of lexical diversity, and lexical sophistication in the form of word frequency scores, primarily, and the assessment of bigrams. For the Language Use category, the focus is on word order, syntactic variety and complexity, and morphological accuracy, which is most notable in

Spanish in verb conjugation and number and gender agreement between adjectives and nouns. These are best reflected in measures of strength of association of bigrams, and bigram dependencies. If the learners used proper morphology to link lexicogrammatical items (e.g., verb-objects, adj-nouns), these would be assessed for strength of association against the reference corpus and could provide an explanation of the variance of this category's scores. In the following section, there is a summary of the variables chosen for the two final models.

Lexical Diversity. For lexical diversity measures, study one provided clear evidence that MATTR index and the lemma operationalization with full verbal information (mood, tense, person, and number) were the combination that provided the best fit for the model and that explained the most variance in the scores of both the Vocabulary and Language Use categories. While MTLD models with full verbal information explained 21.5% of the variance for Vocabulary scores, and 25% for Language Use scores, the MATTR models with the same operationalization explained 24.7% of the variance for Vocabulary scores, and 28.5% for Language Use. The differences between the models as assessed by AIC scores showed differences between 17 and 19 points in favor of the MATTR models. AIC differences greater than 2 points are considered indicative of meaningful differences in model quality (Burnham & Anderson, 2002). Regarding different operationalizations, the lemma operationalization with full verbal information consistently predicted 2 to 3 percentage points more than any other operationalization.

Word Frequency. Word frequency scores can help determine how sophisticated a word is based on how frequently it is used in the reference corpus. For this measurement, study two provided clear evidence that the content word calculation and the simple lemma operationalization were the combination that provided the best fit for the model and that

explained the most variance in the scores of both the Vocabulary and Language Use categories. While models with all words calculations and the simple lemma operationalization explained .4% of the variance for Vocabulary scores, and .1% for Language Use scores, the models with content words calculations and the same operationalization explained 3.3% of the variance for Vocabulary scores, and 2% for Language Use. The differences between the models as assessed by AIC scores showed differences between 8 and 12 points in favor of the models with content word calculations. AIC differences greater than 2 points are considered indicative of meaningful differences in model quality (Burnham & Anderson, 2002). Regarding different operationalizations, the lemma operationalization consistently predicted 2 to 3 percentage points more than any other operationalization when assessing content words.

Bigram SOA. Bigram strength of association measures differ in exactly how they calculate strength of association and what learner information that provides for assessment. While MI provides information on how strongly two words are associated relative to change, and T-score offers data on how reliably two words co-occur based on frequency and variance, Delta P measures use directional predictability and evaluate the bigrams bidirectionally by calculating what are the chances of word B occurring when preceded by word A, and the chances of word A occurring when followed by word B. Consequently, MI highlights rare but exclusive bigrams or word pairs, and can sometimes overestimate rare combinations. T-scores emphasize frequent and common collocations, but it may be less sensitive to rare but meaningful bigrams. And Delta P can capture the asymmetry in the associations, which can provide more fine-grained details on the learners' development trajectory. For these measures, study 2 provided suggestions on which operationalizations to use for each index, but all indices were used in the final models. For T-scores, the raw operationalization for bigrams accounted for 4% of the variance in Vocabulary

scores, and 5% in Language Use; and overall, the raw operationalization explained 2-3 percentage points more than other operationalizations. The differences between the models across T-scores operationalizations as assessed by AIC scores showed differences between 6 and 12 points in favor of the models with raw words when calculating T-scores. For MI scores, the raw operationalization for bigrams accounted for 4.5% of the variance in Vocabulary scores, and 6% in Language Use; and overall, the raw operationalization explained 1-4 percentage points more than other operationalizations. The differences between the models across MI scores operationalizations as assessed by AIC scores showed differences between 4 and 18 points in favor of the models with raw words when calculating MI scores. For Delta P scores calculated from left to right, the raw operationalization for bigrams accounted for 3% of the variance in both Vocabulary and Language Use; and overall, the raw operationalization explained 1-3 percentage points more than other operationalizations. The differences between the models across MI scores operationalizations as assessed by AIC scores showed differences between 4 and 11 points in favor of the models with raw words when calculating Delta P scores from left to right. Finally, for Delta P scores calculated from right to left, the simple lemma operationalization for bigrams accounted for 0.5% of the variance in both Vocabulary and 0.4% for Language Use; and overall, the simple lemma operationalization explained less than 1 percentage points more than other operationalizations. The differences between the models across MI scores operationalizations as assessed by AIC scores showed differences between 1 and 2 points in favor of the models with simple lemmas when calculating Delta P scores from right to left.

Bigram Dependency SOA. Three dependencies were evaluated for this category: Object dependencies, Adjective dependencies and Adverb dependencies. For each, T-scores, MI scores,

and Delta P scores were evaluated with all operationalization. Based on the results from study two, the differences across operationalizations were meaningful enough to make decisions for each SOA calculation; however, most calculations accounted for less than 1% of variance in the scores for Vocabulary and Language Use categories. An exception to this was T-scores which explained between 2-3% of variance for these categories across dependency types. Overall, T-scores showed to explain more variance in all dependency types when calculated with raw words. Mi scores explained more variance with the lemma operationalization for Object dependencies, with lemma with partial verbal information for Adverb dependencies, and with raw words for Adjective dependencies. Delta P calculation from left to right explained more variance with lemma for Object dependencies, and with lemma with partial verbal information for both Adverb and Adjective dependencies. Finally, Delta P calculation from right to left explained more variance with lemma for Object and Adverb dependencies, and raw words for Adjective dependencies. Below, Table 47 outlines all variables selected as predictors for final models based on the results of studies one and two.

Table 47

Summary of predictors for final model

Predictors	Operationalization
MATTR	Lemma with full verbal information
Content words frequency	Simple lemma
Bigram T-score	Raw words
Bigram MI	Raw words
Bigram Delta P1	Raw words
Bigram Delta P2	Simple lemma
Bigram Object T-score	Raw words
Bigram Object MI	Simple lemma

Bigram Object Delta P1	Lemma with partial verbal information
Bigram Object Delta P1	Simple lemma
Bigram Object Delta P2	Simple lemma
Bigram Adverb T-score	Raw words
Bigram Adverb MI	Lemma with partial verbal information
Bigram Adverb Delta P1	Lemma with partial verbal information
Bigram Adverb Delta P2	Simple lemma
Bigram Adjective T-score	Raw words
Bigram Adjective MI	Raw words
Bigram Adjective Delta P1	Lemma with partial verbal information
Bigram Adjective Delta P2	Raw words

Results

After checking for linearity with Vocabulary and Language Use scores, the T-score for Adjective dependencies and the T-score for bigrams were removed due to lack of linearity. Finally, a final linear regression model for Vocabulary scores and one for Language Use scores were created, each with the variables in Table 48 as predictors.

Table 48

Summary of predictors for final model

Predictors	Operationalization
MATTR	Lemma with full verbal information
Content words frequency	Simple lemma
Bigram MI	Raw words
Bigram Delta P1	Raw words
Bigram Delta P2	Simple lemma
Bigram Object T-score	Raw words
Bigram Object MI	Simple lemma

Bigram Object Delta P1	Lemma with partial verbal information
Bigram Object Delta P1	Simple lemma
Bigram Object Delta P2	Simple lemma
Bigram Adverb T-score	Raw words
Bigram Adverb MI	Lemma with partial verbal information
Bigram Adverb Delta P1	Lemma with partial verbal information
Bigram Adverb Delta P2	Simple lemma
Bigram Adjective MI	Raw words
Bigram Adjective Delta P1	Lemma with partial verbal information
Bigram Adjective Delta P2	Raw words

After removing the non-linear variables, Variance Inflation Factors (VIF) were calculated among all predictor variables to check for multicollinearity. All variables show low to moderate VIF values, indicating small to moderate correlations and no multicollinearity issues (Menard, 2001). Table 49 below shows the results in detail.

Table 49

Variance Inflation Factors (VIF) for each predictor in the full model

Predictor	VIF
MATTR (lemmaMTPN)	1.342
Content words frequency (lemma)	1.287
Bigram MI (raw)	1.501
Bigram Delta P1 (raw)	1.255
Bigram Delta P2 (lemma)	1.266
Bigram Object T-score (raw)	2.008
Bigram Object MI (lemma)	1.485
Bigram Object Delta P1 (lemmaMT)	3.214
Bigram Object Delta P1 (lemma)	3.011

Bigram Object Delta P2 (lemma)	1.783
Bigram Adverb T-score (raw)	1.624
Bigram Adverb MI (lemmaMT)	1.777
Bigram Adverb Delta P1 (lemmaMT)	1.403
Bigram Adverb Delta P2 (lemma)	1.884
Bigram Adjective MI (raw)	1.437
Bigram Adjective Delta P1 (lemmaMT)	1.300
Bigram Adjective Delta P2 (raw)	1.227

Finally, two linear models were created to examine the relationship between human judgment scores for the Vocabulary category and the indices analyzed in Table 50. Afterwards, the models were examined with the MuMIn::dredge() package to find the optimal model from the variables in the original model.

In the optimal model for Vocabulary scores, the predictor variables were MATTR, content words frequency, bigram Delta P from left to right, bigram Object dependency T-score, bigram Object dependency Delta P from right to left, bigram Object dependency Delta P from left to right, and bigram Adverb dependency T-scores. The model indicated that there is a significant relationship with a medium effect ($p < .001$, R^2 adjusted = .290), and explained approximately 29% of the variance in Vocabulary scores. Table 50 shows that 20% of the explained variance can be attributed to the MATTR index of lexical diversity, 7.2% to the bigram dependency SOA measures in the form of T-scores and Delta P scores, 1.7% to content word frequency, and 1.4% to bigram SOA scores in the form of Delta P from left to right.

Table 50

Optimal linear model for Vocabulary scores

Predictor	Relative importance	Estimate	SE	t	*p*
MATTR (lemmaMTPN)	0.201	20.728	2.114	9.805	0.001
Bigram Adverb T-score (raw)	0.023	-0.031	0.012	-2.594	0.010
Bigram Object Delta P2 (lemma)	0.021	0.074	3.544	0.021	0.983
Content words frequency (lemma)	0.017	0.531	0.282	1.884	0.060
Bigram Delta P1 (raw)	0.014	17.438	9.529	1.830	0.068
Bigram Object T-score (raw)	0.020	-0.052	0.017	-3.009	0.003
Bigram Object Delta P1 (lemmaMT)	0.006	13.694	5.392	2.540	0.011

In the optimal model for Language Use scores, the predictor variables were MATTR, content words frequency, bigram Object dependency T-score, Bigram Object MI, bigram Object dependency Delta P from left to right, bigram Object dependency Delta P from right to left, bigram Adverb dependency T-score, bigram Adverb dependency MI, and bigram Adjective dependency Delta P from right to left. The model indicated that there is a significant relationship with a medium effect ($p < .001$, R2 adjusted = .328), and explained approximately 32.8% of the variance in Language Use scores. Table 51 shows that 22% of the explained variance can be attributed to the MATTR index of lexical diversity, 10% to the bigram dependency SOA measures in the form of T-scores, MI, and Delta P scores, 1% to content word frequency.

Table 51

Optimal linear model for Language Use scores

Predictor	Relative importance	Estimate	SE	t	P
MATTR (lemmaMTPN)	0.222	21.758	2.129	10.220	0.001
Bigram Object MI (lemma)	0.031	0.575	0.287	2.007	0.045
Bigram Adverb T-score (raw)	0.028	-0.041	0.012	-3.374	0.001
Bigram Object T-score (raw)	0.017	-0.050	0.017	-2.986	0.003
Bigram Object Delta P2 (lemma)	0.015	2.445	3.401	0.719	0.473
Bigram Adjective Delta P2 (raw)	0.012	4.357	2.143	2.033	0.043
Content words frequency (lemma)	0.011	0.421	0.280	1.505	0.133
Bigram Adverb MI (lemmaMT)	0.004	0.326	0.185	1.765	0.078
Bigram Object Delta P1 (lemmaMT)	0.003	11.115	5.277	2.106	0.036

Discussion

This chapter examined the relationship between features of lexicogrammatical use and second language Spanish writing quality in the form of human judgment scores by assessing 400 essays written by L1 English speakers learning L2 Spanish. To explore the relationship between features of lexicogrammatical use and L2 Spanish writing quality in the form human judgment scores, different indices of lexicogrammatical use (MATTR scores, content word frequency, bigram SOAs and bigram dependencies SOAs) were calculated for all 400 L2 Spanish essays. The studies on the previous chapter provided the different operationalizations for the tokens to be analyzed by these indices of lexical richness in order to be able to create linear models that would leverage the operationalizations that explained the most variance in human judgment scores. After assessing normality, linearity, and collinearity of the indices calculated, and determining which operationalizations best fit each index, two linear models were created to

analyze the relationship between these indices and writing quality in the form of human judgment scores. Two final optimal models were reached to explain the variance in writing scores for the categories of Vocabulary and Language Use. For the Vocabulary category, the optimal model included MATTR, content words frequency, bigram Delta P from left to right, bigram Object dependency T-score, bigram Object dependency Delta P from right to left, bigram Object dependency Delta P from left to right, and bigram Adverb dependency T-scores. For the Language Use category, the optimal model included MATTR, content words frequency, bigram Object dependency T-score, Bigram Object MI, bigram Object dependency Delta P from left to right, bigram Object dependency Delta P from right to left, bigram Adverb dependency T-score, bigram Adverb dependency MI, and bigram Adjective dependency Delta P from right to left.

These results, discussed in more detail below, provide a window into how raters perceive the quality of L2 Spanish writing. While the findings do not reveal the cognitive or linguistic processes that learners engage in while composing their essays, they do shed light on the features that may influence a rater's perception during assessment. Because lexical richness indices are used here as predictors of human judgment scores, the goal is not to measure learner development directly, but to better understand which features (such as lexical diversity, collocation strength, grammatical dependencies, or verbal inflection) contribute to variation in rater-assigned scores. In turn, this can inform which aspects of L2 writing are more or less salient to raters as they evaluate learner texts. This is relevant not only for interpreting the findings in this study, but also for future work on developing automated assessment tools for L2 Spanish writing. Understanding what raters tend to notice or prioritize can guide the design of systems that aim to approximate human scoring and provide meaningful feedback. Finally, these results also have implications for future research on features of L2 Spanish lexicogrammatical use that

can be indicators and predictors of successful L2 writing, which in turn have implications for assessment and pedagogical practices.

Vocabulary Scores

The optimal model examining the relationship between indices of lexical richness and Vocabulary scores explained 29% of the variance in scores. The most variance (20%) was explained by the index of lexical diversity MATTR when analyzing tokens operationalized as lemmas with full verbal information. This could indicate that, for Spanish, verbal diversity could play a role in how humans perceive lexical diversity when assessing an essay. Although in English it could be argued that lexical diversity indices with full verbal information may inflate scores, in Spanish it could be understood as a clearer reflection of lexical diversity and morphological development. This is most likely due to the difference in the verbal inflectional systems of English and Spanish. Since English verbs change slightly in the present tense for person (*run, runs, teach, teaches*), and only once or twice for past tenses (*walked, taught, ran, forgotten*), employing operationalizations beyond the lemma form could potentially inflate diversity scores slightly by accounting for changes that are not significant indicators of language development. However, in Spanish, given that the verbal inflectional system is much richer and it takes many courses to be exposed to them all, it is possible that adding verbal information to the lemmas provides a more fine-grained examination of the learner's language development. Since in Spanish we cannot presume that knowing one verbal form (*hablar, yo hablo, tú hablas*) indicates knowing the verbal forms for all persons, tenses or moods for that verb (*hablé, hablaríamos, hablemos, hablaron*, etc.), analyzing each as one unique form can provide more information of the developmental stage of the learner at that point. For instance, in the example in Table 52 below from a Beginner Spanish learner who scored 4 out of 20 in the vocabulary

category, and 2 out of 20 in the Language Use category, the MATTR score is 0.51. In the essay, there are many types of repetition that suggest developing grammatical competence. There are more than 25 instances of the use of the pronoun “él”, the verb “ser” is the main verb used in the essay with its present indicative third person singular conjugation, along with the verb “gustar”, and noun phrases like “el presidente de los Estados Unidos”. There are also repetitions of common nouns such as “hermanos”, “abuelos”, “perros”, or descriptive adjectives that are overused instead of replacing them with synonyms.

Previous research with similar results (Consolini & Kyle, 2024) have provided pedagogical suggestions that can be implemented at the novice level of instruction, such as teaching vocabulary through word families to elaborate in lexical connections and nuanced descriptions, or by providing early instructional emphasis on sentence coordination with simple conjunctions such as “y” and “pero”. These suggestions are aligned with the idea that even at early stages of development, learners can benefit from noticing and practicing meaningful patterns in vocabulary and sentence structure. However, it is important to strike a balance between cognitive load and scaffolding. At the novice level, it is developmentally appropriate for learners to produce repetitive language, both lexically and syntactically. Nonetheless, there is space to provide support that enhances the connectivity of their emerging lexical and structural systems. One recommendation is to move away from long, decontextualized vocabulary lists within thematic units and instead introduce smaller clusters of semantically related words along with their most frequent collocations. Although thematic units are useful to explore specific topics, they often tend to provide large lists of vocabulary in which not all items may be as relevant for the language functions that learners are acquiring in novice stages. Instead, presenting thematic units with a reduced list of unique vocabulary items along with semantic

networks for each item and their most frequent collocations could be valuable for their writing development. With this approach, it is possible to expose learners to a wider lexical variety that is contained in semantic clusters or small networks and that start emphasizing the connectivity through synonyms and similar meanings, while exposing them to frequent collocations that can foster more natural writing and communication.

Table 52

Example of an Essay from a Beginner learner

Feature	Value / Content
MATTR	0.52
Vocabulary score	4
Language Use score	2
Tokens	211
Unique tokens	75
Essay	Barack Obama es intelligente. El le gusta jugar baloncesto. El es el presidente de los estados unidos. El es medio negro y medio blanco. El es nacido en Honolulu de Hawaii. El gusta concinar y es divertido. El es mas hermanas y hermanos. El es mas abuelas y abuelos. El es mas la tia y el tio. Tambien, el es mas masnastro y padrastro. El le gusta corriendo. El tiene dos las hijas. El tiene una esposa. El no tiene la nieta y el nieto. Es el dos media hermanas de madre y cuatro media de padre hermanos. Es el no los hijos. Es el mas la suegra u el suegro y la nuera - el yerno y la cunada y cunado. Barack Obama es trabajador. El es viejo y tiene pelo negro. El es guapo. El es no tiene perro. El es alto. El donde Hawaii. El vivio Washington DC y blanco casa. El es el cuarenta y cuatro el presidente de los estados. El es contento y activo. El es trabajadar en clase. El le gusta nadando. El es the bisnieto de Charles payne. El es muy bien. El es no prerezoso porque bien presidente de los estados unidos. El es no joven y baja. El es seis y once. El es responsable por el los estados unidos. El es nacido cuatro agosto de diesinueve y sesenta y uno. El es mas universidad en los estados unidos. El le gusta viajar el mundo. El es nueve carros y tres perro y cuatro sillas

Another example essay in Table 53 is from an Intermediate Spanish learner that scored 9 out 20 in the Vocabulary category and the Language Use category, a for which the MATTR score is 0.77. In this essay, there continue to be some repetition, but it is less evident or recurring than in the previous example. Given the prompt of writing about their perfect vacation, the repetition of "vacación perfecta" is not completely unfounded, but other common collocations could be used such as "viaje ideal", "vacaciones de mis sueños", etc. Other repetitions are not perceived as repetitive due to the number of times that is used, but due to its proximity, such in the case of "(…) otro país sería Puerto Rico . Puerto Rico es mi idea de (…)". Here, the use of anaphoric references in the form of pronouns or other nouns, such as "(…) otro país sería Puerto Rico . **Este lugar** es mi idea de (…)". Similarly, the repetition of verbs in different forms may also contribute to the impression of repetitiveness such as "me gusta, me gustaría", when similar verbs can convey the same information (prefiero, disfruto, me encanta).

As mentioned in the previous section, similar results from earlier research (Consolini & Kyle, 2024) have led to pedagogical recommendations that can be applied to the intermediate level of writing development. One suggestion is to place more emphasis on teaching nuanced word families, not only for nouns and adjectives, but also for verbs. At this stage, learners often rely on high-frequency verbs like hacer or tener, which can limit the specificity and naturalness of their writing. Introducing alternatives through semantic clusters, such as realizar, crear, organizar for hacer, or contar, poseer, disfrutar for tener, can help reduce lexical repetition and expand expressive range. Another recommendation is to support the use of anaphoric references, especially through pronouns and combinations of function words and light nouns (e.g., esta situación, ese momento, dicha propuesta), which can replace repeated noun phrases while maintaining clarity and cohesion. These strategies not only support lexical variation but also

promote cohesion across sentences. At the intermediate level, learners can begin to revise their texts with a focus on repetition, vague language, or overly direct reference chains. Activities that draw attention to these patterns can help them gain awareness of how lexical and referential choices affect the overall coherence of their writing. Since it is not uncommon for intermediate learners to plateau in their acquisition development at this stage, these recommendations can add structure to continue growing learners' proficiency and control of the morphosyntactic system in Spanish.

Table 53

Example of an Essay from an Intermediate learner

Feature	Value / Content
MATTR	0.77
Vocabulary score	9
Language Use score	9
Tokens	189
Unique tokens	112
Essay	Pienso que una vacación perfecta es en un país extranjero . Si dejaría a otro país sería Puerto Rico . Puerto Rico es mi idea de un vacación perfecta porque me gusta la música , la comida , la gente , y la cultural . Primero comenzaré mi día y me levanta de mi cama . Me ducharé y me pondré la ropa . Después tendré que llamar mis amigos y mi familia y mi novia por venir a mi casa para comer desayunar . Cuando todos hayan terminado su comer nosotros fuimos a la aireporte en LAX . Cuando la avión aterriza tomaríamos un coche al hotel . El hotel estaría lado de playa . La primera actividad sería comer a la restaurante popular . Yo querría comida que tendría las verduras , y carne falso porque estoy vegan . Después , querríamos tomar mis invitaciónes a un pueblo donde habría bailando , cantando , y las tiendas chiquitas . Cuando la noche termina y el sol saldría , querría pasar el tiempo por la playa e en la océano . A mi me gustaría todos las mañanas . Yo iría con los peces . Después nosotros comería , nos pondríamos la ropa y fuimos en las a caminatas , o compraríamos . Que es mi definición de una vacación perfecta .

Finally, an example essay in Table 54 from an Advanced Spanish learner that scored 14.5 out 20 in the Vocabulary category and the Language Use category, the MATTR score is 0.82. In this essay, there is really very little to say about lexical repetition. Even though some of the words do repeat minimally (*historia, sacrificio, dificultad, quehaceres*), these repetitions are not perceived as repetitive because the instances are relatively separated from each other, and they are repeated to fulfill grammatical or rhetorical purposes. Overall, these essays show the lexical diversity of learners and the differences in developmental stages. Furthermore, these examples may be indicative of why the MATTR scores explained more variance when the token were operationalized as lemmas with full verbal information. The diversity of verb use is seen not only in the use of different verbs (*gustar* vs. *encantar*), but also in the variety of tenses, moods, and persons for which those verbs are conjugated. In the Beginner level essay, there are few verb types and they are all conjugated for the second person singular in the present tense and the indicative mood. There are no other conjugation combinations. In the Intermediate level essay, there are more diverse verb types, the verbs are conjugated for first person singular and plural, and third person plural, and they are also conjugated for different tenses, such as future or conditional. Finally, in the Advanced level essay, there are a variety of verbs conjugated in the past, present and conditional tenses, and in the indicative and subjunctive moods for both the first-person singular and plural, and the third person plural. This suggests that verbal information plays a role in the perception of repetition and lexical diversity, and that providing that verbal information to the indices that measure lexical diversity can make them more robust to approximate human judgment scores.

Table 54

Example of an Essay from an Advanced learner

Feature	Value / Content
MATTR	0.82
Vocabulary score	14.5
Language Use score	14.5
Tokens	269
Unique tokens	151
Essay	Como una estudiante de primera generación, quienes mis padres tuvieron que separarse para sacrificar todo por mi y que yo pueda tener mejores oportunidades, me impacta bastante verlos desesperados tratando de obtener cualquier ayuda financiera para poder seguir permaneciendo aquí en los Estados Unidos, nuestra historia no ha sido la mejor ni la peor que se ha relatado, pero si ha sido original. Y a la vez, pienso que ha sido una historia hermosa, puesto que hemos ido en contra de todas las adversidades que la vida se ha encargado de poner frente nuestro. Cuando vine a los Estados Unidos, no supe que la situación seria tan difícil como resultó siendo, no pude disfrutar de mi niñez como cualquier otra niña / joven lo hubiera hecho, tuve que tener responsabilidades a una corta edad, pero todas esas experiencias han formado la persona quien yo soy ahora y eso es lo maravilloso de mi historia. Si no hubiera sido por mi viaje a los Estados Unidos, siento que nunca habría aprendido el valor de la vida, a organizar mi tiempo, a ser independiente y saber valorar el sacrificio de mis padres. Cuando vivía en Bogotá, solía tener todo hecho para mi, es decir, mi mama solía hacer todos los quehaceres, mientras que yo solo estudiaba, como " spoiled " por así decirlo, cuando vine a los EE. UU y por algunas circunstancias tuve que vivir sola por algún tiempo, tuve que aprender a hacer todos los quehaceres, llevar la administración de mis cuentas, pagar la renta, y todo lo que involucra, nunca pensé que seria tan difícil pero lo logré, lo superé y ahora me encuentro en una de las universidades mas prestigiosas de los Estados Unidos.

Furthermore, 7% of the 29% explained by the model was accounted by bigram dependency SOA measures. Particularly, different measures of Object dependencies, which provide an insight into how accurately learners produce verb-object constructions. In the case of bigram object dependencies SOAs, the operationalizations that explained the most variance in scores were raw words and simple lemmas. This is most likely because the relationship between the verb and the object in each bigram is more relevant to the measurement than the conjugation of the verbs involved in those bigrams. This is due to verb-object collocations being typically the same regardless of the person, number, tense or mood in which the verb is conjugated. Finally, approximately 2.3% of the 7% variance explained by bigram dependency SOA measures were in the form of Adverb dependencies, which can provide an insight into how accurately learners produce verb-adverb or adjective-adverb constructions. Similarly to Object dependencies, in the case of adverb dependencies, the operationalization that explained the most variance was raw words. This also indicates that the syntactic dependencies between adverbs and verbs or adverbs and adjectives does not require detailed verbal information to assess dependency and raw words or lemmas can be used. The results from the lexical sophistication measures in relation to Vocabulary scores indicate that more proficient L2 writers tend to use more strongly associated bigram Object and Adverb dependencies. Finally, content word frequency explained only 1.7% of the variance for Vocabulary scores, indicating that the frequency of use of the lexical items used in the essays can explain a small part of the lexical sophistication of each essay.

Overall, the results from lexical diversity measures are in line with previous research in both English and Spanish that have shown that more proficient writes tend to use a more varied array of lexical items, while lower-proficiency writers tend to reuse the same lexical items more often within their writings (Kyle & Crossley, 2015; Díez-Ortega & Kyle, 2024). Similarly, the

results involving lexical sophistication measures in the form of bigram object dependencies SOA measures are somewhat in line with previous research in English and Dutch (Kyle & Eguchi, 2021; Paquot, 2019; Rubin et al., 2021); however, in these previous studies, other dependency relationships (e.g., adjective dependencies) were also predictive or correlated with writing proficiency, while in the present study only object and adverb dependencies explained Vocabulary scores variance. This difference in the Adjective dependency results in the current study may stem from the fact that adjective use may plateau early in L2 writing development since noun-adjective constructions in Spanish as straightforward are commonly learned in early stages of development, while more complex noun-adjective combinations may not occur very frequently even in more advanced levels.

Language Use Scores

The optimal model examining the relationship between indices of lexical richness and Language Use scores explained 32.8% of the variance in scores. The most variance (22%) was explained by the index of lexical diversity MATTR when analyzing tokens operationalized as lemmas with full verbal information. This could indicate that a diverse use of lexical items is tied to proper use of related language features, such as proper word order in general and in collocations, and proper use of morphology rules when conjugating verbs or inflecting adjectives to match gender and number. Similarly, there may be a relationship between having a diverse productive vocabulary and employing a diverse array of syntactic patterns. However, this last relationship is outside of the scope of this study, and it is merely a speculation.

The Language Use category as evaluated by human raters focuses on word order, morphology, and sentence variety and complexity. The MATTR index of lexical diversity can provide insight into the morphology employed by learners, especially when the tokens are

operationalized as lemmas with full verbal information. The more diverse and accurate conjugation of verbs may partially be a reflection of higher scores on Language Use concerning proper morphology. Furthermore, 10% of the 32.8% of variance explained by the Language Use model is explained by bigram dependency SOA measures in the form of T-scores, MI, and Delta P scores. This is 3 percentage points more variance explained than in the Vocabulary model. This is likely to bigram dependencies better reflecting the human raters' assessment of sentence variety and complexity, which is a focus of the Language Use category. Moreover, another difference with the Vocabulary model, is that the Language Use model includes a variety of measures of bigram dependency SOAs and includes the three types of dependencies studied in this research project and in previous research: Object, Adverb and Adjective dependencies. From the 10% variance explained by bigram dependency SOAs, 6% is explained by Object dependencies, 3% by Adverb dependencies, and 1% by Adjective dependencies. Similarly to the Vocabulary model's results, Object dependencies explain the most variance related to dependency bigrams. As mentioned before, previous research in English has found a relationship between bigram dependency SOAs scores and writing proficiency. However, in Spanish, this is emerging research, and more studies are necessary to make claims about the relationship between bigram dependencies and writing quality, especially with Adjective and Adverb dependencies. Finally, content word frequency accounted for only 1% of the variance for Language Use scores, suggesting that the sophistication of lexical items assessed by how frequently or infrequently a word is used may be less relevant for the Language Use category than it is for the Vocabulary category.

When evaluating the essay examples provided in Tables 52, 53, and 54 through the lenses of the Language Use category, the Beginner level essay shows not only lexical repetition, but

also syntactic repetition. Most sentences are simple and uncoordinated or subordinated, they showcase the same structural pattern on subject plus verb (*Él es… Él tiene… Él le gusta…*). Morphologically, there is little variety of inflections, not only on verbs which are always conjugated in the present indicative for the second person singular, but in adjectives as well, since most are inflected for the same person and number. The Intermediate level essay showcases more variety of syntactic structures by introducing coordination and some subordination, but it relies heavily on similar subject plus verb structures (*Yo querría… Yo iría… Yo…*) mostly conjugated for the same person and number. Other syntactic structure that repeats several times is the use of prepositions and conjunctions to start a sentence, such as *Después, yo…*, *Cuando yo...*, etc. Morphologically, there are small errors and a repetitive use of the conditional tense for the first-person singular. Finally, in the Advanced level essay, there is some repetition of relative clauses, however, they are varied in content and form, and they are used thoughtfully throughout the essay. Moreover, there is a diversity of tenses and moods for different persons, with the inclusion of periphrastic structures (e.g., *se ha encaragado de, me encuentro en*). This provides rich morphological and syntactic variety.

Overall, the model for Language Use showed results that, similarly to the results from the Vocabulary model, are in line with previous research in English that have used bigram dependencies SOAs measures and lexical sophistication measures to predict writing quality (e.g., Granger & Bestgen, 2014; Garner et al., 2019; Kim et al., 2018). Additionally, the results from both models are in line with previous recent research in Spanish multivariate approaches to modeling the relationship between lexical richness and writing proficiency (Consolini & Kyle, 2024; Díez-Ortega & Kyle, 2024; Schnur & Rubio, 2021). These studies showed a strong relationship between indices of lexical diversity and measures of proficiency, showing that more

proficient writers or speakers tend to use more diverse vocabulary. However, the current research study provides new information to this body of research. Previous research in L2 Spanish writing has not evaluated indices of lexical richness in relation to scores of human judgments, and they have also not evaluated bigram dependencies as part of lexical sophistication measures. Furthermore, this dissertation adds to the previous body of research by evaluating different operationalizations of word normalization to carry out measures of lexical richness, and provides valuable information for future research in second language Spanish writing development and assessment.

5 CONCLUSION AND OUTLOOK

This dissertation's goal was to examine in more depth that until recently the modeling of lexicogrammatical features of L2 Spanish writing in the form of indices of lexical richness and their relationship to human judgment scores of writing quality. Based on previous research modeling indices of lexical richness and their relationship to L2 writing proficiency in English (Crossley & McNamara, 2012; Kyle & Crossley, 2015; Bulté & Housen, 2012; Jarvis, 2013; Lu, 2012; Salsbury et al., 2011; Kim et al., 2018; Bestgen & Granger, 2014), a series of studies were conducted with L2 Spanish essays to find operationalizations and indices that would best predict L2 writing quality in the form of human judgments. The first two studies in this dissertation reviewed four different operationalizations for indices of lexical diversity, and indices of lexical sophistications. Based on the results of these first two studies, a third study was conducted to evaluate the percentage of variance in human judgment scores of L2 Spanish writing quality that can be explained by indices of lexical richness. From these studies, a series of suggestions emerged for modeling lexical richness and writing quality in L2 Spanish, which contribute to the continuous development of the field of Spanish L2 writing assessment. A summary of the outcomes of the three studies presented in this dissertation are outlined below.

Operationalization Of Morphological Normalization and Lexical Richness

An important contribution of this dissertation is the review of a series of different operationalizations for evaluating L2 Spanish tokens when calculating indices of lexical richness. In chapter 3, two studies are conducted to establish which operationalization provides the best linear model fit to explain the variance in Vocabulary and Language Use scores provided by human raters. After creating and evaluating 240 linear models between each index of lexical richness (N=20) and each operationalization (N=4) for each scoring category (N=3), suggestions

of operationalizations for each index were provided. These suggestions were made by evaluating all linear models and identifying the operationalization for each index and each scoring category that provided the most variance explanation and that showed the best model fit for the data analyzed. The results from these first two studies provide information to future researchers modeling indices of lexical richness in Spanish L2 writing, or investigating models that can provide automatic assessment of large data sets of Spanish L2 writing. A summary of the findings from study one and two can be found below.

Summary of Findings

Research questions 1a-1d: Operationalizations for lexical diversity indices. The first set of research questions was answered by study one and it is comprised of the following questions:

1- To what extent are different operationalizations of morphological normalization when calculating lexical diversity more or less predictive of L2 Spanish writing quality in the form of human judgments?

a. To what extent is lexical diversity predictive of L2 Spanish writing quality in the form of human judgments when tokens are morphologically normalized as raw words?

b. To what extent is lexical diversity predictive of L2 Spanish writing quality in the form of human judgments when tokens are morphologically normalized as lemmas with parts of speech information?

c. To what extent is lexical diversity predictive of L2 Spanish writing quality in the form of human judgments when all tokens are morphologically normalized as lemmas, except for mood and tense information in verbs?

d. To what extent is lexical diversity predictive of L2 Spanish writing quality in the form of human judgments when all tokens are morphologically normalized as lemmas, except for mood, tense, person, and number information in verbs?

The results indicated that different operationalizations of morphological normalization provide different percentages of explanation for human scores variance when calculating indices of lexical diversity in the form of MTLD and MATTR calculations. For both indices of lexical diversity, the operationalization of tokens in the form of lemmas with full verbal information was the operationalization that provided better model fit and that explained the most variance when compared to the operationalization of tokens in the form of raw words, simple lemmas, or lemmas with partial verbal information. In this study, "partial" and "full" verbal information refer to verb forms used in the essays tagged with mood and tense, and mood, tense, person and number, respectively. However, the results also show that the difference in variance explained between operationalizations was between 1-2 percentage points. For example, while lemmas with full verbal information when calculating MATTR scores explained 24.7% of the variance in Vocabulary scores and 28.5% in Language Use scores, other operationalizations explained 23% (lemmas with partial verb information), 22.7 (simple lemmas) and 23.2% (raw words) of the Vocabulary score variance, and 27.2% (lemmas with partial verb information), 26.7% (simple lemmas, raw words) of the Language Use score variance. These are small but meaningful differences in R2 scores (Cohen, 1988), particularly when creating models with the same predictors and outcome variables but different operationalizations.

Regarding previous research, these results add to previous studies in the field. Studies focusing on L2 English have examined different operationalizations when calculating indices of lexical diversity and have found lemmas to be an optimal operationalization for evaluating L2 English writings (e.g., Crossley et al., 2011; Jarvis, 2013; Kyle & Crossley, 2015). The results from study one suggests that simple lemmas may not be the optimal operationalization when evaluating L2 Spanish writings. This is mostly likely due to the difference in the verbal inflectional system between English and Spanish.

Overall, the results from study one suggests that some of the established methodology for modeling writing quality and features of lexicogrammatical use may need to be revised when conducting research in L2 Spanish. These results also provide evidence for the need to carry out similar studies in other L2s that may differ morphologically, or in other terms, from English.

Research questions 2a-2d: Operationalizations for lexical sophistication indices. The second set of research questions was answered by study two, and it is comprised of the following questions:

2- To what extent are different operationalizations of morphological normalization when calculating lexical sophistication more or less predictive of L2 Spanish writing quality in the form of human judgments?

a. To what extent is lexical sophistication predictive of L2 Spanish writing quality in the form of human judgments when tokens are morphologically normalized as raw words?

b. To what extent is lexical sophistication predictive of L2 Spanish writing quality in the form of human judgments when tokens are morphologically normalized as lemmas with parts of speech information?

c. To what extent is lexical sophistication predictive of L2 Spanish writing quality in the form of human judgments when all tokens are morphologically normalized as lemmas, except for mood and tense information in verbs?

d. To what extent is lexical sophistication predictive of L2 Spanish writing quality in the form of human judgments when all tokens are morphologically normalized as lemmas, except for mood, tense, person, and number information in verbs?

The results indicated that different operationalizations of morphological normalization provide different percentages of explanation for human scores variance when calculating indices of lexical sophistication. Since there are many indices of lexical sophistication that were part of study two, results will be discussed in three broader categories within lexical sophistication: word frequency measures, bigram strength of association measures, and bigram dependencies.

Results show that for both indices of word frequency measures (all words and content words), the operationalization of tokens in the form of simple lemmas was the operationalization that provided a better model fit and that explained the most variance when compared to the operationalization of tokens in the form of raw words, or lemmas with partial or full verbal information. The results also show that the difference in variance explained between operationalizations was between 2 percentage points. For example, while simple lemmas when calculating content word frequency scores explained 3.3% of the variance in Vocabulary scores and 2% in Language Use scores, other operationalizations explained 0.6% (raw words), 1.1%

(lemmas with partial information) and 1.2% (lemmas with full information) of the Vocabulary score variance, and 0.0% (raw words), 0.1% (lemmas with partial information) and 0.9% (lemmas with full information) of the Language Use score variance.

Furthermore, results also show that for all four indices of bigram strength of association (T-score, MI, Delta P in each direction), the operationalization of tokens in the form of raw words was the operationalization that provided a better model fit and that explained the most variance when compared to the operationalization of tokens in the form of simple lemmas, or lemmas with partial or full verbal information; except for the calculation of Delta P from right to left, for which the lemma operationalization was preferred. The results also show that the difference in variance explained between operationalizations was between 2-3 percentage points, with T-score exhibiting the most variance between operationalizations (2-3 percentage points), MI showing more moderate differences between operationalizations (1-2 percentage points), and Delta P showing very small differences between operationalizations (less than 1 percentage points).

Lastly, results for bigram dependencies showed a variety of operationalizations and indices mix for different bigram dependencies. For instance, for Object dependencies, the raw operationalization was preferred for T-scores, while the lemma operationalization with partial verbal information was preferred for MI scores, and the simple lemma operationalization was preferred for Delta P calculations. For Adverb dependencies, the lemma operationalization with partial verbal information and the raw operationalization of T-score both showed very similar results and negligible differences in R2 scores. For MI scores, lemmas for the Vocabulary category and lemmas with partial verbal information for the Language Use category were preferred. And for Delta P scores, simple lemmas and lemmas with partial verbal information

were preferred depending on the direction of the calculation. Finally, for Adjective dependencies, all operationalizations across categories showed very little difference, but overall, they all prefer raw operationalizations for T-score and MI, while for Delta P a combination of raw words and simple lemmas is preferred, depending on the direction of the calculation.

Regarding previous research, these results continue to add to previous studies in the field. Studies focusing on L2 English have typically employed lemmas when analyzing indices of lexical sophistication (e.g., Eguchi & Kyle, 2020; Kim et al., 2018). Similarly, the results from study one suggests that a combination of raw words and simple lemmas may be the optimal operationalization when evaluating L2 Spanish writings in relation to indices of lexical sophistication, although certain indices may provide better model fit when adding partial verbal information to the lemmas. These results may be different from English, most likely due to the varying morphological nature of verbs in Spanish, additionally to wider variation in adjective and adverb inflections to account for number and gender agreement.

Overall, the results from study two suggests that some of the established methodology for modeling writing quality and features of lexicogrammatical use may not be appropriate when conducting research in L2 Spanish lexical sophistication indices. However, this may depend on the lexical sophistication indices used, and therefore case by case evaluations should be made when dealing with languages that may have verbal inflectional systems more complex than the one in English, or more agreement inflectional variations.

Lexical Richness as Predictor of L2 Spanish Writing Quality: A Multivariate Approach

Another contribution of this dissertation is the addition of research employing multivariate approaches to modeling lexical richness in L2 Spanish. In chapter 4, a third study

was conducted to establish which indices of lexical richness provide the best linear model fit to explain the variance in Vocabulary and Language Use scores provided by human raters. Two models were created with each of the 20 indices of lexical richness and their respective operationalizations based on the results from studies one and two, minus two indices that were nonlinear. Afterwards, the optimal model for each scoring category was reached by using the Mumin::dredge() package, and one optimal model each was found for Vocabulary and the Language Use categories. These models explained the variance in human judgment scores related to lexicogrammatical features used in the evaluated essays. The results from this third study provide information to future researchers and provides some pedagogical implications for the development of lexical richness in L2 Spanish learners writing competence. A summary of the findings from study one and two can be found below.

Summary of Findings

Research questions 3: Modeling lexical richness in L2 Spanish. The last research question was answered by the third study:

3-To what extent are features of written lexical and lexicogrammatical use in the form of lexical sophistication and lexical diversity predictive of L2 Spanish writing quality in the form of human judgments?

The results indicated that a combination of lexical sophistication indices and the MATTR index for lexical diversity can account for up to 30% of the variance in Vocabulary and Language Use scores provided by human raters. Additionally, results from the third study showed that a different combination of lexical sophistication indices were part of the optimal model to predict score variance for the category of Language Use when compared to the

category of Vocabulary. This indicated that rubric information and format, along with rater training, are important elements when modeling indices of lexical richness and writing quality in the form of human judgment scores. Furthermore, results show that lexical diversity accounted for most of the variance in each model (between 20 and 22%), indicating diversity as one of the most salient features of lexical richness that the indices are sensitive to. The second highest amount of variance was accounted by indices measuring strength of association in bigram dependencies. This was most prominent in the Language Use model, which was comprised of 7 indices of bigram dependency SOA, compared to the Vocabulary model, which was comprised of 4 indices of bigram dependency SOA, almost 50% less. Moreover, from all types of bigram dependencies, Object dependencies were more prominent in both models, while Adverb dependency indices appear once and twice in the Vocabulary and Language Use models, respectively, and Adjective dependencies only appear in one index in the Language Use model. This indicated that syntactic dependencies are also a strong indicator of writing quality, mainly in the form of Object bigram dependencies, along with lexical diversity.

Regarding previous research, these results are somewhat in line with research in English and Dutch L2 that showed that more proficient writers tend to use more strongly associated bigram dependencies (Kyle & Eguchi, 2021; Paquot, 2019; Rubin et al., 2021). However, in these previous research studies, all three types of bigram dependencies are predictive of writing quality, while in the current study mostly Object dependencies are relevant predictors. This may stem from the fact that adjective use may plateau early in L2 writing development since noun-adjective constructions in Spanish as straightforward are commonly learned in early stages of development, while more complex noun-adjective combinations may not occur very frequently even in more advanced levels.

Overall, the results from the third study suggests that a multivariate approach to modeling features of lexicogrammatical use in L2 writing Spanish can predict up to 30% of the variance in writing quality scores in the form of human judgments. Moreover, these multivariate models benefit from tailored operationalizations for each index.

Contributions

This dissertation contributed to the field of corpus linguistics in two ways. The first contribution stems from studies one and two, which provide a thorough assessment of four different types of operationalizations for morphological normalization that account for the main lexicogrammatical feature that is different between English and Spanish: the verbal inflectional system. This contribution not only provides new suggestions for operationalizations when working with L2 Spanish writings, but it opens the door to the possibility of other operationalizations across different indices being better fits for other L2s. The results of the first two studies support previous research in L2 English corpus linguistics and writing assessment that have shown that different operationalizations may have an impact on the calculations of lexical diversity and lexical sophistication indices. Additionally, the results of this dissertation also expand previous research to indicate that operationalizations with partial and full verbal information in the form of mood, tense, person, and number may provide more optimal models when working with L2s that have rich verbal inflectional systems (e.g., romance languages).

The second contribution stems from the third study in this dissertation project that assembled different models to evaluate which indices of lexical diversity and lexical sophistication can explain the most variance in L2 Spanish writing quality scores in the form of human judgments. This adds to an emerging body of research in L2 Spanish writing development and assessment, and provides insights that can contribute to pedagogical

instructional decisions for Spanish writing development. Moreover, results from this dissertation also contribute information regarding the use of human judgment scores and the training of raters, suggesting that a clear rubric that specifies clearly what concepts are being evaluated in each category can provide more fine-grained model predictions, as shown in the differences between the Vocabulary and Language Use categories and their respective models. Finally, these results also add and expand previous research in L2 writing assessment by incorporating bigram dependencies measures of strength of association in a multivariate approach to explain writing quality variance, which had been done in other L2s before, but not Spanish. Results in this area suggest that Spanish writing assessment models can also benefit from bigram dependency SOA measures, although more studies are necessary to confirm to what extent.

Finally, the third contribution from this dissertation project is the collection and upcoming release of 400 L2 Spanish essays that have been rated by human raters. This is a significant contribution to the field of L2 Spanish corpus linguistics since human judgments of writing quality are not freely available to study their relationship with indices that calculate lexical sophistication and lexical diversity. It is important to have human judgments of writing quality to continue to refine indices and better understand what these measures reflect. Overall, the human judgments of professionals specialized in second language writing is what guides the development of automatic assessment tools and informs researchers at large what qualities or features of second language writing are relevant when examining writing proficiency, quality, and development.

Implications

The findings discussed in this dissertation have implications for the fields of second language acquisition, second language Spanish writing assessment and second language Spanish pedagogy.

Second Language Acquisition

Regarding second language acquisition, the findings in this dissertation support usage-based theories of language learning (e.g., Ellis, 2002), in which input frequency and meaningful use of language are central. More precisely, the findings of bigram dependencies explaining variance in Language Use scores emphasize the fact that learners develop more target-like grammatical structures over levels of proficiency, or time. This may indicate that the more they are exposed to more strongly syntactically associated bigrams, the more they produce them in their writings. Similarly with more strongly associated collocations, or acquisition and productive use of more complex verb forms that include more tenses, moods, and subjects with different person and number categories.

Second Language Spanish Writing Assessment

Regarding writing assessment, the results of the studies in this dissertation project provide insight to modeling automatic tools of writing assessment for L2 Spanish with consideration of the operationalization of morphological normalization when working on large datasets. The implications for L2 Spanish emphasize the need for operationalizations to contain full or partial verbal information when calculating certain indices of lexical richness but warn that these operationalizations may need to be examined on a case-by-case basis. Furthermore, the study of operationalizations of morphological normalization in Spanish implies that other L2s

that have lexicogrammatical or morphosyntactic differences with English may also benefit from evaluating what operationalizations may be better fits for the predictive models of second language writing. Finally, rater training and careful assembly of a detailed rubric may contribute to more fine-grained examination of lexicogrammatical or morphosyntactic features.

Second Language Spanish Pedagogy

Regarding second language Spanish pedagogy, results have implications for targeted instructional practices related to vocabulary breadth and depth, and syntactic variety. Firstly, results suggest that teaching vocabulary in word families may enhance nuanced lexical links, and may offer more possibilities for the application of synonyms or nuanced lexical choices. This is a practical suggestion that can be applied from beginner levels of proficiency. As vocabulary is introduced by topic or theme, small word families can be created for key lexical items. Secondly, early introduction to basic syntactic variety may also contribute to more proficient productive use of syntactic structures. Simple coordination with two different conjunctions can simplify repetition in enhancing writing flow even from beginning stages of writing development. Finally, an introduction to different variations in person in verb conjugation from early stages of writing may also aid in the development of more diverse writing that avoids repetition and presents information in a more concise way.

Limitations

The present dissertation and the studies that comprise it are not exempt from limitations, as with most studies. Firstly, the variety of genres in the essays analyzed could be expanded. Although the sample of 400 essays is a robust sample, the essays are comprised of only descriptive writings. Adding other academic genres such as argumentative essays or narratives

retellings could provide a more comprehensive understanding of how different indices of lexical richness can account for variance in human judgment scores of writing quality. Differences in the relationship between lexicogrammatical features and writing quality or proficiency have been shown to be significant in L2 English studies (e.g., Crossley and McNamara, 2012; Kyle and Crossley, 2016; Lu, 2011).

Another limitation that could be addressed in future studies is the integration of other layers of linguistic analysis such as syntactic complexity or discourse-level features, which have not been researched in Spanish as a second language. Although different operationalizations of morphological normalization can capture fine-grained differences at a lexical level, other syntactic and discourse level measures can provide a better comprehensive assessment of writing quality. For instance, the use of embedded clauses, coordination or passive constructions could play a role in writing quality development and human judgement assessments. Similarly, discourse-level features like cohesion, rhetorical structure, or thematic progression also affect perceptions of writing proficiency but fall outside the scope of what morphological forms can measure. This limitation aligns with critiques in L2 writing research (e.g., Lu, 2011; Norris & Ortega, 2009; Biber et al., 2016), which emphasize that fully capturing writing development requires integrating multiple layers of linguistic analysis, including syntax and discourse.

Finally, another limitation is the use of one corpus from one specific institution. This may result in very specific samples that may or may not be generalizable to other populations since different institutions may teach language classes with different approaches. Relatedly, replicating the types of studies present in this dissertation with other Spanish-speaking populations such as heritage speakers can also provide valuable information about writing development that may be valuable not only when aiding heritage speakers in the classroom, but also may add to the

understanding of language acquisition mechanics in populations that are not technically L2 learners but that share L2 learners' background language, in this case English in the US.

Outlook

As the field of second language assessment grows its analytical tools with the help of natural language processing tools and more sophisticated statistical models, it is important to remember that replication is at the core of scientific progress. In this context, expanding the research carried out in the last two decades in L2 English to other L2s is paramount to continue to obtain a better and more comprehensive understanding not only of writing development and assessment techniques, but also for the discovery of overarching themes and patterns in second language acquisition in development at large. Although emerging studies in other L2s in the field are growing (Consolini & Kyle, 2024; Díez-Ortega & Kyle, 2024; Hao et al., 2023; Vold, 2023), it is hoped that researchers will replicate L2 English studies in other languages and evaluate operationalizations of difference lexical diversity indices to uncover better suited models for other L2s and provide insights into second language assessment, writing, and acquisition.

APPENDICES

APPENDIX A: ESSAY RATING RUBRIC

Scoring Guidelines

Familiarize yourself with the rubric by scoring some practice essays (please, see these attached in the folder "Practice essays", do not practice with the final sample for research). The purpose of this exercise is to become familiar with the rubric, but not to depend on the rubric to make micro decisions. Rather, the goal is to feel comfortable with what type of essay falls in what range of scores. This minimizes the time spent on each essay, which in turn reduces over-thinking about the analysis of each essay. Research shows that the more time we spend re-viewing, re-considering, and re-analyzing writing, the more we tend to make mistakes or mis-compensate for errors and successes. Assessing writing with an analytic rubric involves a degree of intuition after the first or second read of the essay, but it should not call for a detailed or in-depth read of the essay. By familiarizing ourselves with the rubric and practicing, we can develop a clear sense of what scoring category an essay falls under by reading it once or twice and thus avoiding time-consuming in-depth readings that can cause errors in scoring.

Don't compare essays when scoring. It can be tempting to provide a higher score to a given essay when the previous one was score very low because, by comparison, the new one is "so much better". This is also why practicing is effective. While practicing, we can make comparisons among the essays in the practicing sample to train our intuition as to what a good, great, or poor essay looks like. However, at the time of scoring the sample for research, each decision should be based solely on the essay we are scoring at the moment.

Beware of personal biases when it comes to style. Remember that we are trying to score for different given and established categories (organization, content, mechanics, vocabulary, etc.) and not the way we would personally prefer the essay to be. Of course, there is an element of style when we evaluate "academic register". Nonetheless, there should not be points taken off for things we would have said differently if these things are not grammatically incorrect, syntactically hard to comprehend, or unnatural to the point of obscuring the message. There is a category that mentions "awkward" syntax – for this, I encourage you to ask, if you feel hesitant, whether the syntax is awkward on itself (needs to be re-read to be fully understood, it is convoluted, it is confusing), or if it is clear but not our style or a style of writing we are comfortable with.

	Content		Organization		Vocabulary		Language Use		Mechanics
20 18 16	-Thorough and logical development of thesis -Substantive and detailed - No irrelevant information	20 18 16	- Excellent overall organization - Clear thesis statement - Substantive introduction and conclusion - Excellent use of transition words - Excellent connections between paragraphs - Unity within every paragraph	20 18 16	- Very sophisticated / advanced vocabulary - Excellent choice of words with no errors - Excellent range/variety of vocabulary - Idiomatic vocabulary - Academic register	20 18 16	- No major errors in word order or complex structures - No errors that interfere with comprehension - Only occasional errors in morphology - Frequent use of complex sentences - Excellent sentence variety	20 18 16	- Little to no spelling errors - Little to no punctuation errors
15 13	- Good and logical development of thesis - Fairly substantive and detailed - Almost no	15 13	- Good overall organization - Clear thesis statement - Good introduction and conclusion - Good use of transition words. - Good connections between paragraphs	15 13	- Somewhat sophisticated / advanced vocabulary - Attempts, even if not completely successful, at sophisticated vocabulary - Good choice of words with some errors that don't obscure	15 13	- Occasional errors in awkward order or complex structures - Almost no errors that interfere with comprehension - Attempts, even if not completely successful, at a	15 13	- No more than a few spelling errors in less frequent vocabulary - No more

11	irrelevant information	11	- Unity within most paragraphs	11	meaning - Adequate range of vocabulary but some repetition - Approaching academic register	11	variety of complex structures - Some errors in morphology - Frequent use of complex sentences - Good sentence variety	11	than a few punctuation errors
10 8 6	- Some development of thesis - Not much substance or detail - Some irrelevant information	10 8 6	- Some general coherent organization - Minimal thesis statement or main idea - Minimal introduction and conclusion - Occasional use of transitions words - Some disjointed connections between paragraphs - Some paragraphs may lack unity	10 8 6	- Unsophisticated / not advanced vocabulary - Limited word choice with some errors obscuring meaning - Repetitive choice of words - No resemblance to academic register	10 8 6	- Errors in word order or complex structures - Some errors that interfere with comprehension - Frequent errors in morphology - Minimal use of complex sentences - Little sentence variety	10 8 6	- Some spelling errors in less frequent and more frequent vocabulary - Several punctuation errors
5 3	- No development of thesis - No substance	5 3	- No coherent organization - No thesis statement or main idea - No introduction and conclusion - No use of transition words	5 3	- Very simple vocabulary - Severe errors in word choice that often obscure meaning - No variety in word choice	5 3	- Serious errors in word order or complex structures - Frequent errors that interfere with comprehension	5 3	- Several spelling errors even in frequent

0	or details - Substantial amount of irrelevant information	0	- Disjointed connections between paragraphs - Paragraphs lack unity	0	- No resemblance to academic register	0	- Many errors in morphology - Almost no attempt at complex sentences - No sentence variety	0	vocabulary - Many punctuation errors

Rubric adapted from Connor-Linton, J., and Polio, C. (2014). Comparing perspectives on L2 writing: Multiple analyses of a common corpus, Journal of Second Language Writing, 26, 1-9, https://doi.org/10.1016/j.jslw.2014.09.002.

Notes on the difference between morphological errors and spelling errors (Language Use vs Mechanics):

- Morphological errors may indicate difficulties with morphology rules, while spelling errors may indicate difficulties with phonological awareness or orthographic knowledge.

- Morphological errors: Errors in the formation and use of words based on their shape and meaning. These include:

incorrect plurals (e.g., viruses, ciudads, etc);

incorrect gender/number agreement for adjectives (e.g., tres choches rojas, dos coche rojos, cuatro coche rojas, etc);

incorrect verb conjugations (the subjunctive, the past, the future, number agreement on verb conjugation, etc.); and

incorrect use of derivations (diminutives, nominalization, affixes, etc.)

- Spelling errors: Misspellings of individual words, often due to confusion with similar sounds or letter patterns. These can:

involve mistakes in the individual letters used to spell the word regardless of the relationship to other words; and

do not change the meaning of the word, but make it difficult to read

- If there is hesitation on weather it is a morphological or spelling error, these questions can be of guidance:

In the present incorrect form, does the meaning of the word change in the given context? If yes, it's a morphological error

In the present incorrect form, does the form of the word show incomplete knowledge of how to perform correct word formation regardless of meaning change? If yes, it's a morphological error

In the present incorrect form, could the word be sound out/read out loud and sound correctly and it's the meaning preserved in context? If yes, it's a spelling error

In the present incorrect form, are all or most of the letters part of the word but in a different order, or are there letters missing without creating a new or unintended meaning in context? If yes, it's a spelling error

References Cited

Allan, D. (1980). *Probability theory: A concise course*. Dover Publications.

Berton, M., & Sánchez, L. (2023). Effects of passive vocabulary knowledge and task type on lexical sophistication in L2 writing. In I. Checa (Ed*.), Lexical diversity in creative writing of L2: A quantitative approach* (pp. 45–62).

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

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

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, *9*(1), 2–20.

Biber, D., Gray, B., & Poonpon, K. (2016). Grammatical complexity in academic English: Linguistic change in writing. Cambridge University Press.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Bowden, H. W., Gelfand, M. P., Sanz, C., & Ullman, M. T. (2010). Verbal inflectional morphology in L1 and L2 Spanish: A frequency study. *Language Learning, 60*(1), 44-87. https://doi.org/10.1111/j.1467-9922.2009.00551.x

Bulté, B., & Housen, A. (2021). Syntactic and lexical complexity in L2 English academic writing: Syntactic and lexical complexity in L2 English academic writing. *Journal of English for Academic Purposes, 51*.

Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference: A practical information-theoretic approach (2nd ed.). Springer.

Castañeda-Jiménez, G., & Jarvis, S. (2013). Exploring lexical diversity in second language Spanish. In K. L. Geeslin (Ed.), *The handbook of Spanish second language acquisition* (pp. 498-513).

Church, K. W., & Hanks, P. (1989). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16*(1), 22-29.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Routledge.

Connor-Linton, J., & Polio, C. (2014). Understanding L2 writing contexts: The case of Korean EFL university writers. *Journal of Second Language Writing, 25*, 1-16. https://doi.org/10.1016/j.jslw.2014.05.001

Consolini, C., & Kyle, K. (2024). Lexical richness and proficiency in L2 Spanish: A multivariate approach. *Studies in Second Language Acquisition, 46*(1), 123-145.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics, 17*(2), 94-100. https://doi.org/10.1080/09296171003643098

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading, 35*(2), 115–135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing, 32*, 1–16.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing, 28*(4), 561–580. https://doi.org/10.1177/0265532211408039

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2013). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics, 34(*5), 1–22.

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

de la Torre García, N., Ainciburu, M. C., & Buyse, K. (2021). Morphological complexity and rated writing proficiency: The case of verbal inflectional diversity in L2 Spanish. *International Journal of Applied Linguistics, 172*(2), 290-318.

Díez-Ortega, M., & Kyle, K. (2024). The development of lexical richness in L2 Spanish: A longitudinal study. *Language Learning, 74*(2), 345-372.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Lawrence Erlbaum Associates.

Eguchi, M., & Kyle, K. (2020). Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *The Modern Language Journal, 104*(4), 381–400. https://doi.org/10.1111/modl.12637

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics, 27*(1), 1–24.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the Phrasal Teddy Bear. Annual Review of Applied Linguistics, 32, 17–44. https://doi.org/10.1017/S0267190512000025

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

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of second language writing, 4*(2), 139-155.

Engelhard, G. (2013). Invariant measurement: Using Rasch models in the social, behavioral, and health sciences. Routledge.

Educational Testing Service (ETS). (2012). *Best Practices for Constructed-Response Scoring*. ETS.

Evert, S. (2005). The statistics of word cooccurrences: Word pairs and collocations. (Doctoral dissertation, Universität Stuttgart).

Fergadiotis, G. (2015). Lexical diversity in spoken and written discourse: A comparative study. *Journal of Quantitative Linguistics, 22*(3), 187-214. https://doi.org/10.1080/09296174.2015.1027890

Fernández-Dobao, A., & Herschensohn, J. (2019). The role of L1 literacy in late L2 learning: Evidence from Spanish-English bilinguals. *Studies in Second Language Acquisition, 41*(4), 797-821. https://doi.org/10.1017/S0272263119000206

Fernández Vítores, D. (2023). *El español: una lengua viva. Informe 2019.* Instituto Cervantes.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). SAGE Publications.

García Salido, M., & Garcia, P. (2018). Lexical complexity in L2 Spanish writing: A cross-sectional study. *Journal of Spanish Language Teaching, 5*(2), 134-150.

Garner, J., Crossley, S. A., & Kyle, K. (2019). N-gram measures and L2 writing proficiency. *System, 80*, 176-187. https://doi.org/10.1016/j.system.2018.11.007

Granger, S. & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching, 52*(3), 229-252. https://doi.org/10.1515/iral-2014-0011

Gries, S. T. (2013). *Statistics for linguistics with R: A practical introduction* (2nd ed.). Mouton de Gruyter.

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

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique*. Presses universitaires de France.

Guo, Y., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality using computational indices of linguistic features. *Modern Language Journal, 97*(1), 19–37.

Hao, Y., Jin, Z., Yang, W., Wang, X., & Liu, H. (2023). To predict L2 writing quality using lexical richness indices: An investigation of learners of Chinese as a Foreign Language. *System, 118*. https://doi.org/10.1016/j.system.2023.103123.

Honniball, C., Smith, J., & Lee, M. (2020). Lexical diversity and writing quality in L2 English: A corpus-based study. *Applied Linguistics, 41*(3), 345–367.

Hyland, K. (2003). Second language writing. Cambridge University Press.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19*(1), 57-84. https://doi.org/10.1191/0265532202lt220oa

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*(s1), 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing, 34*(4), 537-553. https://doi.org/10.1177/0265532217710632

Jarvis, S., & Hashimoto, B. (2021). Lexical diversity and the need for norms. *Annual Review of Applied Linguistics, 41*, 1-15. https://doi.org/10.1017/S026719052100001X

Johnson, R. L., Penny, J. A., & Gordon, B. (2000*). Assessing Performance: Designing, Scoring, and Validating Performance Tasks.* Guilford Press.

Jung, Y., Crossley, S., & McNamara, D. (2019). Predicting second language writing proficiency in learner texts using computational tools. *Journal of Asia TEFL, 16*(1), 37.

Kanwit, M. (2019). Mood selection and its relationship to lexical knowledge. *Applied Psycholinguistics, 40*(1), 1–29.

Kellogg, R. T., & Raulerson, B. A. (2007). Improving the writing skills of college students. *Psychonomic Bulletin & Review, 14*(2), 237–242.

Kim, Y. H., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency and writing quality. *Language Testing, 35*(3), 319–340.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing writing proficiency using lexical and grammatical features. *Journal of Second Language Writing, 28*, 1–11.

Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing, 34*, 12–24. https://doi.org/10.1016/j.jslw.2016.10.003

Kyle, K., & Eguchi, M. (2021). Bigram strength measures and L2 writing quality: A meta-analysis. *Language Testing, 38*(4), 567-590.

Kyle, K., Crossley, S. A., & McNamara, D. S. (2018). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing, 35*(4), 515–535.

Kyle, K., Crossley, S., & Jarvis, S. (2021). Assessing lexical sophistication using word, bigram, and dependency indices. *Modern Language Journal, 105*(1), 1-20. https://doi.org/10.1111/modl.12686

Kyle, K., Sung, H., Eguchi, M., & Zenker, F. (2024). Evaluating evidence for the reliability and validity of lexical diversity indices in L2 oral task responses. *Studies in Second Language Acquisition, 46*(1), 278–299. doi:10.1017/S0272263123000402

Kobayashi, H. and C. Rinnert. (2012). Understanding L2 writing development from a multicompetence perspective: Dynamic repertoires of knowledge and text construction. *L2 writing development: Multiple perspectives,* 101-134.

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal, 25*(2), 21-33. https://doi.org/10.1177/003368829402500202

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

Lee, J., Kim, Y. H., & Crossley, S. A. (2021). The predictive validity of lexical indices on L2 writing quality. *Language Learning, 71*(4), 1166–1200.

Leki, I., Cumming, A., & Silva, T. (2008). A synthesis of research on second language writing in English. Routledge.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7* (4), 328. http://www. rasch. org/rmt/rmt74m. htm.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? Rasch Measurement Transactions, 16(2), 878.

Nation, I. S. P. (2011). Research into practice: Vocabulary. *Language Teaching, 44*(4), 529-539.

Nicol, J., & Greth, D. (2003). Production of subject-verb agreement in Spanish as a second language. *Experimental psychology, 50*(3), 196.

Manchón, R. M. (Ed.). (2011). *Learning-to-write and writing-to-learn in an additional language.* John Benjamins Publishing Company.

Marrero, V., & Aguirre, C. (2003). Plural acquisition and development in Spanish. Linguistic theory and language development in *Hispanic languages,* 275-296.

Meara, P., & Bell, H. (2001). P-Lex: A Simple and Effective Way of Describing the lexical Characteristics of Short L2 Tests. *Prospect, 16*(3), 5-19.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381-392. https://doi.org/10.3758/BRM.42.2.381

Monteiro, K., Crossley, S. A., & Kyle, K. (2020). The development of lexical richness in L2 Spanish writing. *Hispania, 103*(1), 17-32. https://doi.org/10.1353/hpn.2020.0002

Montrul, S. (2004). *The acquisition of Spanish: Morphosyntactic development in monolingual and bilingual L1 acquisition and adult L2 acquisition.* John Benjamins Publishing Company.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555–578. https://doi.org/10.1093/applin/amp044

Paquot, M. (2018). Phraseological complexity in learner writing: A multidimensional approach. *International Journal of Corpus Linguistics, 23*(2), 200–226.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research, 35*(1), 121–145. https://doi.org/10.1177/0267658317694221

Plakans, L., Gebril, A., & Bilki, Z. (2019). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing, 36(*2), 161–181.

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language learning,* 471. 101-143.

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

Radden, G., & Dirven, R. (2007). *Cognitive English grammar*. John Benjamins Publishing.

Richards, J. C. (1987). The role of vocabulary teaching. *TESOL Quarterly, 21*(1), 83–94.

Read, J. (2000). Assessing vocabulary. Cambridge University Press.

Rubin, M., Paquot, M., & Bestgen, Y. (2021). Dependency bigrams and L2 proficiency: A corpus-based study of syntactic complexity. *Language Learning, 71*(S1), 267–306.

Sánchez-Gutiérrez, C. (2022). The UC Davis Corpus of Written Spanish, L2 and Heritage Speakers (COWSL2H). University of California, Davis. https://github.com/ucdaviscl/cowsl2h

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford University Press.

Schnur, E., & Rubio, F. (2021). Lexical diversity and L2 Spanish proficiency: A corpus-based study. *Hispania, 104*(2), 245-260.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*(2), 99-123. https://doi.org/10.1177/026553229401100202

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

Schafer, J. C. (2015). Vocabulary knowledge and its impact on L2 writing proficiency. *Journal of Second Language Writing, 29*, 1–15.

Tracy-Ventura, N. (2017). L2 Spanish writing development: A longitudinal perspective. *Spanish in Context, 14*(1), 89-112.

Tracy-Ventura, N., Mitchell, R., & McManus, K. (2021). The LANGSNAP corpus: A longitudinal study of L2 Spanish and French. *International Journal of Learner Corpus Research, 7*(1), 1-28.

Tversky, A. (1977). *Features of similarity. Psychological Review, 84*(4), 327–352

Ure, J. (1971). Lexical density and register differentiation. *Applications of Linguistics, 12*, 443-452.

Vold, E. T. (2023). Development of lexical richness among beginning learners of French as a foreign language. *Nordic Journal of Language Teaching and Learning, 10*(2).

Waldvogel, D. A. (2014). An analysis of Spanish L2 lexical richness. *Academic Exchange Quarterly, 18*(2), 8.

Whitley, M. S. (2002). *Spanish/English contrasts: A course in Spanish linguistics* (2nd ed.). Georgetown University Press.

Yule, G. U. (1944). *The statistical study of literary vocabulary*. Cambridge University Press.

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing, 47,* 100-128. https://doi.org/10.1016/j.asw.2020.100505

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language*. Harvard University Press.