CorreGram:
Using Corpus Data to Develop Student-Adaptable
Automated Corrective Feedback for the
L2 Spanish Language Classroom

By

SAMUEL S. DAVIDSON
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Linguistics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Kenji Sagae, Chair

Claudia H. Sánchez Gutiérrez

Zhou Yu

Committee in Charge

2024

To Angie and Papa.
Your undying support made this possible.
And to Mama.
I wish you could have been here to see this.

Contents

Abstract vi

Acknowledgments vii

Chapter 1. Introduction 1

1.1. Linguistic diversity in the language classroom 1

1.2. Automated Corrective Feedback 2

Chapter 2. Pedagogical Background 6

2.1. Learners studied 6

2.2. Automated Written Corrective Feedback 11

Chapter 3. Technical Background 16

3.1. NMT-based GEC 20

3.2. Sequence tagging-based GEC 22

3.3. Use of Augmented Learner Data 23

3.4. Application of Generative LLMs to GEC 24

Chapter 4. COWS-L2H Corpus 28

4.1. Current Spanish Learner Corpora 28

4.2. COWS-L2H 29

4.3. Longitudinal data 30

4.4. Error Annotation 31

4.5. Correction & Automated Error Tagging 32

4.6. Additional corpora used 33

4.7. Studies conducted using COWS-L2H 34

Chapter 5. COWS-L2H Study I: Lexical development 35

5.1. Measuring lexical development 35

5.2. Previous work 43

5.3. Methods and implementation 46

5.4. Discussion 53

Chapter 6. COWS-L2H Study II: Syntactic development 55

6.1. Motivation and approach 56

6.2. Linguistic Features 58

6.3. Experiments 62

6.4. Tracking Writing Skills' Evolution 63

6.5. Understanding Linguistic Predictors 65

6.6. Discussion 67

Chapter 7. COWS-L2H Study III: Errors by demographic group 69

7.1. Error annotation 69

7.2. Parallel corrected text 70

7.3. ERRANT 72

7.4. Method and implementation 76

7.5. Statistical analysis 77

7.6. Key differences and findings 78

7.7. Lower proficiency learners 79

7.8. Advanced L2 and Heritage learners 81

7.9. Discussion and implications 83

Chapter 8. GEC Approach & Implementation 85

8.1. Initial tests of error correction using COWS-L2H 86

8.2. Synthetic data generation 88

8.3. Synthetic data generation with LLM models 94

8.4. Model development and adaptation 95

8.5. Results 100

Chapter 9. Feedback Generation 105

9.1. Adaptability and Error Prioritization 106

9.2. Template-based feedback generation 107

9.3. LLM generated feedback 109

Chapter 10. AWCF Implementation and Use 113

10.1. Web App Design and Implementation 113

10.2. Provision of feedback to students 117

10.3. Initial Testing 120

Chapter 11. Conclusion 123

Appendix A. Llama 2 fine-tuning prompt 125

Appendix B. Claude multi-step feedback generation 129

Bibliography 132

Abstract

Automated corrective feedback (ACF), in which a computer system helps language learners identify and correct errors in their writing or speech, is considered an important tool for language instruction by many researchers. Such systems allow learners to correct their own mistakes, thereby reducing teacher workload and potentially preventing issues related to grammatical error fossilization. Research in this area has led to the development and widespread adoption of tools such as Grammarly for English learners. However, research in grammatical error correction (GEC) and other forms of ACF in languages other than English has been much more limited. This dearth of research is in part due to the large demand for English instruction, but is also driven by the limited training data available for non-English languages. However, a new corpus of learner Spanish collected at UC Davis, COWS-L2H, provided me with an opportunity to explore development of ACF for students studying Spanish. In my dissertation work, I explore the error patterns present in writing by students of Spanish in COWS-L2H, and use this information to inform a novel data augmentation technique to generate synthetic data for training language models capable of correcting learner errors in Spanish text. I then use this synthetic data, along with learner data from COWS-L2H, to train an AI-based GEC model for Spanish learners that is adaptable to learner L1 and proficiency level. Finally, I explore how this automatically corrected writing can be used to present feedback to learners in a pedagogically motivated way. To that end, I combine the GEC model trained using data from COWS-L2H with hand-written templates and feedback produced by generative LLMs to craft appropriate feedback for learners using the system. The end goal is a grammar-checker that is able to not only explain why something a student wrote is potentially incorrect, but is also able to guide the student to make the correction themselves. I demonstrate this novel system, CorreGram, and further discuss details of its implementation and proposals for how the system may be effectively utilized in the language classroom.

Acknowledgments

I would first like to thank my advisor, Dr. Kenji Sagae, without whom my academic journey would likely have taken a much different path. Kenji recognized my interest in the intersection of computation and language early on and has consistently provided the support and guidance I needed. I would also like to thank Dr. Zhou Yu, who early in my PhD career allowed me to dive headfirst into practical system development, bridging the gap between research and systems for real-world use, and Dr. Claudia Sánchez Gutiérrez, who brought me into the fold of the COWS-L2H project from the very start, giving me the data I needed to make this dissertation possible.

Thank you to Duolingo for your generous support of this project through the Duolingo Research Award.

I would like to thank my wife, Angela, who has been my most stalwart companion through this ups and downs of the PhD journey, and my father, Stewart Davidson, who helped lift me up from a low place to make this achievement possible. I would also like to thank my sister, Emily Davidson, and my brother, Brian Smith, who've always supported me. Thank you to my mother, Alice Davidson, who wasn't able to make it to see the end of this journey, but who helped, supported, and loved me until her last day. Though I never got the haircut she unceasingly requested. And finally, I would like to thank my friends at UC Davis who have helped along the way, both professionally and personally - Zoey Liu, Dian Yu, Tom Hardy, and Sophia Minnillo, among many others.

I couldn't have gotten here without the support of everyone mentioned above, or without the help of many other people. I hope you're happy with the results. Thank you!

CHAPTER 1

Introduction

Providing meaningful, constructive feedback to students as part of a second language education program is a challenging and time-consuming task for language instructors. The quality and appropriateness of feedback can be affected not just by the content of the feedback itself, but also by the proficiency level, linguistic background, confidence, and other individual characteristics of the target student. For this reason, not all feedback is equally beneficial to all students. This dissertation explores the diversity of students in a large university Spanish language program, how that diversity affects student writing and the types of errors students make, and finally discusses the development of an automated written corrective feedback system that takes these factors into account when generating corrective feedback for student writing. I show that effective feedback for learners of Spanish can be automatically generated with modern natural language processing techniques, and that taking the linguistic properties and error patterns found in real-world student data into consideration when implementing such a system results in better system performance and more targeted feedback for individual students.

1.1. Linguistic diversity in the language classroom

One key challenge faced by language instructors in a large university setting is the diversity of linguistic experience among students, which affects both student expectations and learning outcomes when studying a second, third, or Heritage language. In the United States, most universities enroll a large number of international students whose first language is not English. By far the largest group of such students is native speakers of Mandarin from China [iie.org]. Additionally, students from the United States come to the table with a diverse set of linguistic experiences. The most common non-English native language for American students is Spanish, given the large Spanish-speaking population in the US [Bureau, 2018]. This is especially evident in states with large Spanish-speaking populations, such as California [Statisticalatlas.com]. When these students (Spanish-speaking students who grew up in a non-Spanish-dominant area) are studying Spanish, which is the focus of the present dissertation, we refer to them as "Heritage learners." Thus, three primary groups of students are frequently seen in Spanish language instruction courses in American universities: Monolingual L1 (first-language) speakers of English, bilingual L2 (second-language) speakers of English (whose most common L1 is Mandarin, followed by Hindi), and Heritage speakers of Spanish who are also bilingual in English [Lacorte and Suárez-García, 2016]. Of course, with these broad definitions, some overlap between these student groups is likely. Given the differences in linguistic background in the student population, a major question arises: how does a student's prior language experience affect her acquisition of Spanish? While understanding these differences is important in its own right, a better understanding of the variation seen across the learner population may contribute to the development of improved teaching methods. Additionally, the underlying reason that I pursue these studies is to model specific aspects of variation between student groups which can be utilized in developing automated feedback systems, proficiency tests, and other evaluation tools which are tuned to students' linguistic experience. Although the question "how does linguistic experience impact language acquisition" is far too broad to address as a whole, researchers have begun to tease apart various aspects of this question. For example, Montrul and Ionin [2012] offers a review of experimental research comparing the acquisition of Heritage learners to that of L2 learners in terms of phonology, lexicon, and morphosyntax. While corpus studies examining specific aspects of Spanish Heritage, L2 and L3 (third or greater language) learner speech and writing are prevalent, corpus studies which explicitly study variation between these learner groups are much more limited, likely due to the sparsity of high-quality learner corpora of Spanish. This dissertation seeks to explore some of the differences in learning trajectories based on demographic factors like L1, previous language learning experience, and proficiency level. I then apply these findings to help inform a student-adaptable written corrective feedback model for students learning Spanish.

1.2. Automated Corrective Feedback

Automated corrective feedback (ACF), in which a computer system helps language learners identify and correct errors in their writing or speech, is considered an important tool for language instruction by many researchers [Li et al., 2015, Tatawy, 2002, Ranalli, 2018]. ACF tools can provide both automated synchronous (provided while students write) and asynchronous (provided after writing is complete) feedback to students, allowing learners to correct their own mistakes [Shintani, 2016], thereby reducing teacher workload and potentially preventing issues related to grammatical error fossilization [Tajeddin et al., 2017]. This research has led to the development and widespread adoption of tools such as Grammarly[1] for English learners. However, research into the development of grammatical error correction (GEC) and other forms of ACF for the classroom in languages other than English has been quite limited. This dearth of research is in part due to the large demand for English instruction, but is also driven by the limited training data available for non-English languages. Even for English, the amount of training data is often not sufficient to train the large neural models now used for GEC. To overcome these challenges, I propose a method for building synthetic training data that more closely replicates the error patterns seen in learner data.

Specific steps are necessary in generating synthetic data which attempts to replicate the lexical, syntactic and error distributions of Spanish learners at different levels of proficiency and with diverse linguistic backgrounds. First, I need a large quantity of text from corpora or online sources which are somewhat similar to learner text in terms of length, lexical patterns and use of various morphosyntactic forms, such as verb tense and subordination. The most obvious source of such data is real learner text drawn from existing unannotated, uncorrected learner corpora, such as CEDEL2 [Lozano et al., 2009]. However, even the largest error-annotated learner corpora (all of which are in English), such as the Cambridge English Write & Improve + LOCNESS corpus [Bryant et al., 2019], are generally considered too small to train an automated feedback system from scratch [Grundkiewicz and Junczys-Dowmunt, 2019]. Even with the advent of large pretrained models such as T5 [Raffel et al., 2020a], the most successful systems for GEC are trained in a multi-stage process that involves using synthetic data to align the model with the GEC task [Rothe et al., 2021]. When attempting to train such a GEC model for Spanish, the size and number of available learner corpora drops dramatically. Thus, much training text must be selectively extracted from web data, such as that available in the Corpus del Español [Davies, 2016]. However, this data is unlikely to replicate the linguistic patterns seen in learner writing. In an attempt to remedy this situation, I propose to filter the data based on the distribution of lexical and syntactic features to generate a large corpus of text which "looks like" learner data. Once I have gathered a corpus of such text, I need to inject errors into the text which replicate, as nearly as possible, the distribution of errors seen in real

[1]https://www.grammarly.com/

learner data. In order to achieve these goals, I must first identify, as specifically as possible, how the target student groups vary across several dimensions: lexical usage, syntax, and error rates.

While recent work [Li et al., 2015, Stevenson and Phakiti, 2014] has shown that automated corrective feedback shows promise as a tool for improving student writing competency, a key criticism of current automated corrective feedback systems for language education is the fact that most such systems take a "one-size-fits-all" approach to identifying and suggesting corrections for potential errors. That is, many systems are designed with little-to-no consideration of the proficiency and error patterns of the students who will be using the system, rather targeting some abstract "typical" student. Thus, I propose to use student error patterns to inform a GEC-based automated written corrective feedback (AWCF) system for adult learners of Spanish, particularly those learning Spanish in an academic setting, that is adaptable to a student's L1, proficiency, and previous language learning experience. The proposed system is, to my knowledge, the first GEC-based, data-driven error correction system designed for the Spanish classroom. I argue that an automated feedback system which is better able to adapt to both student linguistic experience (such as that presented in Nadejde and Tetreault [2019]) and the instructor's pedagogical goals will prove beneficial to students and result in better learning outcomes. The code for the proposed feedback system is available on GitHub[2].

In order to better understand the linguistic diversity of students in a university Spanish language program, I first report on several studies designed to answer the following questions:

(1) Does the lexical diversity and density of Heritage learners differ significantly from that of advanced L2 learners, or from that of Spanish-dominant individuals?

(2) Are syntactic patterns extracted from longitudinal data sufficient to reveal differences in syntactic usage across proficiency levels?

(3) How do the error rates of English-dominant L2 learners of Spanish differ from those of non-English-dominant L3 counterparts?

(4) How do the error rates of Heritage learners of Spanish differ from those of their advanced L2 counterparts?

Once I have established a baseline understanding of these factors, I seek to answer the following research questions related to automated corrective feedback for Spanish language learners:

[2]https://github.com/ssdavidson/gec_feedback_spanish

(1) Does a grammatical error correction system that is adapted to student attributes, such as L1 and proficiency, identify and correct errors more effectively than a baseline, "one-size-fits-all" system that is otherwise identical?

(2) Does synthetic data built to replicate the linguistic and error patterns observed in student data improve the GEC performance of a large language model (LLM) relative to a model fine-tuned on real-learner data alone?

(3) How can we effectively present the errors identified by the GEC model to learners in a pedagogically motivated manner?

In the following chapters, I answer these questions by demonstrating the linguistic diversity found across learners in a large university language program. I then show that by taking this diversity into account when developing automated feedback systems for students, I am able to create a writing assistant that is effective in correcting student errors and is able to provide targeted feedback to language learners.

CHAPTER 2

Pedagogical Background

To provide context for the studies presented in this dissertation, this section first discusses the student populations studied to establish the potential utility of student-adaptable AWCF systems, as well as language transfer in these student populations. Additionally, I discuss the potential benefits and drawbacks of using AWCF in the second language classroom, including studies of the long-term impact of using AWCF and students' tolerance to system errors.

2.1. Learners studied

2.1.1. Heritage Learners. As the Spanish speaking population of the United States increases, Heritage learners of Spanish - students who come to the classroom with exposure to Spanish from the home environment - make up an increasingly large segment of many Spanish program's students [Montrul, 2010a]. To understand the number of potential Heritage learners in American universities, it is estimated that approximately 13.3% of Americans over the age of 5 speak Spanish at home, with the population largely concentrated in the West and Southwest United States [Bureau, 2018]. For example, 28.8% of Californians speak Spanish at home [Statisticalatlas.com] and 24.8% of undergraduates enrolled in the University of California system self-identify as Hispanic [Paredes et al., 2021]. While not all students who self-identify as Hispanic are speakers of Spanish, these numbers make clear the overall prevalence of the Spanish-speaking community in many higher education institutions. While Spanish-English bilinguals have long represented a significant proportion of students at many universities, Spanish departments have become more cognizant of the fact that these students have needs which differ from those of their L2 learner peers. In response to this need, Spanish departments at many Hispanic-serving institutions have developed courses specifically designed to help Heritage learners of Spanish to retain Spanish and to introduce these learners to Spanish in an academic register. A Google search reveals that, while the majority of schools which offer Heritage Spanish are located in the western United States, Heritage courses are now being offered at schools across the country.

Although Spanish departments have sought to address the specific needs of Heritage learners by developing courses which cater to them, this one-size-fits-all approach for Heritage learners is problematic, as Heritage learners are a linguistically and culturally diverse group whose proficiency in Spanish, and the dialect and register of Spanish they use, varies widely from individual to individual. The diversity of Heritage learners can be seen in the many definitions of "Heritage learner" or "Heritage speaker" which have been offered by researchers. The most widely used definition, offered by Guadalupe Valdés, defines a Heritage Learner as a "language student who is raised in a home where a non-English language is spoken, who speaks or at least understands the language, and who is to some degree bilingual in that language and in English" [Polinsky and Kagan, 2007]. While this definition is useful to language educators in particular, many researchers consider this view of Heritage learners to be too narrow. Polinsky and Kagan [2007] point to two conceptions of Heritage learners which have been proposed in the literature, and which they term "broad" and "narrow" definitions. The "broad" conception emphasizes the connection between cultural and linguistic heritage. For example, Fishman [2001] and Van Deusen-Scholl [2003] both argue that a student's status as a Heritage learner should be based on their familial and cultural connections to the language in question. Van Deusen-Scholl refers to such students as "learners with a heritage motivation" [Van Deusen-Scholl, 2003]. Polinsky and Kagan argue that the view taken by Fishman and Van Deusen-Scholl is too broad, since it focuses on a student's motivation for studying a language, rather than on linguistic knowledge. For example, a student who is learning a language for the first time as an adult may be culturally motivated to do so, but that does not make that student a Heritage learner/speaker under Valdés' definition, which requires that the student actually acquired the language in question in the home. This dissertation will use the "narrow" definition of Heritage learners proposed by Polinsky, and used by other researchers interested in Heritage language as a form of early bilingualism such as Rothman [2009].

In addition, Lynch [2008] argues that the implementation of Heritage learner programs has somewhat impeded research into Heritage speakers' abilities, especially with reference to L2 learners, by implying a dichotomous relationship between these two groups, when no such dichotomy exists. Heritage learners are often treated as native speakers when this label may not reflect the true extent of their linguistic abilities [Lynch, 2008]. For example, Rothman [2009] notes that the linguistic competence of Heritage speakers "will differ from that of native monolinguals of comparable age."

Silva-Corvalán [1994] states that Heritage speakers of Spanish in the United States exist along a continuum ranging from standard Spanish to limited, "emblematic" usage expressing social and cultural identity. Silva-Corvalán also points out that many Heritage speakers exhibit limited domain knowledge of their Heritage language, as the use of minority languages is frequently limited to domestic and social interactions. This limited domain knowledge results in difficulty with abstract topics, such as politics and science, and with complex syntactic constructions for many Heritage speakers [Lynch, 2008]. Further, Lynch argues that the limited social use of minority languages results in simplified grammatical systems that introduce "innovative, that is, non-normative, elements at the lexical and discourse levels," and that these innovative patterns are conditioned by the dominant language [Lynch, 2008]. That is, Heritage language tends to adopt usage patterns from the majority language surrounding it. These influences result in Heritage learners speaking a tongue that can be markedly different from that spoken by non-Heritage native speakers of the language. O'Grady et al. [2011] point out that past research has shown many Heritage learners to have specific linguistic deficiencies when compared to native speakers. The extent of this type of borrowing and "deficiency" is debated, though. For example, Pousada [1979] find that there is no evidence of convergence between the verb systems of Puerto Rican Spanish and English in a language contact situation. Conversely, Montrul [2004] argues for convergence in the subject and object expression of adult simultaneous bilinguals of English and Mexican Spanish in the United States. Thus, the impact of bilingualism on the language of Heritage speakers, and which linguistic features are most subject to convergence, remains an open question.

It should be noted that the majority of students whom I consider in my evaluation of Heritage learners report speaking only Spanish at home. Thus they are likely highly proficient in Spanish as a spoken language, and are enrolled in the Heritage course series primarily to acquire a more academic register of Spanish and to improve their writing skills. Thus, these learners' backgrounds may not align with the subjects studied in much of the previous literature on Heritage language acquisition and use. That said, the COWS-L2H corpus also contains a number of students who report some degree of Spanish exposure in the home, but who choose to enroll in the non-Heritage course series designed for less proficient learners; however, I do not include these students in the the present analysis due to need to more thoroughly examine the linguistic background of individual participants.

2.1.2. L3 Learners of Spanish. Universities in the United States have, on the whole, a large number of international students whose first language is not English. During the 2018-2019 academic year, approximately 1.1 million international students were enrolled in U.S. degree-granting institutions [iie.org], accounting for 5.5% of total U.S. higher education enrollment. Among these students, the largest sub-population is students from China; approximately 370,000 Chinese students were enrolled in U.S. universities in 2019, accounting for approximately 34% of all international students. The proportion of Chinese students is much higher at some institutions; for example, as of Fall 2023, 18,602 undergraduate students from mainland China enrolled in the University of California system; this means that approximately 6.3% of UC undergraduates are Chinese nationals. By contrast, India, which is the next highest student-sending country, sent only 4,461 undergraduates in Fall 2023 [universityofcalifornia.edu, 2023].

This large population which speaks Mandarin and related Chinese languages poses a challenge to language instructors due to possible effects of these students' L1 on the way that the students learn the target language. When dealing with such a large population of students whose L1 is not English, do language instructors need to modify their teaching methods to accommodate L1 transfer that is different from that of the majority English-speaking students [Cummins, 2008]? For example, Mandarin does not use articles in its syntactic system in the same way as many Indo-European languages, including English and Spanish [Snape, 2009]. When English-speaking students are learning Spanish, they have prior exposure to the use of articles in their L1, despite differences in the gender systems of English and Spanish [Ionin and Montrul, 2010]. Mandarin speakers, on the other hand, come to Spanish from an L1 which has given them little experience with the use of articles in their native tongue [Ionin and Montrul, 2010]. Does this difference affect the way in which Mandarin speakers acquire the use of Spanish articles relative to their native English counterparts? But, one must also consider the fact that these students have been exposed to the use of articles through learning English, which all have presumably studies as an L2 as a prerequisite to admission to an American university. How does the students' knowledge of article usage from their L2 interact with syntactic transfer from their L1 [Cai and Cai, 2015]? Which linguistic system has the greater impact on these students' acquisition of Spanish as an L3 [Rothman and Cabrelli Amaro, 2010]? To answer these questions, an in-depth analysis of their writing is needed.

2.1.3. Linguistic transfer. While many studies have considered the effect of L1 syntactic transfer on students who are acquiring an L2 [Karim and Nassaji, 2013], fewer studies have considered these effects when a student is learning an L3 in an L2 medium classroom. However, those studies that have analyzed the effect of linguistic transfer in the context of L3 learning have disagreed about the effect of L1 transfer when acquiring an L3 language. For example, Rothman and Cabrelli Amaro [2010] and Bardel and Falk [2007] investigate four competing views of the effect of L1 transfer on L3 acquisition:

(1) an L1 effect for all adult language acquisition [Schwartz and Sprouse, 1996].
(2) the L2 transfer, or status factor, hypothesis, which posits that learning an L2 blocks L1 transfer when acquiring an L3 [Williams and Hammarberg, 1998].
(3) the Cumulative Enhancement Model [Flynn et al., 2004] which hypothesizes transfer from all previously learned languages.
(4) the Typological Primacy Model [Rothman, 2011] which predicts that L3 learners will transfer the linguistic system from their L1 or L2 based on which system is typologically most similar to the L3 they are learning.

Rothman and Cabrelli Amaro [2010] provide data which supports the L2 status factor hypothesis, arguing that the acquisition of an L2 can, in certain cases, effectively block L1 transfer when learning an L3. Similarly, Bardel and Falk [2007] state that "syntactic structures are more easily transferred from L2 than from L1 in the initial state of L3 acquisition." The results of the present analysis of Spanish L3 learner writings also seem to support these findings.

2.1.4. Effect of target language proficiency. While the linguistic background of learners is one major consideration in the types of errors learners make, another key component to developing an effective AWCF system is adapting to the proficiency of learners [Ranalli, 2018, Bitchener and Ferris, 2012, Nadejde and Tetreault, 2019]. For example, Bitchener and Ferris [2012] argue that unfocused feedback - feedback which suggests that an error has been made without suggesting a solution - may result in cognitive overload for lower-proficiency students. On the other hand, more advanced students may benefit from unfocused feedback in that it does not seek to constrain their writing style; rather, it forces them to decide how to best resolve a potential error themselves, thereby reinforcing language learning [Bitchener and Ferris, 2012]. While a first year Spanish student may make many grammatical and stylistic errors in their writing, pointing out errors that are beyond the scope of their learning objectives may serve to confuse the student and reduce the utility of the AWCF system. This weakness can be overcome by designing a GEC-based AWCF system that provides output aligned to student L1, proficiency, and specific pedagogical goals. Additionally, the limited research available that investigates the role of corrective feedback in improving the use of specific linguistic forms has shown that simple errors, such as gender agreement and use of past tense, are more readily learned from corrective feedback [Bitchener and Ferris, 2012], though it should be noted that the study in question used teacher-provided asynchronous feedback rather than synchronous AWCF.

The linguistic diversity and differing proficiencies of students brings to light one of the major criticisms of most AWCF systems: that they use a "one-size-fits-all" approach which "takes little or no account of individual differences" [Ranalli, 2018]. According to Ranalli [2018], "Current-generation AWE tools are not designed to differentiate among users' with different L2 proficiencies, L1s, writing skills, or educational backgrounds." This weakness of many AWCF tools leads to the issue, pointed out by Koltovskaia [2020], that more proficient students under-utilize and lower proficiency users over-rely on AWCF recommendations. Bitchener and Ferris [2012] specifically called on researchers to investigate the impact of L2 proficiency and other student-specific factors on the utility of AWCF for language learning.

2.2. Automated Written Corrective Feedback

When proposing to build a ACWF system for L2 and Heritage language learners of Spanish, one must ask how such a system will benefit students and instructors in the classroom setting. According to Ferris [2012], a longstanding debate exists around the benefits of corrective feedback (in most studies, provided by instructors) to L2 student writing development. For example, Truscott and Hsu [2008] finds no lasting improvement to the number of grammatical and lexical errors in student writing when students who have been previously provided corrective feedback are asked to write without the aid of feedback. However, Bitchener and Ferris [2012] more closely investigate Truscott and Hsu's claims and find that, over time, students who are given corrective feedback actually do make fewer lexical and grammatical errors. Corrective feedback is considered an essential part of second language learning by many researchers. For example, Gass [1991] and Ellis [2002] see CF in the role of "noticing"; in order to acquire a second language, learners must be able to notice the differences between their production and the correct form in the target language [Tatawy, 2002]. Of additional note is the fact that very little research exists related to the use of written corrective feedback in Heritage learner writing instruction. For example, Park et al. [2016] focuses on the provision of corrective feedback to Heritage learners of Korean whose relative proficiency in Korean is much lower than the proficiency of those Heritage learners of Spanish involved in the present research.

So, the benefit of corrective feedback in terms of improving student grammar and lexical choice remains an open question. But, as any writing instructor would tell you, there is much more to writing than grammar and word choice. As Ferris [2012] points out, "L2 writing is a (sub)discipline informed by language and composition studies as well as other disciplines, including rhetoric, communication, and cultural studies, among others." Even if a student is able to write a grammatically "perfect" essay, if her communicative and rhetorical skills are lacking, the composition may still be ineffective. It is teaching these communicative and rhetorical skills that Truscott and Hsu [2008] and Truscott [1996] argue should be the goal of instructor feedback in the L2 writing classroom. However, students sometimes "fail to meet practical goals because of their lack of progress in producing more linguistically accurate texts" [Ferris, 2012], thus supporting the utility of feedback related to grammatical and lexical choices. However, Ferris [2012] further points out that most writing instructors view corrective feedback as one tool in a larger toolbox of pedagogical methods designed to improve overall student writing effectiveness.

As with the use of corrective feedback in general, the use of automated written corrective feedback in the second language classroom has a relatively short but controversial history. Providing CF to students, especially in a written format, is an *extremely* time-consuming prospect for instructors [Shintani, 2016], and the potential automation of the provision of feedback concerning grammatical and lexical errors can free up instructor time to more effectively focus on the providing instruction related to rhetorical and composition skills [Li et al., 2015]. Particularly, AWCF can provide the type of real-time feedback to students which is simply impossible for instructors to provide, thus allowing students to immediately take advantage of the proposed suggestions and gain more confidence in their independent expressive abilities [Barrot, 2021]. Heift and Hegelheimer [2017] further explain that the "usefulness of computer-generated corrective feedback largely lies in enabling learner self-study and practice of the target language by identifying and explaining error sources and, with regard to L2 essay writing, allowing for draft revision". This type of rapid feedback is in many ways similar to the "dynamic feedback" method described in Hartshorn et al. [2010], in which students write frequent short essays, are provided with immediate feedback, and given an opportunity to revise their writing based on this feedback. According to Hartshorn et al. [2010], this type of feedback is particularly effective in instilling lasting benefits to student writing. AWCF, such as the model I propose, allows for exactly this type of real-time corrective feedback which would otherwise be unavailable to students. However, some researchers [Cheville, 2004] and teacher groups, such as the National Council of Teachers of English [NTCE, 2014] oppose the use of ACF and other types of computer-mediated automated assessment. According to the NTCE:

> Automated assessment programs do not respond as human readers. While they may promise consistency, they distort the very nature of writing as a complex and context-rich interaction between people. They simplify writing in ways that can mislead writers to focus more on structure and grammar than on what they are saying by using a given structure and style [NTCE, 2014].

In a study motivated by the NTCE statement and other critics of the use of automated feedback, Li et al. [2015] demonstrate the utility of automated written feedback in improving student writing, as indicated by both number of correct revisions made by students and by teacher evaluation of the feedback system. Additionally, Stevenson and Phakiti [2014] demonstrate modest improvements to error rates in student texts written using AWCF; however, the authors note that there is little evidence that these improvements transfer to the students' writing when not using AWCF [Li et al., 2015, Bitchener and Ferris, 2012]. Stevenson and Phakiti [2014] state that more research is needed to establish that AWCF actually leads to improvement in overall student writing proficiency.

Another key criticism of current AWCF systems is that they use a "one-size-fits-all" approach that "takes little or no account of individual differences" [Ranalli, 2018]. According to Ranalli [2018], "Current-generation AWE tools are not designed to differentiate among users' with different L2 proficiencies, L1s, writing skills, or educational backgrounds." This weakness of many AWCF tools leads to the issue pointed out by Koltovskaia [2020] in which more proficient students under-utilize and lower proficiency users over-rely on AWCF recommendations. The prevalence of the "one-size-fits-all" approach to AWCF is despite previous work in the use of corrective feedback generally which indicates that corrective feedback is most effective when instructors consider not only proficiency, but also factors such as previous language learning experience and the nature of the learning environment [Ferris and Hedgcock, 2013]. This finding raises an important question: how can AWCF systems, specifically those driven by GEC, be better adapted to differences in learner groups, such as proficiency level, L1, and previous language learning experience, to improve the ability of students to utilize AWCF systems? For example, Bitchener and Ferris [2012] specifically called on researchers to investigate the impact of L2 proficiency and other student-specific factors on the utility of AWCF for language learning. Additionally, it should be noted that nearly all research in ACF has focused on English as a second language; this dissertation proposes a GEC-based AWCF system built for use by Spanish language learners. However, the proposed method should be applicable to any language for which a small amount of annotated learner text is available for error distribution analysis, and for which a larger amount of unannotated text is available for generation of synthetic training data.

2.2.1. Synchronous and asynchronous feedback. One of the major advantages of AWCF, beyond the potential time savings to instructors [Stevenson and Phakiti, 2014], is the fact that it facilitates synchronous feedback to student writers - that is, tagging of errors and suggesting corrections in near-real-time while students are writing [Dikli, 2006]. In a study which analyzed differences between synchronous (SCF) and asynchronous (ACF) corrective feedback, Shintani [2016] found:

> (1) SCF created an interactive writing process similar in some respects to oral corrective feedback; (2) both the SCF and ACF promoted noticing-the-gap, but self-correction was more successful in the SCF condition; (3) focus on meaning and form took place contiguously in the SCF condition while it occurred separately in the ACF condition; and (4) both types of feedback facilitated metalinguistic understanding of the target feature, reflecting the unique features of writing (i.e., its slow pace, its permanency and the need of accuracy) [Shintani, 2016].

Shintani [2016]'s study did not use AWCF to provide synchronous feedback, but rather used Google Docs to allow instructors to provide synchronous feedback as students wrote. However, their findings clearly highlight the benefits of synchronous feedback, assuming that the feedback provided is of sufficiently high quality and well targeted to the student's proficiency level.

2.2.2. Error tolerance in AWCF. The question of high quality feedback raises an additional question regarding the use of GEC-based AWCF in the language classroom: what is the error-tolerance of students and instructors using AWCF systems. Unfortunately, I am unable to find significant research into how much error students are willing to accept when using AWCF and how much errors made by such systems impact their usability. The fact is that the errors identified and corrections suggested by a GEC model will include some number of errors; for example, one state-of-the-art GEC system, GECToR [Omelianchuk et al., 2020] achieves a precision of 78.9 and a recall of 58.2 (for an $F_{0.5}$ 73.6) on the BEA-2019 English GEC shared task test set [Bryant et al., 2019]. The relatively low recall for these systems, indicating that the systems are missing a large portion of the errors identified by human annotators in the texts, is one reason that general-purpose, domain agnostic GEC systems have not been fully integrated into commercial AWCF systems such as Grammarly and Criterion. While the usability of general-purpose GEC for AWCF may still be somewhat limited by system performance, a more fine-grained analysis is necessary to determine which error types these systems are good at identifying and correcting, and if their performance with this subset of error types outperforms more well-established and less computationally expensive statistical and rule-based models. This analysis should be contrasted with the types of corrections made by instructors. Instructors are unlikely to correct every error in a student text, either to target particular learning objectives or to avoid overwhelming students [Hendrickson, 1980]. Similarly, an automatic system that does not catch every error may still be quite useful, if the system is consistent, targets more impactful errors types, and shows high precision.

CHAPTER 3

Technical Background

In addition to the practical and pedagogical considerations outlined in the previous chapter, several technical aspects of data-driven grammatical error correction influence the development of workable GEC systems for the language classroom. Data-driven GEC, the machine learning problem of identifying and correcting writer errors in text, is a difficult task due to the non-deterministic nature of error correction. For many types of errors, there are multiple possible ways to correct the target error; how to resolve the error is largely a matter of choice for the corrector. Traditional methods of dataset and prediction evaluation, such as inter-annotator agreement and BLEU score [Papineni et al., 2002], are not applicable to the GEC task [Bryant and Ng, 2015] due to the fact that agreement between annotators or alignment to a specific correction is not necessary for an edit to be an appropriate correction of an error. While the correction of some errors is relatively straightforward, such as subject-verb agreement, even these scenarios are open to interpretation and subject to corrector choice. Should we change a pronoun to make it agree with the verb, or change the verb to agree with the pronoun? For example, should the agreement error in “They reads the paper” be corrected to “They read the paper” or to “She reads the paper”? While many instructors may assume the student used the wrong verb form, an incorrect choice of pronoun is a reasonable alternative. More difficult decisions arise when considering non-standard forms that, while acceptable in certain registers or dialects, do not conform to an academic standard. For example, do we correct a student who writes “I get to work at 8am” versus “I arrive at work at 8am”? While this choice may be considered stylistic rather than grammatical [Fraser and Hodson, 1978], the line between the two is often not clear. These questions are of particular relevance in the L2 teaching field, in which grammatical error correction tools have the potential to provide rapid feedback to students and expand instructor bandwidth.

Fraser and Hodson [1978], make the following distinction between grammar and usage:

> Each language has its own systematic ways through which words and sentences are assembled to convey meaning. This system is grammar. But within the

general grammar of a language, certain alternative ways of speaking and writing acquire particular social status, and become the conventional usage habits of dialect groups.

While this distinction is easy to maintain in rule-based GEC systems, neural and statistical systems learn transformations from training data; thus, if a correction to style or usage is prevalent in the training data, that correction will likely be enforced by the resulting system. This is particularly true of language model based systems which learn the structure of the language in an unsupervised manner from large amounts of unannotated text.

In an effort to ameliorate the issue that multiple possible valid corrections causes when evaluating GEC systems, several benchmark GEC datasets, particularly those commonly used to evaluate English GEC, provide corrections written by several independent annotators. For example, the BEA 2019 dataset [Bryant et al., 2019] provides three reference corrections for each sentence in its evaluation set. This approach improves the odds of an edit generated by the GEC system being evaluated being present in at least one of the reference sentences, and thus judged as valid [Bryant et al., 2019]. To demonstrate this point, Bryant and Ng [2015] augmented the CoNLL 2014 GEC dataset [Ng et al., 2014] (which contains two reference corrections for each evaluation sentence) with eight additional reference corrections, for a total of ten possible corrections for each evaluation sentence. They evaluate their system using the $\mathrm{F}_{0.5}$ measure, which weights recall at half the value of recall, indicating a preference for systems that may miss corrections, but are more accurate with the predictions they make. See Formula 3.1 for a formal definition. With their additional reference corrections, Qorib and Ng [2023] achieve an $\mathrm{F}_{0.5}$ of 85.21, as opposed to 71.12 on the original two-reference evaluation set. This improvement makes clear a key weakness in GEC evaluation, and demonstrates that GEC systems that are performing "poorly" on some evaluation set, may be generating far more valid correction edits than the evaluation metrics would indicate. It is important to understand this weakness in current evaluation methods of GEC systems when deciding if a system is "good enough" for practical use, and that in all cases human evaluation should be the final judge of a system's efficacy.

$$F_{0.5} = ((1 + 0.5^2) * Precision * Recall)/(0.5^2 * Precision + Recall) \tag{3.1}$$

3.0.0.1. *Rule-based and Classifier-based GEC.* The simplest GEC systems, and, until quite recently, those most commonly encountered by end-users in tools such as Microsoft Word and Google Docs, use hand-crafted rules, regular expressions, and classification to correct specific grammatical errors. For example, a simple regular expression can be used to enforce the correct use of "a" or "an" depending on the following phoneme. Similarly, spelling correction can be implemented by checking words in a text against a dictionary list, then identifying the most-similar word based on edit-distance. However, many grammatical errors, such as subject-verb agreement mismatches, are too complex to be identified and corrected with simple string matching rules. To overcome this challenge, most rule-based GEC systems take advantage of part-of-speech (POS) information and parse trees derived from automated tagging and parsing algorithms [McCoy et al., Sidorov et al., 2013]. For example, as discussed in Bryant et al. [2017] and Sidorov et al. [2013], subject-verb mismatches can be identified by a rule which checks that the nominal subject (*nsubj*) of the sentence has the same person and number as the verb. The grammatical relation between the subject and the verb can be obtained from a dependency parser, while the person and number are often encoded in POS tags. The example in Figure 1.1 demonstrates how dependency and POS tag data can be used in rules to identify grammatical errors. Sidorov et al. [2013] further demonstrates the use of a rule to identify the correct verb form from a verb list to resolve identified subject-verb agreement errors. Rules to correct other types of errors can be constructed in a similar way.

A major benefit of rule-based GEC systems is that they do not require training data to implement. This fact allows rule-based systems to be quickly built to identify and correct specific errors in low-resource languages and domains with little available training data [Bryant et al., 2017]. Additionally, the rules tend to be straightforward to implement and can be precisely targeted [Bryant et al., 2017]. Such rules are effective when designing a system which targets a specific type or set of errors. However, designing a system which is able to identify and correct a broad range of errors can quickly become unfeasible due to the number of rules, as each error type requires its own set of potentially complex hand-crafted rules.

While rules must be implemented manually, statistical classifiers attempt to learn the function which determines the appropriate token in a given position from training data. Like rule-based methods, classifier-based methods must restrict the class of errors which they attempt to correct in order to limit the number of categories into which a token can be classed [Bryant, 2019]. Much work has focused on the correction of errors in article, verb, and preposition usage [Rozovskaya and Roth, 2014]. These parts-of-speech are particularly appropriate for classifier-based GEC, as the number of potential targets is relatively small. For example, an article classifier only need determine if the article before a verb should be *a*, *an*, *the*, or omitted. Similarly, a verb classifier, which relies on information drawn from verb inflection databases, need only determine whether the given verb *eat* should be classified as *eat*,*eats*, *ate*, *eating*, or *eaten*. Obviously, this task becomes more complex in languages which have richer verb morphology. The learning of the classification function can be done using various statistical approaches, such as naive Bayes, logistic regression, decision trees, support vector machines, and other statistical methods [Bryant, 2019].

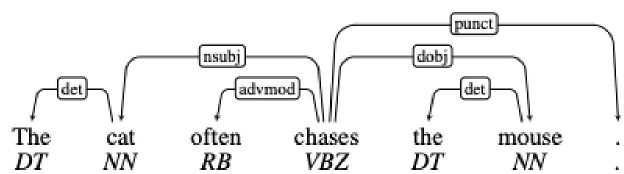

FIGURE 3.1. "cat" is the nominal subject (*nsubj*) of "chases" and the POS tags show that both words are singular (*NN* (singular noun) and *VBZ* (third person singular verb)) rather than *NNS* (plural noun) and *VBP* (non-3rd person present). Figure taken from Bryant et al. [2017].

3.0.0.2. *Language Model-based GEC.* Using language models to identify and correct errors relies on the fact that, when analyzed using a well-trained language model, the probability of an ungrammatical sentence should be lower than that of a grammatical one [Bryant and Briscoe, 2018]. For example, the probability of the sentence *"I seed the cat" should be lower than the probability of "I saw the cat." Current techniques in LM-based GEC involve correcting errors for only a limited subset of items in a given sentence; for example, Bryant and Briscoe [2018] target only non-word errors (spelling errors which result in a non-word), morphological errors such as noun number and verb tense, and articles and prepositions. The method used by Dahlmeier and Ng [2012], Lee and Lee [2014] and Bryant and Briscoe [2018] involves creating confusion sets for each token in a sentence which falls into one of their predefined categories based on the part-of-speech of the token. They create the confusion sets using various external resources, such as spell-checkers for non-words and inflection databases for morphological errors. Once the confusion sets have been generated, they iterate through the various changes proposed in the confusion sets and re-score the sentence using the trained LM. They then choose the resulting sentence with the lowest LM perplexity score as the best correction of the target sentence [Bryant and Briscoe, 2018].

Most available work which uses LMs to rank proposed corrections use N-gram LMs such as KenLM [Heafield, 2011]. However, more recent work utilizes large neural language models like BERT [Devlin et al., 2018] to rerank proposed corrections [Kaneko et al., 2019], or to generate corrections using BERT's masked LM framework [Li et al., 2020]. Li et al. [2020] propose a two-stage process in which they first label each token in the sequence with one of four labels: remain, substitution, insert, delete. They then mask all tokens labeled "substitution" and insert a mask token for the "insert" labels, allowing the BERT model to propose corrections for these items.

They key advantage of LM-based GEC is that it does not require annotated training data of human-corrected sentences; confusion sets can be built for any language for which the necessary linguistic resources such as spell-checkers exist. In Bryant and Briscoe [2018]'s system, annotated data is used only for system tuning, though system tuning is not a strict requirement of an LM-based GEC system. The drawbacks of this type of system are that 1) they require linguistic resources such as spell-checkers and inflection databases to generate confusion sets necessary to generate alternate proposals for scoring; and 2) since theses systems, at least as proposed, make changes only on the token level, they are not capable of correcting multi-word grammatical and stylistic errors. Additionally, as pointed out by Bryant [2019], probability is not always a perfect proxy for grammaticality; for example, the sentence "I is the ninth letter of the alphabet" would have a lower probability than "I am the ninth letter of the alphabet" according to most LMs, despite the fact that the former sentence is actually the more appropriate in context [Bryant, 2019].

3.1. NMT-based GEC

The most common approach to the GEC task in recent literature frames error correction as a monolingual machine translation task in which the source and target languages are "language with errors" and "language without errors," respectively [Ng et al., 2014]. Grammatical error correction can be viewed as a noisy-channel model, a task to which machine translation is particularly well suited [Flachs et al., 2019]. Specifically, an input sentence containing errors can be said to be a corrupted version of its grammatical counterpart that has been passed through a noisy channel.

The goal of the GEC task, then, is to reconstruct the correct sentence from the erroneous input [Bryant et al., 2023].

Because GEC is often framed as a machine translation task, the development of GEC systems has largely mirrored progress in MT more broadly. The first MT-based models for GEC were built using statistical machine translation (SMT) frameworks (for example, Yuan and Felice [2013], Felice et al. [2014], Junczys-Dowmunt and Grundkiewicz [2014]). While these systems were a major improvement over the previous rule-based and classifier-based GEC systems, they often generate ungrammatical output and make unnecessary corrections to align their output with common wording seen in their training data [Bryant et al., 2023]. As SMT gave way to deep learning in machine translation (and other sequence-to-sequence applications), so have neural networks become the dominant approach to GEC. Yuan and Briscoe [2016] demonstrated the potential for this approach by applying and RNN-based encoder-decoder architecture with an attention mechanism to English GEC. More recently, Transformers [Vaswani et al., 2017] have become a common architecture for development of GEC systems, as demonstrated by Junczys-Dowmunt et al. [2018] and Grundkiewicz et al. [2019], among many others.

While framing the problem of error correction as a monolingual translation task is promising, the approach requires parallel training data [Rei et al., 2017], which if not publicly available, must be created by manually correcting text containing errors or by artificially generating errors in grammatical text. Kasewa et al. [2018] demonstrate the potential of using of artificial errors to train a GEC system; however, their method requires real-world parallel text to train their noise model used to generate artificial errors in grammatical text. Similarly, Xie et al. [2018] use a noising model trained on a "seed corpus" of parallel sentences to build a GEC system trained on artificially generated parallel noised data. Junczys-Dowmunt et al. [2018] show that effective neural GEC can be achieved with a relatively small amount of parallel training data when techniques such as transfer learning are employed. Each of these approaches were applied to error correction in English only. Grundkiewicz and Junczys-Dowmunt [2019] expands this work, demonstrating a GEC system for German and Russian that uses small corpora of corrected text to fine-tune a baseline system trained on artificial data. Most transformer models are pretrained either on a language modeling task [Junczys-Dowmunt et al., 2018] or, more commonly, on this type of synthetically generated erroneous/correct sentence pairs [Grundkiewicz and Junczys-Dowmunt, 2019, Stahlberg and Kumar, 2021].

Recent work in NMT-based GEC harnesses the predictive power of pretrained language models, such as RoBERTa and T5, which are further fine-tuned on GEC data (real, synthetic, or both). For example, Stahlberg and Kumar [2021] and Rothe et al. [2021], achieve state-of-the-art results on the JFLEG [Napoles et al., 2017] and CoNLL-2014 [Ng et al., 2014] datasets, respectively. Both models achieve these results by combining innovative methods of synthetic data generation with large pretrained transformer language models. Rothe et al. [2021] presents an effective multilingual GEC model based on the "text-to-text-transfer-transformer" (T5) [Raffel et al., 2020a] model. They further pretrain the model on a large number of synthetic erroneous-correct sentence pairs that cover four languages, before finetuning on real learner data in the target language. In this dissertation, I adopt this approach, using mT5 [Xue et al., 2021] as the base model, which I pretrain on synthetic Spanish GEC data and fine-tune on corrected learner data from the COWS-L2H corpus [Davidson et al., 2020].

3.2. Sequence tagging-based GEC

As with many NLP tasks, the current state-of-the-art in GEC involves using large masked language models such as BERT [Devlin et al., 2018]. For example, Omelianchuk et al. [2020] achieve an $F_{0.5}$ of 73.6 on the combined Write & Improve [Yannakoudakis et al., 2018] and LOCNESS [Granger, 1998] test corpus used for the BEA 2019 Shared Task on Grammatical Error Correction [Bryant et al., 2019]. Specifically, Omelianchuk et al. [2020]'s GECToR reframes the GEC task as a sequence labeling task rather than a sequence transformation task. For example, the transformation ***{go→goes}*** would instead be tagged as **$VERB_FORM_VB_VBZ**, indicating to indicate the change of a base verb form to a third person singular form. Their best performing sequence tagging model encodes input sentences using a fine-tuned version of the XLNet language model [Yang et al., 2019], then passes the encoding through additional softmax layers to assign tags to tokens. The fact that the sequence tagging approach uses an encoder only, rather than an encoder-decoder architecture used in seq2seq models, results in significantly reduced runtimes compared to alternative neural GEC approaches [Kaneko et al., 2019].

Omelianchuk et al. [2020] uses a three-stage fine-tuning protocol. Starting with a pretrained Transformer encoder, they first train the model on a large dataset of synthetically generated grammatical errors. Next, they fine-tune on actual learner data that has been annotated for errors. Finally, they fine-tune on a mixture of correct and erroneous sentences, which causes the model to more correctly predict the **$KEEP** tag [Omelianchuk et al., 2020]. Once predicted, error tags can be applied using post-processing to generate the target corrected tokens. The information in the error tags is also useful in providing feedback to learners about the types of errors they made, and in making the edits generated by the system more transparent.

While the tagging model proposed by Kaneko et al. [2019] and Omelianchuk et al. [2020] is able to achieve impressive results in GEC tasks, the approach is limited in that, even with iterative decoding that allows multiple changes to the source text, such models are ultimately dependent on token-level transformations rather than larger, stylistic changes that may help improve learner writing. Additionally, the fact that the error tags are atomic on the token level means that only one change can be made to a given token in a single run of the system. In Omelianchuk et al. [2020], the authors run the system iteratively to be able to apply multiple corrections to individual tokens. Finally, the relative complexity of the error-tagging approach, as compared to the seq2seq approach, makes this approach somewhat less attractive. First, parallel error data must be preprocessed to convert all corrections between two parallel sentences into a sequence of tags, with a tag generated for every token in the input sentence. Next, once error tags are generated, the output must be post-processed to apply the predicted changes to the input sentence. This step requires outside resources, such as a verb conjugation dictionary, to determine the correct verb form to use based on the generated tag. Given that not all languages have such resources readily available, the tagging approach may not be suitable to lower-resourced languages.

3.3. Use of Augmented Learner Data

Much previous work on adapting error correction systems to user attributes, such as L1 and proficiency level, have depended on large corpora of annotated learner data [Chollampatt et al., 2016, Nadejde and Tetreault, 2019]. However, such corpora simply do not exist for most languages, and even for English the largest such corpus, the Cambridge Learner Corpus (CLC) [Nicholls, 2003], is not publicly available. To help address the dearth of quality training data for learner GEC, much previous work has demonstrated the efficacy of synthetic training data in the development of user-adapted GEC systems. For example, Takahashi et al. [2020] show that training data created by injecting errors which simulate the error patterns seen in learner data improves a neural GEC system's ability to identify and correct errors in learner text, when compared to previous methods of generating artificial data, such as those presented in Xie et al. [2018] and Grundkiewicz and Junczys-Dowmunt [2019]. The use of artificially generated training data is even more necessary when accounting for fine-grained learner attributes, since with each attribute you wish to include, you reduce the number of examples from the target group. As with Takahashi et al. [2020], Stahlberg and Kumar [2021] use error rates extracted from learner corpora using the ERRANT system to inform their synthetic error generation method for data augmentation. They train a model that generates an "errorful" sentence given an error tag and a correct sentence. This method thus allows them to adapt their error augmentation to the error rates extracted from real learner data; their results show a significant performance improvements for English GEC. Here, I expand the work of Stahlberg and Kumar [2021] to Spanish, a language that has much richer morphology than English, and which I hypothesize will require more fine-grained error categories to effectively use error rates to inform data augmentation. Additionally, my dissertation explores generating artificial training data that replicates the errors made by students of various linguistic backgrounds and proficiency levels, thereby creating a GEC model that is tuned to these student attributes. Although I use L1 and proficiency level to categorize students, the approach could be expanded to account for any identifiable learner attribute which is shared by a reasonably-sized cohort of students.

3.4. Application of Generative LLMs to GEC

Recently, pretrained large language models (LLMs) such as BERT, XLNet, and T5 have heavily influenced nearly all areas of NLP, vastly improving the fluency of generated language and demonstrating advanced, seemingly human-like, capabilities [Minaee et al., 2024]. These models have also had a profound impact in GEC tasks, being used in all currently leading models on popular benchmarks like the BEA-2019 [Bryant et al., 2019] and CoNLL-2014 Ng et al. [2014] evaluation sets. However, unlike many areas of NLP, the state-of-the-art in GEC has been less influenced by the release of large generative language models like Claude2[3] and GPT-4[4]. While

[3] https://www.anthropic.com/news/claude-2
[4] https://openai.com/index/gpt-4/

multiple papers have explored the possibility of using in-context learning with pretrained generative LLMs to correct learner text, the reported performance varies widely across test datasets and prompts. Specifically, large generative LLMs are "known for over-correction where results obtain higher recall measures than precision measures" [Zeng et al., 2024]. That is, generative LLMs tend to make more edits than human annotators. For example, Coyne et al. [2023] report that the performance of GPT-3.5 and GPT-4 on GEC tasks largely depends on the types of edits seen in the evaluation dataset. When evaluating on the JFLEG dataset [Napoles et al., 2017], Coyne et al. [2023] achieve state-of-the-art GLEU scores using GPT models (GLEU is the standard evaluation metric used for evaluation of the JFLEG benchmark). However, when evaluating on the BEA-2019 dataset [Bryant et al., 2019] the authors report $F_{0.5}$ scores that are more than 20 points below those achieved using a modified version of the GECTOR model [Yasunaga et al., 2021]. In evaluating these results, the authors state their belief that this marked difference in performance relates to the types of edits that occur in the two datasets. In JFLEG, edits are intended to improve fluency, not just make the sentences strictly grammatical: the corpus "uses holistic fluency edits to not only correct grammatical errors but also make the original text more native sounding" [Napoles et al., 2017]. The BEA-2019 dataset, on the other hand, includes primarily minimal edits designed only to make ungrammatical sentences grammatical [Coyne et al., 2023]. According to Coyne et al. [2023], GPT-3.5 and GPT-4 are adept at rewriting sentences to make them sound more fluent (and grammatical), but they tend to over-correct when the goal is minimally editing source sentences for grammatical errors only. Similarly, Fang et al. [2023] report that ChatGPT is adept at making sentences sound more fluent, but that the model fails to adhere to the principle of minimal edits, even when this principle is clearly explained in the prompt. I found the same trend when analyzing output from GPT-4 when I used it for Spanish GEC with data from the COWS-L2H corpus. Thus, while GPT-3.5 and GPT-4 can be quite effective in sentence editing tasks, care must be taken when minimal edits are desired. For this reason, the model I primarily use for experiments in this dissertation is a fine-tuned version of the mT5 model.

While pretrained generative LLMs struggle to abide by the principle of minimal edits required by many GEC tasks, other work has shown them useful in related applications. For example, Kobayashi et al. [2024a] demonstrates that generative LLMs can be effectively used as evaluators of output from other GEC models. Specifically, they use GPT-4 to assess the quality of corrections from the SEEDA dataset [Kobayashi et al., 2024b] on several dimensions; for the edit-level: edit difficulty and edit impact, and for the sentence level: grammaticality, fluency, and meaning preservation [Kobayashi et al., 2024a]. They report that in all dimensions, GPT-4's judgments show higher correlation with human judgments than alternate metrics such as the ERRANT $F_{0.5}$ score [Bryant, 2019]. The correlation is strongest when the model is asked to judge the overall fluency and grammaticality of the edited sentence. This finding aligns with other work related to the use of generative LLMs in the GEC task; that is, that these models are particularly proficient at fluency-related tasks. Overall, Kobayashi et al. [2024a] argue that larger-scale generative LLMs, like GPT-4, may be a useful alternative to the rule-based metrics currently used to evaluate GEC systems.

One additional application that generative LLMs have found in the realm of GEC and AWCF is their utility in generating explanations for edits made by other GEC models. From a pedagogical standpoint, simply providing proposed error corrections to students with no explanation or context is sub-optimal [Ferris, 2012, Barrot, 2021, Ellis et al., 2006], as discussed in detail in Chapter 2. Thus, to effectively integrate GEC models into an AWCF system, such as the one proposed in this dissertation, feedback must be generated based on the edits made by the GEC model. For example, if a GEC model suggests the insertion of a missing preposition, the AWCF system must provide feedback explaining why the preposition is needed in that position. Until recently, most AWCF systems used a complex set of feedback templates to provide targeted implicit feedback to learners (see, for example, Liang et al. [2023]). This approach, while effective, is brittle and limited in scope, given that only so many error types can feasibly be considered. Additionally, writing a large number of templates is time consuming for developers, leading to economic constraints that may effect the availability of effective AWCF systems for lower-resourced languages. Recent work, however, has applied the generative power of LLMs to create fluent, grounded explanations of error corrections that can be used to provide implicit feedback to writers. Initial steps in this direction have been taken by two recent works - Song et al. [2023] and Kaneko and Okazaki [2023] - which use GPT-4 and GPT-3.5, respectively, to write explanations of error corrections that can then be presented to learners. Song et al. [2023] shows that GPT-4 is not effective at identifying and explaining errors directly from parallel original/corrected sentence pairs; when presented with a dataset consisting of such sentence pairs, GPT-4 was only able to identify only 60.2% of errors, and was then only able to effectively explain 67.5% of those identified [Song et al., 2023]. Thus, the authors propose a two-step prompting strategy in which the GPT-4 is first asked to identify all atomic edits between parallel sentences; these atomic edits are then appended to the original sentence pair as input to a second prompt that asks GPT-4 to generate explanations for why each edit is needed. Human evaluation shows that this approach is quite effective at generating correct explanations; 93.9% and 98% of error explanations were judged as correct in German and Chinese, respectively. In this dissertation, I expand the approach described in Song et al. [2023] to Spanish by combining a template-based feedback approach with LLM-generated explanations to improve feedback diversity and to expand the number of error types that can be effectively covered by the proposed AWCF system, while preserving the control afforded by template-based feedback.

CHAPTER 4

COWS-L2H Corpus

While annotated learner corpora of English are widely available, large learner corpora of Spanish are less common, and as a result, the field has seen little data-driven research on the developmental processes that underlie Spanish language learning, or on the development of computational tools to assist teachers and students of Spanish. This may come as unexpected, considering the fact that there exists a relatively high demand for learning Spanish; in 2013, fifty-one percent of students enrolled in university language courses in the United States studied Spanish [AAAS, 2016] and there are over 21 million learners of L2 Spanish across the globe [Cervantes, 2019]. This paucity of non-English data is especially evident with respect to error annotated and corrected data in Spanish for use in error analysis or training of automated feedback systems.

4.1. Current Spanish Learner Corpora

Text written by fluent Spanish speakers is widely available (e.g. Wikipedia and other corpora, such as the Corpus del Español [Davies, 2002] and the various corpora organized by the Real Academia Española). Additionally, several corpora of transcribed spoken Spanish produced by L2 learners, such as CORELE [Campillos Llanos, 2014], the Corpus Oral de Español como Lengua Extranjera, and SPLLOC [Mitchell et al., 2008], the Spanish Learner Language Oral Corpus, are available to the research community. However, few corpora of written learner Spanish are available for use by researchers, and those corpora of written learner Spanish that have been compiled do not include parallel corrected text or annotations to facilitate analysis of error patterns or the training automated correction systems. For example, CAES, the Corpus de Aprendices de Español [Rojo and Palacios], is searchable via a concordancer, but it is not error-annotated and raw text is not easily downloadable. Other potentially promising corpora of L2 Spanish for researchers include Aprescrilov [Buyse and González Melón, 2012] and the "Corpus Escrito del Español como L2", or CEDEL2 [Lozano et al., 2009], which contain approximately 1 million and 750,000 tokens, respectively. In addition, CEDEL2 contains a subset of essays written by native Spanish-dominant participants for comparative purposes. Where appropriate, I use Corpus del Español and CEDEL2 for analysis of native writing.

4.2. COWS-L2H

Due to the paucity of available data in learner Spanish, researchers in the Spanish and Linguistics departments at the University of California, Davis developed the Corpus of Written Spanish of L2 and Heritage Learners (COWS-L2H) [Yamada et al., 2020, Davidson et al., 2020]. This corpus of over 1,367,000 words is composed of 5,383 personal essays written by 1,934 unique participants enrolled in various levels of undergraduate Spanish instruction at UC Davis. The corpus contains 4,804 essays from L2 (and L3) learners of Spanish in various instruction levels ranging from beginner (SPA 1, 2, 3) to upper division, as well as 579 essays written by Heritage learners at three levels of instruction (SPA 31, 32, and 33). The distribution of the essays across levels, as detailed in Table 4.1 is uneven due to the distribution of students enrolled in Spanish courses; because more students enroll in beginning Spanish courses than in advanced levels, a larger number of essays submitted to the corpus come from these beginner-level courses. During each academic quarter (ten weeks of instruction), participants are asked to write two essays in Spanish that adhere to a minimum of 250 and a maximum of 500 words, though students enrolled in Spanish 1 are allowed to write essays with a lower minimum word count, due to the fact that many of these students are true beginners in L2 Spanish who possess relatively little vocabulary and grammatical resources of their own. Participants are asked to write each essay they submit in response to one of two short prompts. Participants of all levels followed the same two prompts during the same academic quarter, to allow lexical and syntactic comparisons across levels which are not influenced by topic variation in the writing samples. Both prompts are presented with a distinct brevity, to allow for a broad degree of creative liberty and open-ended interpretation on the part of the writer. To test the effect of prompt on student writing and promote diversity in our corpus, we periodically change the prompts presented to students. To date we have presented eight essay prompts. For the first set of compositions, collected from 2017 to 2018, participants were asked to write about "a famous person" and "the perfect vacation." For essays collected from 2018 to 2019, the prompts were "a special person in your life" and "a terrible story". For more recent compositions, collected from early 2020 late early 2021 ask students to write "a description of yourself" and "a beautiful story".

Course Level	**Essays**	**Tokens**
Beginner (SPA 1-3)	3,018	711,630
Intermediate (SPA 21-22)	562	150,279
Composition (SPA 23-24)	883	248,746
Heritage (SPA 31-33)	579	164,723
Upper Division	222	62,404
Unknown level	119	29,476
Total	5,383	1,367,258

Table 4.1. Summary of corpus composition by instruction level.

Finally, our current collection effort, begin in early 2021, asks students to describe "a place you dislike" or to write a description of a Charlie Chaplin movie clip we ask students to watch. We have collected an average of 900 essays in response to each of the prompts we have used to date (with the exception of the those most recently collected).

Given the diverse backgrounds of our students, identifying the specific variety of Spanish in the essays is challenging. This is especially true for those students who enroll in courses for Heritage speakers. However, our courses are generally taught using a standard variety of academic Spanish, so we expect this to be the predominant variety in the corpus. Students provide information about their linguistic background that is included as metadata in the corpus; this metadata may elucidate variability in usage resulting from students' past experience with Spanish. The metadata also allows us to test the effects of variables such as L1 on student writing. Finally, the linguistic metadata facilitates the use of subcorpora, filtered by specific student attributes, to conduct targeted analysis of potential errors and language development. Additionally, these subcorpora may be useful in the training of NLP systems; for example, as mentioned previously, Nadejde and Tetreault [2019] demonstrate that grammatical error correction systems benefit from adaptation to L1 and proficiency level. Similar improvements to GEC performance based on adapting GEC systems to L1 and proficiency level are reported by Rothe et al. [2021] and Zeng et al. [2024].

4.3. Longitudinal data

One of most important features of the COWS-L2H corpus is the inclusion of large amounts of longitudinal data submitted by students who participated in the project for multiple quarters, allowing researchers to study the writing development of individuals or cohorts of students as they progress through a university Spanish language program. While the LANGSNAP corpus

Number of Texts	Num. of participants
1 text	298
2 texts	998
3 texts	99
4 texts	308
5 texts	47
6 texts	112
7 texts	14
8 texts	34
9+ texts	25
Total	1,935
Total 3+	639
Total 6+	185

Table 4.2. Summary of COWS-L2H Longitudinal Participation

[Tracy-Ventura et al., 2016] contains longitudinal written data gathered from L2 learners of Spanish, the scale of the longitudinal data in COWS-L2H and the annotations available in the corpus make it a unique resource to researchers. Overall, a total of 250 students participated for at least 3 quarters, and 30 students participated for 5 or more quarters. In terms of number of essays submitted, 639 participants submitted 3 texts or more, while 185 wrote 6 texts or more. These numbers are summarized in Table 4.2 Due to the nature of our prompt-based data collection method, many students who participated over multiple quarters responded to the same prompt several times at different points in time, making the resulting essays readily comparable.

In an effort to make this longitudinal data more valuable to language researchers, all essays submitted by students who participated for more than 3 quarters (consisting of 1,628 essays) have been corrected and annotated for selected errors by graduate-level Spanish instructors, as described in more detail below. Thus the COWS-L2H corpus represents the first dataset of error-corrected and annotated longitudinal text written by L2 and Heritage learners of Spanish available to the research community.

4.4. Error Annotation

In an effort to facilitate research in language development, the designers of the corpus and their research partners have annotated a large portion of the corpus for several specific types of grammatical error of interest to project collaborators. Out of the 5,383 texts in COWS-L2H, 2,948 have been both holistically corrected (providing parallel text for machine learning training) and error annotated for the following types of errors:

- Incorrect gender and number agreement in nouns, adjectives and articles
- Undue presence or absence of personal pronouns or articles
- Incorrect use of the prepositions por and para or the verbs ser and estar
- Incorrect word order in noun-adjective pairs
- Verb form errors related to:
  - Mood, tense or aspect selection
  - Person and number selection
  - Regularization of irregular verbs

As previously mentioned, we have annotated all longitudinal data of students who participated in more then 3 quarters, as well as a portion of the cross-sectional data, for a variety of error types of interest to project collaborators. Additionally, we have designed a detailed error annotation schema which can be readily adapted to other error types of interest to researchers using the corpus. The instructions provided to annotators, which describe each annotation target in detail, are available in the corpus GitHub repository. We encourage researchers interested in adding additional error annotations to the corpus to use the proposed annotation schema, available in the corpus repository, to facilitate inclusion in the corpus.

Annotators are graduate student instructors of Spanish, either at the university where the essays were collected (UC Davis) or at a partner university in Spain (Universidad de Salamanca). To ensure reliability of the annotations, we conducted multiple rounds of training with annotators, allowing them to discuss areas of disagreement. Once we achieved an acceptable Cohen's $\varkappa$, we proceeded to annotation of the remaining target data. Overall, our annotators achieved a Cohen's $\varkappa$ of 0.62, indicating reasonable inter-annotator agreement.

4.5. Correction & Automated Error Tagging

Manual error annotation, while accurate and reproducible, is an extremely time consuming process. We therefore, supplement our error annotation with parallel corrected text, from which we are able to extract a much larger set of errors based on instructor corrections of student essays. When creating the corrected versions, annotators followed the principle of minimal edits. They were

Course Level	Essays	Sentences	Tokens
Beginner	1,566	35,980	432,290
Intermediate	356	7,537	110,490
Composition	438	8,598	141,353
Heritage	337	5,487	110,313
Other	225	4,112	68,393
Total	2,922	61,714	862,839

Table 4.3. Summary of corrected essays in the COWS-L2H corpus. Note that these numbers do not count both of the double-corrected essays, hence the discrepancy between the 70,397 sentence count above and the numbers in this table.

instructed to correct spelling, grammatical and lexical errors while refraining from making significant stylistic changes to the original texts, particularly avoiding changes that alter the meaning of a sentence.

The present version contains 2,922 corrected essays for 70,397 corrected sentence pairs. Our corpus represents the largest parallel dataset of corrected Spanish text available to researchers. The distribution of corrected essays by student level is shown in Table 7.2. 572 essays have been corrected by two annotators, providing two references for more flexible evaluation of GEC systems.

The parallel nature of the corrected essays allows errors to be automatically tagged using tools such as ERRANT [Bryant et al., 2017], which I have modified and demonstrated to work effectively with Spanish in Davidson et al. [2020]. As such, the parallel data can be used to study errors that have not been specifically annotated, providing far more flexibility when conducting error analysis. To measure annotator agreement, we compare the corrections made by two different annotators using ERRANT, achieving an $F_{0.5}$ of 0.53, with precision of 0.54 and recall of 0.51.

4.6. Additional corpora used

To examine the lexical richness and formulaic language use of native Spanish speakers residing in Spanish-majority countries, I use a 2 million token sample of the Web/Dialects portion of Corpus del Español [Davies, 2002]. This freely-available sample contains Spanish texts collected from online articles published in 21 different hispanophone countries. In addition to raw text, the Corpus del Español has been annotated with part-of-speech and lemma tags, making analysis of lexical diversity and lexical density far easier and more reliable.

Finally, to further compare the learners in the COWS-L2H corpus to non-Heritage native speakers of Spanish, I utilize a portion of the CEDEL2 corpus [Lozano et al., 2009] which contains a prompt identical to one used in COWS-L2H (the COWS-L2H prompt was modeled on the CEDEL2 prompt). However, this subcorpus is small, thereby limiting the inferences which can be drawn from this analysis.

4.7. Studies conducted using COWS-L2H

In the following three chapters, I present corpus-driven studies conducted primarily using data from COW-L2H to identify differences between L2, L3 and Heritage learners of Spanish along three dimensions:

- Lexical development, including the acquisition and use of formulaic language.
- Syntactic and morphological development in cross-sectional and longitudinal corpus data.
- Distribution of errors in parallel corrected data. It is important to note that "errors" are defined by instructor corrections, whether or not the corrections represent true errors in production or simply non-standard usage.

Although these studies may seem disparate, they can also in many ways be viewed as a unified whole, both informing and reinforcing their respective findings. Additionally, all three are useful as a means of investigating optimal methods of generating synthetic "learner-like" data for the training of student-adaptable corrective feedback systems. While recent work [Li et al., 2015, Stevenson and Phakiti, 2014] has shown that automated corrective feedback shows promise as a tool for improving student writing competency, to date it has largely been a blunt instrument that does not consider student background or instructor pedagogical goals. I argue that an automated feedback system which is better able to adapt to both student linguistic experience (such as that presented in Nadejde and Tetreault [2019]) and the instructor's pedagogical goals will prove beneficial to students and result in better learning outcomes. The results presented in the following chapters lend further understanding to how the writing of students of different linguistic backgrounds are likely to differ, and are an important step toward building a feedback system which can be adapted to the learning needs of a diverse student population.

CHAPTER 5

COWS-L2H Study I: Lexical development

In the this chapter, I present findings comparing the lexical development of Heritage learners of Spanish to that of their advanced L2 learner peers.

5.1. Measuring lexical development

One known deficiency in Heritage Learners is vocabulary [Montrul, 2012]; however, other related deficiencies, such as Heritage learners' acquisition and use of formulaic language, remain largely unexplored. Recent research [Ellis et al., 2015] has demonstrated that formulaic language is stored as part of a speaker's mental lexicon, so it is plausible that a deficiency in vocabulary would also manifest as a deficiency in formulaic language use. Much research in language learner lexical development is based on the premise that "vocabulary use directly taps into learners' lexical knowledge and is closely aligned with language proficiency and levels of vocabulary acquisition" [Crossley et al., 2017]. However, this view does not take into account the role of holistically stored, multi-word units which play a significant role in language acquisition, and help shape the lexicon of language learners [Paquot and Granger, 2012].

Previous research comparing L2 and Heritage learners has found that, when compared to advanced L2 learners, Heritage learners tend to be stronger in verbal communication but weaker in writing [Montrul, 2012]. This fact is not surprising, given that for many Heritage students, Heritage Spanish courses are their first formal education in Spanish. Interestingly, though, Montrul [2012] also reports that Heritage learners tend to use more low-frequency words in their writing than their L2 learner peers, as they often write using the same vocabulary they would use in speaking, thus demonstrating lexical patterns more similar to those of Spanish-dominant individuals.

In analyzing lexical development among language learners, researchers often employ three simple metrics: lexical diversity, lexical density, and lexical sophistication [Gregori-Signes and Clavel-Arroitia, 2015]. These three metrics are designed to indicate the richness and variety of a student's vocabulary and how well she uses this vocabulary to convey information.

5.1.1. Lexical diversity. Lexical diversity is a relatively straightforward measure of how many unique words a student uses in a text relative to the overall number of words [Johansson, 2008]. The motivation for using such measures is the assumption that a person who uses more diverse vocabulary in his or her writing possesses greater lexical knowledge. In addition, indices of lexical diversity are correlated with issues as diverse as Alzheimer's onset and socioeconomic status [McCarthy and Jarvis, 2010]. The present study explores how lexical diversity can be used to measure differences in vocabulary knowledge between Heritage learners and their L2 learner peers.

The most basic measure of lexical diversity, which was widely used in the past (see, for example, Bates et al. [1991]), is the "type-to-token ratio" (TTR) [Templin, 1957] which simply counts the number of unique words (types) in a text and divides it by the total number of words (tokens). However, TTR shows significant text-length effects; as texts grow longer, their TTRs tend to decrease, thereby greatly reducing the value of TTR as a metric of lexical diversity. The reason for this trend is simple: people, in general, have a limited working vocabulary, and therefore tend to repeat words throughout a given text. So, while the length of a text can grow without bound, the number of unique words used tends to grow at an increasingly slower rate as the text gets longer. In addition, function words, such as articles, as well as common verbs and nouns, are repeated many times in a longer text; while these words continue to contribute to the token count, they stop contributing to the type count after their first appearance. Thus, longer texts generally have lower TTR scores than do shorter texts; this makes comparison of texts using TTR difficult, especially in developmental studies where students are likely to write longer texts as their proficiency increases [Johansson, 2008]. Therefore, Johansson argues, TTR in its basic form is only useful for comparing texts of similar length.

Several modifications of TTR have been proposed to mitigate the text-length effect on TTR scores, allowing researchers and instructors to compare students' lexical diversity in a more meaningful and robust way. One such measure, which I will use for my analysis in this dissertation, is the Measure of Textual Lexical Diversity (MTLD) [McCarthy and Jarvis, 2010], which seeks to normalize TTR for text length, allowing comparison of texts with differing lengths. I will further describe the implementation of MTLD later in this dissertation. Other proposed measures of lexical diversity include VocD [Malvern and Richards, 1997], which compares the predicted decrease in lexical diversity with increasing text length with the observed decrease by randomly sampling portions of text, and theoretical vocabulary [Broeder et al., 1986], which calculates TTR for a fixed length sample from texts of varying length, thereby approximating comparable TTR values.

5.1.2. Lexical density. Another commonly used metric of lexical development is lexical density [Ure, 1971]. In its simplest form, lexical density is a ratio of the open-class words (nouns, verbs, adjectives, adverbs) to total words in a given text. By comparing the number of words carrying significant semantic content with the number of function words in a text, we are able to conceptualize the "information content" of a text; writing with a higher ratio of content words is more informationally dense than a text with relatively more function words (such as pronouns, articles, and prepositions) [Johansson, 2008]. Halliday and Halliday [1989] further refined Ure's concept of lexical density in the context of child language development. Specifically, Halliday discusses the fact that, in children, lexical density decreases as the child's fluency in the language develops. While this may seem counter-intuitive, it is due to the fact that young children begin speaking in short utterances that consist of only content words. As the children begin to speak in longer phrases and, eventually, in sentences, they begin to use more function words, thereby reducing the content word to function word ratio of their language [Halliday and Halliday, 1989]. Interestingly, I find a similar pattern among L2, and to a lesser extent, Heritage learners in this study.

Laufer and Nation [1995] argue that lexical density is not a reliable measure of lexical development, stating that "fewer function words in a composition may reflect more subordinate clauses, participial phrases and ellipsis, all of which are not lexical but structural characteristics of a composition." While this argument may be true, we must consider the fact that corpus studies, such as the present study, analyze aggregate data. An individual text may show high lexical density for the reasons mentioned by Laufer and Nation, but it is unlikely that, as a group, students are going to write in a way that confounds the demonstrated correlation between lexical density and overall lexical development.

In order to improve upon the reliability of lexical density measures, researchers have proposed various modifications of the metric. For example, Wolfe-Quintero et al. [1998] discuss subdividing lexical density by function; for example, they propose noun density as a possible alternative to the more general lexical density. However, none of these variations in the lexical density metric seem to have been widely adopted by the research community. For this study, I will use the basic measure of lexical density as proposed by Ure [1971] in which each orthographic word is treated as an individual item (Halliday, by comparison, tends to treat fixed items like "turn down" as a single lexical item, rather than as one lexical and one function item), and in which lexical (content) words are defined as all nouns, verbs, adjectives, and adverbs.

5.1.3. Lexical sophistication. Lexical sophistication is simply a measure of the proportion of "advanced" words a learner uses in his or her output. Many researchers have defined "advanced" words to mean "infrequent" words [Laufer and Nation, 1995], though others have developed specific "advanced" word lists. For example, the Academic Word List developed by Coxhead [2000] is used in some measures of lexical sophistication [Kyle and Crossley, 2015]. The most common measure of lexical sophistication uses a word frequency list, generally extracted from a large corpus, to assess the sophistication of the words in a text. Using this approach, researchers select some word frequency list (or extract their own from a corpus) and set some frequency cut-off after which words are considered "advanced." More fine-grained metrics, such as the Lexical Frequency Profile [Laufer and Nation, 1995], subdivide more frequent words into bands of frequency - so a word in the top 1000 most-frequent words would be considered less "advanced" than a word in the 1001-2000 most-frequent words.

One important caveat to using lexical sophistication as a measure of lexical development is the fact the definition of "advanced" is at the researcher's discretion, and the researcher must choose an appropriate frequency list to make the results meaningful [Laufer and Nation, 1995]. For example, a list of words considered advanced for a first-year L2 learner may not be appropriate when assessing the lexical sophistication of advanced Heritage learners. A related issue is the effect of genre; the selected word list must be genre-appropriate, or words that are actually quite common in a genre-specific setting may be inappropriately counted as "advanced." As Bell [2003] points out: "problems remain when this form of assessment is applied to the evaluation of the freely-produced language of learners. One reason for this is the difficulty of producing a frequency list which is appropriate to the situation of the learners."

I note that little work has been done in assessing lexical sophistication in Spanish learners. In fact, I found only one study [Waldvogel, 2014] that specifically investigates the lexical sophistication of L2 Spanish learners. I found no studies that specifically analysed the lexical sophistication of

Heritage learners of Spanish. Developing and testing a tool to analyze the lexical sophistication of Spanish L2 and Heritage learners is beyond the scope of this study.

5.1.4. Formulaic language use. According to Bybee [2008], grammar and lexicon are the "cognitive organization of one's experience with language." This emergentist view of language, which states that humans have no innate language capacity beyond our general cognitive abilities, allows for both L1 and L2 to emerge from experience with language alone [Tomasello and Rakoczy, 2003]. As an extension of this theoretical viewpoint, Hall [2017] argues that when language learners repeatedly experience multi-word formulaic language with fixed meaning, they are able to store these strings holistically in their mental language representation in a manner similar to words. Wray [2009] defines formulaic language as follows: "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar." The use of formulaic language consists of using "semi-preconstructed" phrases which are actually single choices by the speaker, though they appear to be analyzable into component parts [Sinclair and Sinclair, 1991]. Formulaic language consists of recurrent multi-word items such as idioms, collocations, and lexical bundles [Conklin and Schmitt, 2012].

The amount of language, written and spoken, which consists of formulaic language is up for debate. Conklin and Schmitt [2012], citing numerous studies, state the following about the prominence of formulaic language:

> Oppenheim (2000) counted the multiword stretches of talk that occurred identically in practice and final renderings of a short speech on the same topic and found between 48 percent and 80 percent (overall mean of 66 percent) consisted of identical strings. Sorhus [1977] calculated that speakers in her corpus of spontaneous English Canadian speech used an item of formulaic language once every five words. Erman and Warren [2000] calculated that 52–58 percent of the language they analyzed was formulaic, and Foster [2013] came up with a figure of 32 percent using different procedures and criteria. Biber et al. [1999] found that around 30 percent of the words in their conversation corpus consisted of lexical bundles, and about 21 percent of their academic prose corpus. Howarth [1998] looked at frequent verbs in a social science/academic corpus and found that they occurred

in either restricted collocations or in idioms in 31–40 percent of the cases. Rayson [2008] found that 15 percent of text is formulaic according to a Wmatrix analysis.

Based on these studies, Conklin and Schmitt draws the conclusion that formulaic language makes up somewhere between one-third and one-half of all discourse. As a result of its prominence in discourse, the ability to correctly use formulaic language is an important step in acquiring native-like competence in any language; the use of collocations, idioms, and other types of formulaic language directly impacts the three dimensions of language proficiency — complexity, accuracy, and fluency [Housen and Kuiken, 2009].

As with lexical development, formulaic language is often studied using large corpora, as identifying words which co-occur with a high frequency is a straightforward procedure given modern corpus methods.

5.1.4.1. *Idioms.* Idioms are a type of formulaic language in which the semantics of the words are non-compositional; that is, the meaning of the phrase cannot be derived from the individual meanings of the words. For example, "kick the bucket" does not literally mean to kick a bucket; the actual meaning "to die" has no clear relationship to the component words. Due to their non-compositional semantic nature, which results in an arbitrary meaning-phrase relationship, idioms are stored in the minds of native speakers in a manner similar to individual words [Tremblay et al., 2011]. Language learners must also learn and store the non-compositional meanings of idioms if they want to achieve native-like language fluency [Ellis et al., 2008].

5.1.4.2. *Collocations.* Collocations are words that occur together at a rate higher than would be expected by chance. In addition, lexical choice restrictions apply with collocations; speakers are not free to replace a member of the collocate with a semantically similar word to get a similar meaning. For example, while one can say "heavy rain," it would sound non-native to say "weighty rain" or "strong rain," even though these have the same semantic meaning in this context. Paquot and Granger [2012] define collocations as "usage-determined lexical combinations that are characterized by restricted co-occurrence of elements." Common types of collocations include verb + noun ("commit suicide" and "answer a question"), adjective + noun (as in "fast food") and verb + adverb (as in "run rapidly"), though any combination may exist that is grammatical in the language in question. Clearly, native speakers are either storing collocation frequencies in some manner or are storing common collocations as fixed lexical units. For example, native English speakers know to say "made a mistake" rather than "did a mistake;" while the latter is perfectly grammatical, any native speaker of English would recognize this immediately as a non-native error.

5.1.4.3. *Lexical bundles.* Lexical bundles are "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" [Biber et al., 1999]. Paquot and Granger [2012] clarify this definition, stating that lexical bundles are "contiguous strings of words of a given length (e.g., bigrams, trigrams)" which occur at a much higher than chance probability in language. The key difference between lexical bundles and collocations is that lexical bundles do not involve restricted lexical choice; rather lexical bundles are simply emergent patterns of fixed-length strings of words that are often repeated. While any fixed-length string can be viewed as a lexical bundle if it is frequently used, lexical bundles often serve as referential markers, text organizers, stance markers, or interactional discourse markers [Biber et al., 2003]. Unlike idioms, lexical bundles are semantically compositional; the meaning of lexical bundles can be derived from the meaning of their component words. And unlike collocations, components of lexical bundles can often be replaced with semantically related words without changing the meaning or creating a non-native-like form. For example, the four-word lexical bundle "it should be noted" can be similarly expressed as "it must be noted," "it should be said," and "it must be said," among the many possible variants. Lexical bundles are quite easy to identify given corpus methods; one need simply look at all n-grams in a corpus and count the number of times each one occurs. Given this information, and the size of the corpus, one can readily calculate the frequency of each n-gram in the corpus. While one can analyze lexical bundles of any length, researchers generally focus of bundles ranging from two to six words [Paquot and Granger, 2012].

5.1.4.4. *Frequency.* When considering the prevalence of a formulaic string, be it an idiom, collocation, or lexical bundle, one must first consider the frequency with which the form occurs. Generally, this is expressed as a normalized frequency, such as occurrences per million words [Bestgen, 2020]. However, because words and word sequences are distributed in a Zipfian manner (the distribution is not normal; some words, such as function words, occur very frequently, while most other words and word sequences occur infrequently), normalizing frequencies in this manner is potentially problematic. Studies have shown that, due to the irregular distribution of words and n-grams, smaller corpora tend to show significantly more high-frequency lexical bundles than do larger corpora [Bestgen, 2020]. While empirical evidence to support this claim is scarce, one must still be careful when comparing corpora of different sizes. In an attempt to control for this effect, I use sub-corpora of approximately equal size when extracting lexical bundles and collocates in this study.

5.1.4.5. *Pointwise mutual information.* Computational linguists often measure the association between words using a measure termed pointwise mutual information (PMI). This measure, derived from information theory, is useful in determining how likely two or more words are to co-occur in a given corpus [Church and Hanks, 1990]. Given two words, the mutual information is a measure of how much more (or less) likely the two words are to occur together than would be expected by chance alone. This is calculated using the following formula, which calculates bigram mutual information:

$$PMI(x,y) = log\frac{P(x,y)}{P(x) \times P(y)}$$

When we are seeking to find the mutual information of a trigram or larger n-gram, we must expand the above formula using the chain rule [Su et al., 1994]. Thus, we calculate trigram mutual information as follows:

$$PMI(x,y) = log\frac{P(x,y,z)}{P(x) \times P(y) \times P(z) + P(x) \times P(y,z) + P(x,y) \times P(z)}$$

PMI tells us how much the presence of one word predicts the presence of another [Church and Hanks, 1990]. Collocations, as defined above, which have lexical choice restrictions, thus tend to have the highest PMI, since the presence of the root word is highly predictive of the collocate (e.g., "think about", in which the presence of "think" is highly predictive of "about"). Lexical bundles which frequently occur will also tend to have higher-than-chance PMI, though bundles which consist of purely grammatical words tend to have lower PMI, as the frequency of the component words are often quite high independent of the bundle. For example, the construction "if it is" is a common lexical bundle, but its PMI would tend to be low as the component words also frequently occur in other contexts. According to Ellis et al. [2008] phrases with higher PMI are more likely to be rated as "formulaic" by human annotators. In this study, I will use PMI to help identify differences in usage patterns between native L1, L2, and Heritage learners of Spanish to determine if, as predicted, the language use of Heritage learners shows similar patterns to that of L2 learners.

5.2. Previous work

As mentioned previously, Guadalupe Valdés defines a Heritage Learner as "language student who is raised in a home where a non-English language is spoken, who speaks or at least understands the language, and who is to some degree bilingual in that language and in English" [Polinsky and Kagan, 2007, Valdés, 2000]. While this definition is useful, it omits an important factor in the Heritage learner equation - the fact that, at least in the United States, most children from non-English speaking homes have been forced, until quite recently, to attend English-medium schools[5]. Thus, their Heritage language was not supported in the educational environment, resulting in significant language attrition for many of these students. O'Grady et al. [2011] acknowledge the issues that public education policy causes for Heritage learners: "in the typical case, Heritage learners receive ample exposure to the home language early in life, only to have that exposure end or undergo a dramatic reduction within a matter of years—often at the point where formal schooling begins" [O'Grady et al., 2011].

O'Grady et al. [2011] further argue that many Heritage language learners suffer specific deficiencies in their language proficiency as a result of the limited exposure that these students have to their Heritage tongue, and point to numerous previous studies which detail specific deficiencies.

> It is by now well established that children in this situation manifest linguistic deficiencies in the heritage language in a wide range of areas, including vocabulary [Polinsky and Kagan, 2007], gender classification [Polinsky, 2008a], tense, aspect, and modality [Montrul, 2002, Lynch, 2003, Silva-Corvalán, 2006, Polinsky, 2008b], case paradigms [Polinsky and Kagan, 2007], the interaction between case and word order [Song et al., 1997], relative clauses [O'Grady et al., 2001, Kim, 2005], agreement [Bolonyai, 2007](for Hungarian), reflexive pronouns [Song et al., 1997, Kim et al., 2009], differential object marking [Montrul, 2004, Montrul and Bowles, 2009], the semantics of unaccusativity [Montrul, 2005], the contrast between overt and null subjects in "pro drop" languages [Montrul, 2004], quantifier placement

[5]In California, bilingual education was largely curtailed by Proposition 228 in 1998, and reinstated by Proposition 58 which was passed by voters in November, 2016.

[Polinsky and Kagan, 2007], and filler-gap dependencies [Polinsky and Kagan, 2007]. [O'Grady et al., 2011]

As stated previously, this section will focus on the development of lexicon and formulaic language in Heritage learners of Spanish, and how their development compares to that of L2 students and native-born Spanish speakers. Both Lynch [2008] and Montrul [2011] argue that Heritage language speakers show deficiencies in their vocabulary development due to the limited context and register of the Heritage language to which they are exposed in the home environment. According to Montrul, "Heritage language speakers know many words in their heritage language, but most often these are words related to common objects used in the home and childhood vocabulary. In fact, heritage language speakers also have significant gaps in their vocabulary and find it difficult to retrieve words they do not use very frequently" Montrul [2011]. When compared to L2 learners, Montrul [2012] finds that there is no overall speed or accuracy advantage for Heritage learners in lexical decision tasks [Montrul, 2012], indicating that the lexical development of Heritage learners may be more similar to L2 learners than to that of non-Heritage native speakers. In addition, limited social use of minority languages results in simplified grammatical systems that introduce "innovative, that is, non-normative, elements at the lexical and discourse levels," and that these innovative patterns are conditioned by the dominant language [Lynch, 2008]. Although little work has focused on the use of formulaic language by Heritage learners, recent research has demonstrated that all speakers are likely to store formulaic language in their mental lexicon alongside individual words [Ellis et al., 2015]. If the lexicon and grammar of Heritage speakers is being negatively impacted by the limited domain of exposure to the Heritage language, it stands to reason that they will also demonstrate similar gaps in knowledge of native-like formulaic language.

Analyzing the lexical development of language learners is an important aspect of assessing overall language proficiency. Beyond the fact that lexical proficiency is a key part of L2 learning [David, 2008], Kyle and Crossley [2015] have demonstrated that, at least for L2 learners of English, those students with higher degrees of lexical diversity are generally judged by human evaluators to have greater lexical and overall language proficiency. Polinsky and Kagan [2007] has demonstrated that similar correlations can be seen in Heritage populations; she reports that, in her study of Heritage learners of Russian, speakers who knew more vocabulary words demonstrated better control of agreement, case markers, and subordination during spontaneous speech [Polinsky, 1995, Montrul, 2011].

Similarly, analyzing the acquisition and use of formulaic language is an important part of assessing language proficiency in L2 learners. Being able to use idioms, collocations, and other recurrent forms effectively is a major part of learning fluent, native-like command of a language [Paquot and Granger, 2012]. L2 learners often have considerable difficulty with collocations and idioms [Nesselhauf, 2003, Paquot and Granger, 2012]. Thewissen [2008] found that even in advanced L2 learners, production of error-free idioms and collocations continues to be difficult. Foster [2013] found that, when compared to native speakers, L2s tend to neglect formulaic phrases in favor of individual words. However, Ping [2009] found that L2 learners of Chinese use four times the number of recurrent 4-grams as do native Chinese speakers. So, while L2s seem to be struggling with formulaic idioms and restricted collocations, they seem to be overusing recurrent lexical bundles. According to Paquot and Granger [2012] "the overall picture that emerges from learner-corpus-based studies is that learners' use of co-occurring combinations is characterized by a mixture of underuse, overuse, and misuse."

Native speakers, on the other hand, seem to gain observable processing advantages from the use of formulaic language, indicating that natives store idioms, collocates, and common lexical bundles in their minds as single units [Sosa and MacFarlane, 2002]. Siyanova-Chanturia et al. [2011] found that natives are able to process figurative idioms significantly faster than non-figurative controls in an eye-tracking task. Similarly, Arnon and Snider [2010] found that natives are able to reliably process common 4-grams faster than non-native speakers. According to Conklin and Schmitt [2012] past studies reliably indicate that the use of formulaic language imparts of processing advantage to native speakers, but that this advantage is much less prevalent among L2s; only the most advanced L2s in Conklin & Schmitt's study showed some processing advantage from the use of formulaic language. So, where do Heritage learners fit into this picture? Are they, like native speakers, receiving a distinct processing advantage from the use of formulaic language? Or has their mastery of formulaic language been impaired by a somewhat restricted experience with their Heritage language?

The use of corpora to study L2 language development is well supported (see Asención-Delaney and Collentine [2011]). However, according to Montrul [2011], there are very few studies which address lexical knowledge of Heritage learners. Further, I have been unable to find any studies which specifically address formulaic language use among Heritage learners. While L2 writers are assumed to have a more limited lexicon than native speakers, both in terms of individual words [Crossley and Skalicky, 2019] and stored fixed expressions and lexical bundles [Paquot and Granger, 2012], the relationship between the lexical development of L2 learners and Heritage learners has not been thoroughly investigated. Montrul [2012] states that "another area of interest that remains highly unexplored is lexical knowledge and representation" in L2 and Heritage learners. The present study seeks to begin to address this gap in the literature. According to Montrul [2011], the "relationship between grammar and the lexicon needs to be explored more closely" as it has important implications for future pedagogical and assessment research.

5.3. Methods and implementation

To analyze the lexical diversity and lexical density of L2 and Heritage learners of Spanish in the COWS-L2H corpus, I first extract essays from the accompanying metadata. Next, all essays in the corpus were automatically part-of-speech (PoS) tagged and lemmatized using FreeLing [Padró and Stanilovsky, 2012]. FreeLing is a good choice as it was initially developed at Universitat Politècnica de Catalunya specifically for processing and analyzing Spanish text; more commonly used NLP toolkits, such as Spacy and Stanford's CoreNLP, were initially focused on processing English text and were later expanded to include additional languages. Further, an informal comparison of the output of FreeLing and Spacy by graduate student researchers in Spanish found the output of FreeLing to be superior to that of Spacy. After processing the text with FreeLing, the resulting output consists of a tab-separated text file in which each line contains a word from the text followed by its corresponding lemma form and PoS tag.

Once the essays are lemmatized and PoS tagged, I am able to analyze lexical density and lexical diversity. Lexical density is measured using a simple count of nouns, verbs, adjectives and adverbs in each text, divided by the total number of tokens in the text. Due to the fact that the COWS-L2H data is annotated for PoS using automated annotation, it is likely that the accuracy of its PoS tags, and the lexical density score derived therefrom, is less reliable than those in hand-annotated corpra, such as the Corpus del Español. It is also important to note that, while FreeLing was developed to process Spanish language texts, the system was trained using native speaker writing; since I am using the system to tag and lemmatize learner writing, the language distribution differences between FreeLing's training data and my corpus data may be a source of additional error. Data drawn from CEDEL2 is also annotated for POS and lemma using FreeLing. Because Corpus del Español is annotated for lemma and PoS tag, automated annotation of this data was not needed to calculate lexical density.

To calculate lexical diversity, I use the Measure of Textual Lexical Diversity (MTLD) [McCarthy and Jarvis, 2010], which seeks to normalize type-token ratio (TTR) for text length, thereby allowing a meaningful comparison of texts with differing lengths. MTLD is the "mean length of sequential word strings in a text that maintain a given TTR value" [McCarthy and Jarvis, 2010]. For each such span of words in the text, the "factor count" is incremented by one and the running TTR is reset to zero. Once the end of the text is reached, the total number of tokens in the text is divided by the total factor count to give the final MTLD score. While the "given TTR value" on which an MTLD is based can be set by the researcher, in their research McCarthy and Jarvis found that the "given TTR value" which results in the most stable MTLD score across texts of varying lengths is 0.720 [McCarthy and Jarvis, 2010]. In a AAAL presentation in 2019, Kristopher Kyle stated that, while the optimal index TTR value is dependent on the texts being evaluated, the default MTLD index of 0.720, as proposed by McCarthy & Jarvis, is generally a stable and reliable choice [Kyle, 2019]. Therefore, I also use an index TTR value of 0.720 in calculating the MTLD of the texts in this study. McCarthy & Jarvis also demonstrate that MTLD is generally only stable for texts with lengths greater than 50 tokens; therefore, I limit my analysis to texts longer than 50 tokens. This restriction results in only a handful of exclusions from my analysis because students in the COWS-L2H corpus are asked to write essays between 250 and 500 words. To calculate MTLD, I use the Python package "Lexical Diversity" by John Frens [Frens et al., 2018]. Because Corpus del Español is drawn from internet data, there are many instances of unusual punctuation throughout the corpus, such as words preceded by hashtags (ex., "#FotoMovil") and stray "@" symbols. In order to consider only words in my analysis, I excluded all non-alphabetic symbols, and strings containing non-alphabetic symbols, from my analysis of lexical diversity.

To compare the differences in lexical diversity between advanced L2 learners and Heritage learners, and between Heritage learners and Hispanosphere residents, I conduct simple independent-sample t-tests. I conduct a similar analysis for lexical density measurements on the same two groups. To analyze the use of formulaic language use of L2 and Heritage learners of Spanish in the COWS-L2H corpus, I extract all 3-grams and 4-grams from essays written by advanced L2 learners (those enrolled in SPA21, 22, 23 or 24 – a total of 104,906 tokens) and Heritage learners (those enrolled in SPA31, 32, or 33 – a total of 67,799 tokens). In order to reduce duplicate n-grams and to avoid n-grams which include punctuation, I removed all punctuation, and lowercased all text. During extraction, I tracked all occurrences of each n-gram to identify the most commonly recurring strings. Thus, I built a list of all 3-grams, and 4-grams used by these two groups of students in descending order of frequency. Once all n-grams were extracted for each group, I calculated the Mutual Information of each n-gram to aid in identifying those phrases which are truly collocates, as defined above, versus those that are recurring lexical bundles. Finally, I calculated the normalized frequency of each n-gram (normalized to occurrences per million tokens).

To identify the n-grams most frequently used by native Spanish speakers living in hispanophone countries, I repeated the above process for a randomly selected 100,000 token sub-corpus of the Corpus del Español, to control for the Zipfian distribution effect described by Bestgen [2020]. To ensure that the 100,000 sub-corpus is a representative sample of the full Corpus del Español, I extracted n-grams from the full corpus (which contains 1,983,822 tokens) for comparison. I compared the full corpus n-grams to those extracted from the randomly selected sub-corpus, and found that these two methods result in a similar most-common n-gram list with similar normalized frequencies.

5.3.1. Findings. As shown in Figure 5.1, the lexical diversity of L2 and L3 students in the COWS-L2H corpus increases at a steady rate as these students progress through the Spanish course series. While this is a testament to the fact that students and instructors are successful in fostering lexical development among L2 and L3 students, the same pattern of lexical growth cannot be seen among Heritage learners (see Figure 5.1). Rather, the average lexical diversity of Heritage learners remains rather flat throughout the three-course Heritage course series. In addition, the difference in mean lexical diversity between students enrolled in SPA24, the most advanced L2 course (mean LD of 56.04) and Heritage learners enrolled SPA31, the first course of the Heritage learner series (mean LD of 59.63) is not statistically significant at the 95% confidence level ($p = 0.1075$).

To compare with our Heritage learners with Spanish-dominant writers, I draw data from CEDEL2 [Lozano et al., 2009], which includes relatively small portion of data written to an identical prompt used in COWS-L2H. Comparing data written in response to the same prompt avoids task

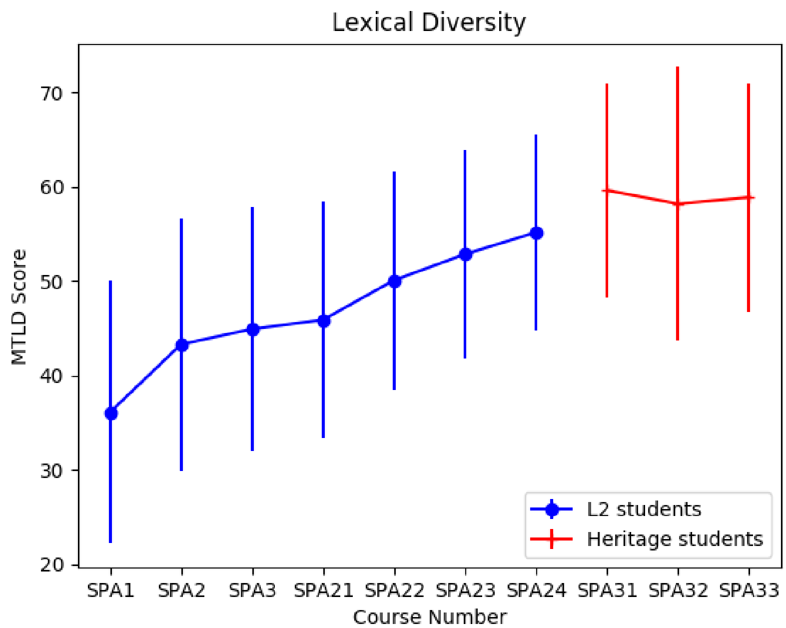

FIGURE 5.1. Lexical diversity by course in COWS-L2H essays.

effects and allows a more direct comparison between student groups. In this subset of data, MLTD values are generally lower than in the full COWS-L2H corpus; however, similar trends emerge, as shown in Figure 5.2. Advanced learners (those enrolled in SPA 21+) show a significantly higher MTLD than their lower-proficiency colleagues (44.0 versus 36.8, p-value: 0.003), while the advanced L2 learners show no significant difference with the Heritage cohort (44.0 versus 44.4, p-value: 0.872). Of note is the fact that both Heritage and advanced L2 learners have MLTD values significantly less than that of Spanish-dominant students whose mean MTLD is 50.26. This is significantly higher than both the advanced L2 (p-value: 0.005) and the Heritage learner (p-value: 0.015) groups. This finding seems to concur with previous work, such as Montrul and Ionin [2012], which indicates that the lexicon is highly susceptible to language attrition in Heritage speaker populations.

As mentioned previously, I find that lexical density in both L2 and Heritage learners in COWS-L2H decreases as students progress through the course series, as shown in Figure 5.3. Interestingly, this pattern is similar to the pattern Halliday and Halliday [1989] found in children acquiring their L1. According to Halliday, in children, lexical density decreases as the child's fluency in the language develops due to the fact that young children begin speaking in short utterances that consist of only content words. As the children begin to speak in longer phrases and, eventually, in sentences, they

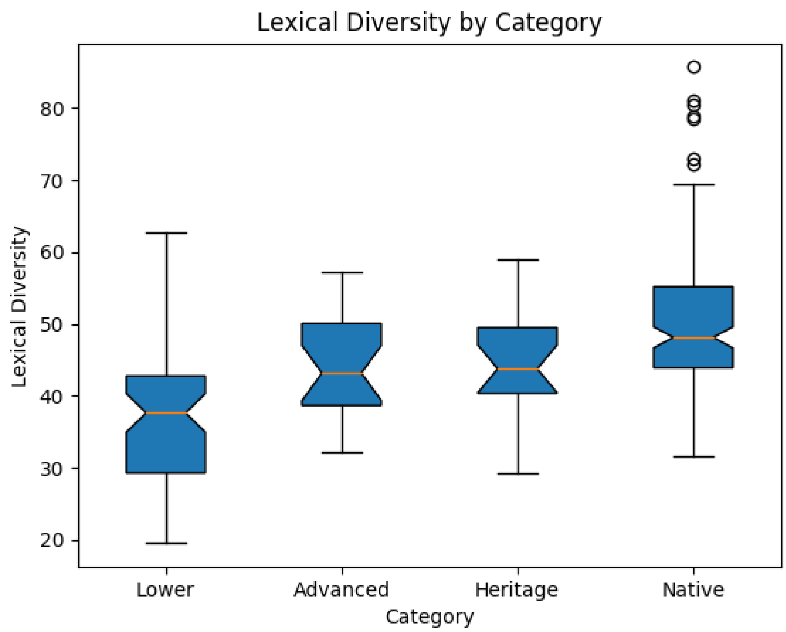

FIGURE 5.2. Lexical diversity by category in Chaplin essays from COWS-L2H and CEDEL2.

begin to use more function words, thereby reducing the content word to function word ratio of their language [Halliday and Halliday, 1989]. Thus, the decreasing pattern seen in L2 and Heritage learners may be indicative of increasing grammatical complexity in sentence formation.

I find that the mean difference between the lexical density of the advanced L2 learners (mean = 0.523) and Heritage learners (mean = 0.534) in COWS-L2H is small, but the difference of .011 is statistically significant (p = 0.016) at the 95% confidence level. In order to compare the Heritage population from COWS-L2H to a Spanish-dominant native speakers, I extract a subset of essays from both CEDEL2 and COWS-L2H, all of which were written in response to the same prompt. Students were asked to watch the same short clip of a movie and write a description of the scene; I refer to this subcorpus as the "Chaplin data." The Chaplin data consists of 69 lower-proficiency student essays (SPA 1-3), 26 advanced essays (SPA 21+), 23 Heritage essays, and 151 essays written by Spanish-dominant native speakers. The lexical density observed in this subcorpus is lower across-the-board than that observed in the corpus-wide data from COWS-L2H, as show in Figure 5.4. The first interesting observation is the fact that there is no statistically significant difference between the mean lexical density of lower-proficiency students (mean = 0.439) and their

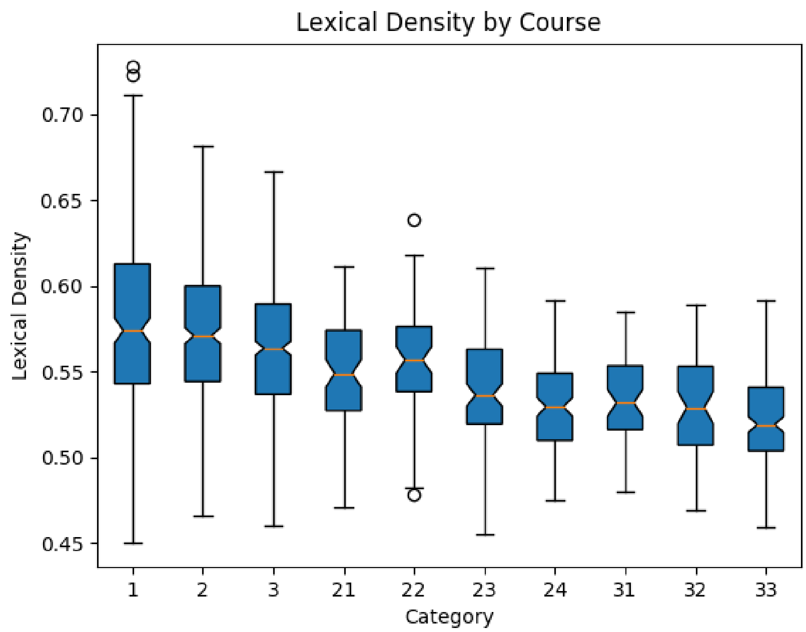

FIGURE 5.3. Lexical density by course in COWS-L2H essays.

more advanced counterparts (mean = 0.441); this finding is surprising, as in the larger corpus data a significant difference is observed between these same two groups (p-val <0.001). Second, the significant difference between advanced and Heritage learners observed in the Chaplin data concurs with observations across the corpus as a whole. In the Chaplin data, the advanced learner mean lexical density is 0.441, while the Heritage mean is 0.424; the p-value of 0.004 indicates that this relatively small difference is nonetheless statistically significant. The most interesting finding, though is the fact that, on the Chaplin task, the lexical density of Heritage learners (mean = 0.424) is very similar to that of Spanish-dominant native speakers (mean = 0.426), and the difference between the two groups is not statistically significant (p-value: 0.690). The present finding that the lexical density of Heritage and Native populations is statistically identical, even in this small sub-corpus, indicates that the Heritage students' writing may be more grammatically sophisticated that previously thought. Further research is needed to elucidate this finding, especially in comparison with the findings regarding syntactic development presented in Chapter 6

On the subject of formulaic language use, clear similarities between the L2 and Heritage groups emerge, as do differences between these two groups and the Spanish-dominant group. While the results are not clear enough to draw any overall conclusions, I found potentially revealing patterns

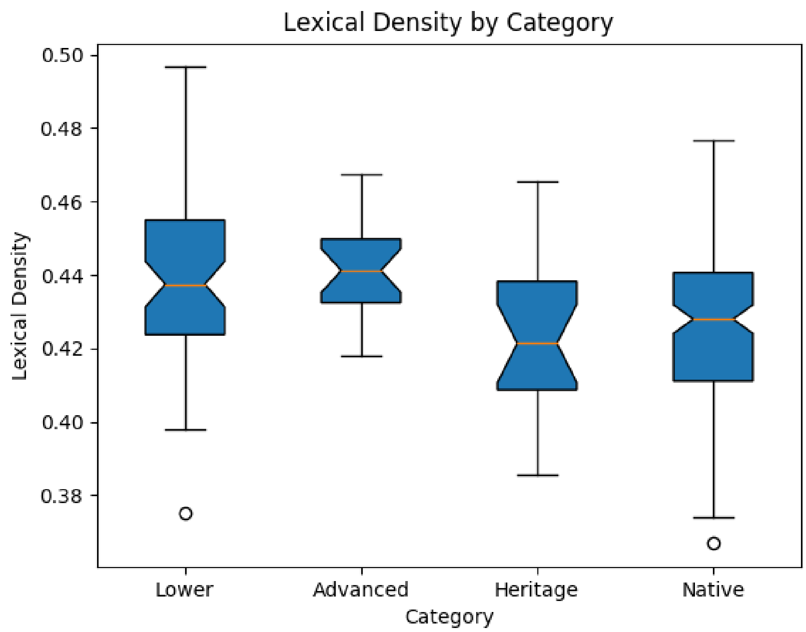

Figure 5.4. Lexical density by category in Chaplin essays from COWS-L2H and CEDEL2.

which require further investigation. First, it is clear that many of the recurrent n-grams in the COWS-L2H data are the result of task effects; that is, we are seeing specific n-grams recur because they are common to the topic which students are being asked to write about, not because learners use these phrases more frequently in general. For example, one of the most common trigrams in the COWS-L2H data, for both L2 and Heritage learners is "una vacación perfecta" with a normalized frequency of 467.08 per million for L2 and 560.48 per million in the heritage data. Of course, Spanish learners in general are not using this phrase with such a high frequency; rather, the high frequency of this particular n-gram is due to the fact that, as one of the prompts in the COWS-L2H data collection, students are asked to describe their idea of a perfect vacation. When I remove those n-grams from the COWS data which are clearly the result of task effects, I still find that learners, both L2 and Heritage, use formulaic language at a higher rate than do native, Spanish-dominant speakers. For example, the lexical bundle "que se llama" is used at a normalized frequency of 18.65 in the Corpus del Español, but this same expression is used at a frequency of 266.91 and 147.49 for L2 and Heritage learners, respectively, in the COWS data. Not all differences are so consistent; the collocation "todo el mundo" is used at a frequency of 125.51 in the native speaker data, and at a rate of 495.68 and 162.24 for L2 and Heritage learners, respectively. While both groups of learners use this collocation more frequently than do native Spanish-dominant writers, the L2 learners use it far more often than do the Heritage learners. On the other hand, with some collocations, such as “al mismo tiempo,” the Heritage learners are more similar in usage to the Spanish-dominant speakers than to L2 learners. However, with the collocation “por los menos,” the L2 learners are actually more similar to the native speakers than are the Heritage speakers. Finally, with the lexical bundle “la mayoría de,” all three groups use the expression at roughly the same rate. Clearly, these inconsistent findings require additional analysis to determine if an underlying pattern of usage exists among these three groups of speakers.

With regard to mutual information, I find that n-grams which represent idioms and lexically restricted collocations, such as “todos los días,” have higher mutual information than do phrases which represent recurrent, but lexically unrestricted, lexical bundles. This finding is consistent with Ellis (2008), who found that phrases with higher PMI are more likely to be rated as “formulaic” by human annotators. In addition, the average mutual information of the common n-grams in the Heritage and L2 data is far higher than the average PMI in the native Spanish-dominant data. The mean PMI for the L2 and heritage learners is 2.52 and 1.93, respectively, while the mean PMI for the native Spanish-dominant data is 0.28. While I have yet to determine the underlying reason for this difference, this may be the result of less varied lexical use among L2 and Heritage learners compared to native speakers. Again, we must be aware that these differences may be the result of task or genre effects; however, at minimum these findings indicate that further research into the development and use of lexicon and formulaic language in Heritage learners is warranted.

5.4. Discussion

Given my findings, I am able to make three specific statements regarding the relative lexical development of L2 and Heritage speakers:

(1) The difference between the lexical diversity scores of Heritage learners and L2 learners of Spanish is statistically insignificant.

(2) The difference between the lexical density scores of Heritage and L2 learners of Spanish is statistically significant, though it is small.

(3) The lexical density of Heritage learners more closely resembles that of Spanish-dominant native speakers than it does L2 learners of Spanish, though the differences between all three groups are relatively small.

These findings potentially confirm my hypothesis that the lexical diversity of Heritage learners may be closer to that of advanced L2 learners of Spanish than it is to that of native Spanish speakers. However, my findings seem to falsify the similar hypothesis regarding lexical density. In addition, while the evidence is sparse, it appears that the lower lexical diversity of L2 and Heritage learners results in higher mutual information in the formulaic language used by these speakers. We must be careful in interpreting these findings, however, as MTLD and other measures of lexical diversity are known to show significant differences based on writing task alone [Yu, 2010]. I cannot rule out the possibility that the difference in MTLD score seen between Heritage and non-Heritage native speakers is due to differences in data collection and task presentation. The same is true with relation to the production of specific, topic-related n-grams in the COWS corpus. Because the COWS participants are writing about specific topics in response to specific prompts, n-grams related to those topics appear far more frequently in the COWS data than would be expected in a dataset covering more diverse topics.

Another possible weakness with the present study is that, in the COWS-L2H corpus, enrollment in Heritage courses is completely voluntary. Given the metadata we collect from students, it is apparent that a fair number of students who have significant childhood exposure to Spanish, and who would likely be classified as Heritage learners under Valdés' definition, are enrolled in L2 courses. This is especially true of the more advanced-level L2 courses. Thus, these Heritage students enrolled in L2 courses could be artificially causing the L2 population to appear more similar to the Heritage group. While we ask students about their linguistic background, and I filter the data to remove students who report Heritage knowledge of Spanish, the diverse linguistic background of the student population may be muddying the waters. Future research could improve upon the present study by categorizing student texts more carefully based on the information provided in demographic surveys. Although this study has tentatively confirmed my hypothesis with regard to lexical diversity and tentatively falsified my hypothesis related to lexical density, much future work is necessary to make these findings more robust and to answer the remaining questions regarding formulaic language use.

CHAPTER 6

COWS-L2H Study II: Syntactic development

Having presented data related to lexical development in L2 and Heritage learners, I now turn to a study of syntactic development in L2, L3 and Heritage learners of Spanish. Specifically, the following study seeks to verify if it is possible to predict the relative order of essays written by the same student from the longitudinal subset of COWS-L2H, and which syntactic features are most important in making these predictions.

In the last few years, research on language acquisition has benefited from the use of Natural Language Processing (NLP) technologies applied to large–scale corpora of authentic texts produced by learners, in both the first and second language context. The empirical evidence acquired from learner corpora, complemented with the increased reliability of linguistic features extracted by computational tools and machine learning approaches, has promoted a better understanding of learners' language properties and how they change across time and increasing proficiency level [Crossley, 2020]. A first line of research has focused on providing automatic ways of operationalizing sophisticated metrics of language development to alleviate the laborious manual computation of these metrics by experts [Sagae et al., 2005, Lu, 2009]. A second line of research has taken the more challenging step of implementing completely data-driven approaches, which use a variety of linguistic features extracted from texts to automatically assign a learner's language production to a given developmental level [Lubetich and Sagae, 2014].

A great amount of work has been carried out in the field of second language acquisition where the study of L2 writings is seen as a proxy of language ability development [Crossley, 2020]. In this respect, much related work is devoted to predicting the degree of second language proficiency according to expert–based evaluation [Crossley and McNamara, 2012] or to modelling the evolution of grammatical structures' competence with respect to predefined grades, such as the Common European Framework of Reference for Languages (CEFRL) [Zilio et al., 2018].

Given the difficulty of defining a unique indicator of linguistic complexity in the context of L2 language development, a great variety of features from all linguistic levels have been used as input for supervised classification systems trained on authentic learner data for different L2s. Such is the case e.g. of Hancke and Meurers [2013] and Vajjala and Lėo [2014], dealing with L2 German and L2 Estonian, respectively, and of Pilán [2018], who also provided a features analysis focused on predictive features extracted from both receptive and productive texts in Swedish L2. They found that a small number of features covering lexical and syntactic aspects are useful predictors across all datasets regardless the type of language learning skill.

In this section, I present an innovative NLP-based stylometric approach to model writing development in learners of Spanish as a second and Heritage language.

This study's approach relies on a wide set of linguistically motivated features extracted from students' essays, which have already been shown relevant for a number of tasks related to modelling the 'form' of a text rather than the content. While the majority of previous studies on the evolution of language proficiency in L2 uses cross–sectional data, this study is the first, to our knowledge, using a longitudinal corpus of Spanish L2 essays to model writing development. Interestingly, a similar approach resulted in the successful prediction of the development of writing competence in a L1 acquisition scenario for the Italian language [Richter et al., 2015].

This study presents, to the best of our knowledge, the first data–driven study which uses linguistic features from student data to model the evolution of written language competence in Spanish as a Second Language (SSL). The results also show that it is possible to automatically predict the relative order of two essays written by the same student at different course levels using a wide spectrum of linguistic features. Finally, the study investigates the importance of linguistic features in predicting language growth at different course levels and whether they reflect the explicit instruction students receive during each course.

6.1. Motivation and approach

Studies of L2 writing have focused on linguistic complexity as an indicator of writing development [Lu, 2011, Ortega, 2003]. This construct, however, is still ill-defined, as evidenced by the divergent measures of complexity utilized in different studies. Typical measures of complexity have been the length of the T-unit [Hunt, 1965], the number of subordinate clauses in a text, or type to token ratios, among others. Instead of considering the construct as being multidimensional [Norris and Ortega, 2009, Bulté and Housen, 2012] and, thus, encompassing an array of different features, most studies have selected one or two of these measures and used them as single indicators of complexity [Bulté and Housen, 2014]. This has prevented the development of much needed research that associates different steps of linguistic and written development with specific sets of characteristics. This situation has also prevented the formation of an in-depth picture of how those specific aspects develop in relation to the grammatical, lexical or stylistic content taught in classes at different language course levels. This second objective of characterizing writing at different proficiency levels may provide useful insights into how writing samples could be used for placement tests or other assessments to determine which language course is best suited to further develop a student's linguistic skills.

In the concrete case of Spanish as a Second Language (SSL), the literature indicates that one of the most difficult aspects to master for learners is the language's complex verbal morphology [Blake and Zyzik, 2016, Salaberry, 1999], given that verbal inflections express a complex cluster of person, number, tense, aspect and mood. Therefore, SSL courses tend to propose a step-by-step introduction to these different aspects of verbal morphology, generally following this order: (1) person and number in the present indicative, (2) past tenses (i.e., imperfect vs. preterite vs. pluperfect), and (3) mood (subjunctive vs. indicative). If this typical instructional sequence had to influence students' writing, it would be expected that learners show an increase in the variety of inflections that they are able to use over time. Nonetheless, several studies also indicate that a linguistic feature that has been learned in class may be mastered in exercises that focus on explicit knowledge but take additional time to unfold in tasks that require more implicit knowledge, such as free writing [Ellis and Shintani, 2013]. This means that a simple classification of students' proficiency based on the presence or absence of features studied in a particular course may not be accurate, as some students may explicitly know the rules for a specific inflectional distinction but still be unable to use them accurately in writing. Taking lack of use in writing as evidence for lack of explicit knowledge could entail that students be mistakenly invited to enroll in courses where those features that do not show in their writing are unnecessarily explained to them again. A better approach would thus be to know what students are able to do when they are enrolled in different courses and, only then, compare those abilities to see which match, or mismatch, the contents seen in that particular class. By using a large set of linguistic features, it is possible to understand which phenomena change across proficiency levels and whether they are explicitly related to the teaching guidelines.

This study aims at tackling some of the still open methodological issues in the literature on Spanish acquisition by decomposing the problem into two main research questions: *(i)* verify if it is possible to predict the relative order of two essays written by the same student at different course levels using a wide set of linguistic predictors automatically extracted from Spanish L2 written productions; *(ii)* understand which typologies of language phenomena contribute more to the identification of writing skills' evolution and whether such properties reflect the teaching guidelines of the courses.

Following the approach devised in Richter et al. [2015] we addressed the first research question as a classification task: given a pair of essays written by the same student and ordered according to the course level (d_1, d_2), we classify whether $C(d_2) > C(d_1)$, where $C(d_1)$ and $C(d_2)$ correspond respectively to the course levels during which the student wrote d_1 and d_2. Specifically, we model the problem as a binary classification task, training a Linear Support Vector Machine (LinearSVM) to predict the relative order of two essays written by the same student using a wide range of linguistic predictors automatically extracted from the POS tagged and dependency parsed essays. We rely on LinearSVM rather than more powerful learning algorithms, such as Neural Language Models, in order to obtain meaningful explanations when the classifier outputs its predictions to anchor the observed patterns of language development to explicit linguistic evidence.

We further extracted and ranked the feature weights assigned by the linear model in order to understand which typology of linguistic features contributes more to the classification task at different course levels. The assumption is that the higher the weight associated with a specific feature, the greater its importance in solving the classification task and, consequently, in modeling the student's written language evolution.

We focused our study on the longitudinal data in the COWS-L2H corpus. We were thus able to model the chronological development of L2 Spanish writing by monitoring how the writing quality of an individual student's compositions increase with time. Student participation is summarized in Table 6.1.

6.2. Linguistic Features

The set of linguistic features considered as predictors of L2 written competence evolution is based on those described in Brunato et al. [2020]. It includes a wide range of text properties, from

Terms Enrolled	Students	Essays	Tokens
2	267	984	290,399
3	111	612	179,306
4	32	242	74,956
5	5	48	13,977

TABLE 6.1. Longitudinal data summary.

Level of Annotation	Linguistic Feature	Label
Raw Text	Sentence Length	tokens_per_sent
	Word Length	char_per_tok
	Document Length	n_sentences
	Type/Token Ratio for words and lemmas	ttr_form, ttr_lemma
POS tagging	Distribution of UD and language–specific POS	upos_*, xpos_*
	Lexical density	lexical_density
	Inflectional morphology of lexical verbs and auxiliaries	verbs_*, aux_*
Dependency Parsing	Depth of the whole syntactic tree	parse_depth
	Average length of dependency links and of the longest link	links_len, max_links_len
	Average length of prepositional chains and distribution by depth	prepositional_chain_len, prep_dist_*
	Clause length (n. tokens/verbal heads)	token_per_clause
	Order of subject and object	subj_pre, subj_post, obj_pre, obj_post
	Verb arity and distribution of verbs by arity	verb_edges, verb_edges_*
	Distribution of verbal heads per sentence	verbal_head_sent
	Distribution of verbal roots	verbal_root_perc
	Distribution of dependency relations	dep_dist_*
	Distribution of subordinate and principal clauses	principal_proposition_dist, subord_dist
	Average length of subordination chains and distribution by depth	subord_chain_len, subord_*
	Relative order of subordinate clauses	subord_post, subord_prep

TABLE 6.2. Linguistic features according to different levels of annotation.

raw text features, to lexical, morpho-syntactic and syntactic properties, which were extracted from different levels of linguistic annotation. For this purpose, the COWS-L2H Corpus was automatically parsed using UDPipe [Straka et al., 2016] trained on the Spanish Universal Dependency Treebank (GSD section), version 2.5. We rely on these features since it has been shown that they have a high predictive power for several tasks all aimed at modelling the linguistic *form* of documents. This is the case for example of the automatic readability assessment task [Dell'Orletta et al., 2011a], of the automatic classification of the textual genre of documents [Cimino et al., 2017], or also of the automatic identification of the L1 of a writer based on his/her language production in a L2 [Cimino et al., 2018]. Interestingly, for all mentioned tasks the set of linguistic features plays a very important role in the classification not only of a whole document but also of each single sentence. This is the reason why, as reported in the following sections, we modelled the prediction of the development of writing skills both as document and sentence classification tasks.

Although we used a state–of–the art pipeline, it is well-acknowledged that the accuracy of statistical parsers decreases when tested against texts of a different typology from that used in training Gildea [2001]. In this respect, learners' data are particularly challenging for general–purpose text analysis tools since they can exhibit deviation from correct and standard language; for instance, missing or anomalous use of punctuation (especially in 1st grade prompts) already impacts on the coarsest levels of text processing, i.e. sentence splitting, and thus may affect all subsequent levels of annotation. Nevertheless, if we can expect that the predicted value of a given feature might be different from the real one (especially for features extracted from more complex levels of annotation such as syntax), we can also assume that the distributions of errors will be almost similar, at least when parsing texts of the same domain. Note also that the reliability of features checked against automatically annotated data was also empirically shown by Dell'Orletta et al. [2011b], who compared morpho-syntactic and syntactic features extracted from a gold (i.e. manually annotated) and an automatically annotated corpus of the same domain (i.e. biomedical language), showing that results are highly comparable.

As shown in Table 6.2, the considered features capture linguistic phenomena ranging from the average length of document, sentences and words, to morpho-syntactic information such as parts of speech (POS) distribution and fine–grained features about the inflectional properties of verbs. More complex phenomena are derived from syntactic annotation and model global and local properties of parsed tree structure, with a focus on subtrees of verbal heads, the order of subjects and objects with respect to the verb, the distribution of Universal Dependencies (UD) syntactic relations and features referring to the use of subordination.

Since it is acknowledged that lexical proficiency plays an important role in predicting L2 writing development [Crossley and McNamara, 2012], we also decided to add a small subset of features that model this property in terms of word frequency. Specifically, we considered the average class frequency of all word forms and lemmas in the essays (*Words Frequency Class*), where the class frequency for each word form/lemma was computed exploiting the Spanish Wikipedia (dump of March 2020) using the following measures: $C_{cw} = \lfloor \log_2 \frac{freq(MFW)}{freq(CW)} \rfloor$, $C_{cl} = \lfloor \log_2 \frac{freq(MFL)}{freq(CL)} \rfloor$, where *MFW* and *MFL* are the most frequent word form/lemma in the corpus and *CW* and *CL* are the considered ones.

Features	SPA 1	SPA 2	SPA 3	SPA 21	SPA 22	SPA 23	SPA 24	SPA 31	SPA 32	SPA 33
Raw Text Properties										
char_per_tok	4.3 ±.27	4.4 ±.27	4.42 ±.26	4.42 ±.26	4.43 ±.25	4.46 ±.23	4.41 ±.22	4.42 ±.25	4.42 ±.28	4.38 ±.3
n_sentences	20.0 ±7.0	24.01 ±7.15	23.57 ±6.87	20.8 ±5.99	20.17 ±5.15	19.54 ±6.33	17.92 ±5.44	16.06 ±4.05	16.31 ±3.78	15.46 ±3.63
tokens_per_sent	10.7 ±3.43	13.16 ±3.52	13.74 ±3.7	15.71 ±3.95	16.43 ±3.59	17.11 ±3.49	19.01 ±4.27	19.95 ±4.16	20.07 ±3.48	20.94 ±4.04
Morphosyntactic information										
lexical_density	.51 ±.05	.5 ±.04	.5 ±.04	.49 ±.03	.48 ±.04	.48 ±.03	.47 ±.03	.48 ±.04	.47 ±.04	.47 ±.04
upos_ADJ	.07 ±.03	.06 ±.02	.06 ±.02	.06 ±.02	.05 ±.02	.05 ±.02	.05 ±.02	.05 ±.02	.05 ±.02	.05 ±.02
upos_ADP	.09 ±.04	.1 ±.03	.11 ±.03	.11 ±.02	.11 ±.02	.12 ±.02	.12 ±.02	.13 ±.03	.12 ±.02	.13 ±.02
upos_NOUN	.16 ±.04	.16 ±.03	.16 ±.03	.16 ±.03	.16 ±.03	.17 ±.02	.17 ±.03	.17 ±.02	.16 ±.03	.16 ±.03
upos_PRON	.07 ±.04	.07 ±.03	.07 ±.03	.07 ±.03	.07 ±.03	.07 ±.03	.07 ±.03	.07 ±.03	.08 ±.04	.08 ±.04
upos_PUNCT	.14 ±.03	.13 ±.03	.12 ±.03	.12 ±.03	.11 ±.03	.11 ±.03	.11 ±.03	.09 ±.02	.09 ±.02	.09 ±.02
upos_SCONJ	.01 ±.01	.02 ±.01	.03 ±.01	.03 ±.02	.04 ±.02	.04 ±.01	.04 ±.02	.04 ±.02	.05 ±.02	.05 ±.02
upos_VERB	.12 ±.04	.12 ±.03	.12 ±.03	.12 ±.02	.12 ±.02	.12 ±.02	.12 ±.02	.13 ±.02	.13 ±.02	.13 ±.03
Inflectional morphology										
aux_mood_Cnd	.02 ±.09	.03 ±.09	.04 ±.12	.03 ±.07	.06 ±.11	.05 ±.11	.04 ±.08	.05 ±.09	.06 ±.12	.04 ±.11
aux_mood_Ind	.97 ±.14	.96 ±.12	.92 ±.15	.94 ±.14	.91 ±.13	.92 ±.14	.94 ±.1	.91 ±.16	.91 ±.12	.93 ±.12
aux_mood_Sub	.01 ±.04	.01 ±.04	.03 ±.07	.02 ±.05	.03 ±.05	.02 ±.08	.03 ±.06	.03 ±.06	.03 ±.06	.03 ±.05
aux_tense_Imp	.05 ±.16	.16 ±.25	.21 ±.26	.21 ±.25	.24 ±.25	.24 ±.26	.22 ±.24	.23 ±.28	.2 ±.27	.24 ±.29
aux_tense_Past	.02 ±.09	.1 ±.15	.09 ±.15	.12 ±.16	.12 ±.14	.11 ±.15	.12 ±.16	.11 ±.16	.12 ±.17	.11 ±.13
aux_tense_Pres	.92 ±.21	.73 ±.32	.69 ±.33	.65 ±.32	.63 ±.3	.65 ±.32	.66 ±.32	.63 ±.34	.66 ±.34	.63 ±.33
verbs_tense_Imp	.02 ±.06	.08 ±.12	.11 ±.13	.13 ±.13	.16 ±.14	.14 ±.15	.13 ±.13	.17 ±.15	.15 ±.15	.14 ±.14
verbs_tense_Past	.11 ±.19	.28 ±.23	.28 ±.22	.3 ±.2	.35 ±.22	.3 ±.22	.31 ±.19	.31 ±.21	.28 ±.18	.33 ±.19
Verbal Predicate Structure										
verb_edges	2.3 ±.36	2.5 ±.32	2.52 ±.3	2.62 ±.35	2.67 ±.28	2.63 ±.28	2.7 ±.32	2.71 ±.29	2.68 ±.26	2.76 ±.27
verb_edges_4	.09 ±.08	.13 ±.07	.13 ±.07	.16 ±.07	.16 ±.07	.15 ±.08	.16 ±.07	.16 ±.06	.16 ±.06	.16 ±.07
verbal_head_sent	1.52 ±.46	1.8 ±.53	1.92 ±.52	2.13 ±.54	2.26 ±.54	2.3 ±.51	2.54 ±.61	2.73 ±.58	2.86 ±.65	2.95 ±.66
Global and Local Parsed Tree Structures										
parse_depth	2.88 ±.65	3.27 ±.62	3.37 ±.61	3.6 ±.63	3.78 ±.55	3.94 ±.64	4.21 ±.69	4.49 ±.65	4.59 ±.67	4.56 ±.62
max_links_len	.65 ±.44	.7 ±.45	.72 ±.42	.96 ±.74	.92 ±.43	.99 ±.42	1.2 ±.68	1.24 ±.53	1.21 ±.42	1.39 ±.72
5rtoken_per_clause	7.17 ±1.56	7.49 ±1.58	7.28 ±1.39	7.52 ±1.51	7.41 ±1.26	7.55 ±1.26	7.62 ±1.24	7.42 ±1.3	7.16 ±1.09	7.26 ±1.32
Order of elements										
obj_post	.67 ±.18	.68 ±.15	.67 ±.15	.64 ±.16	.65 ±.15	.69 ±.13	.69 ±.14	.6 ±.17	.64 ±.17	.6 ±.16
obj_pre	.33 ±.18	.32 ±.15	.33 ±.15	.35 ±.15	.35 ±.15	.31 ±.13	.31 ±.14	.39 ±.16	.36 ±.17	.4 ±.16
subj_pre	.8 ±.19	.84 ±.15	.82 ±.15	.84 ±.15	.84 ±.13	.84 ±.13	.83 ±.13	.81 ±.12	.78 ±.13	.79 ±.14
Use of Subordination										
subord_chain_len	1.06 ±.25	1.15 ±.16	1.18 ±.14	1.21 ±.18	1.24 ±.15	1.24 ±.14	1.26 ±.16	1.29 ±.23	1.33 ±.16	1.32 ±.2
subord_2	.08 ±.14	.11 ±.11	.13 ±.1	.15 ±.11	.17 ±.1	.17 ±.11	.18 ±.11	.19 ±.11	.2 ±.1	.2 ±.1
subord_dist	.24 ±.14	.33 ±.13	.38 ±.12	.4 ±.12	.44 ±.12	.47 ±.12	.5 ±.12	.56 ±.12	.58 ±.08	.57 ±.1

TABLE 6.3. A subset of linguistic features extracted for each course level. For each feature it is reported the average value and the standard deviation.

A first overview of how and to what extent all these features vary across the documents of the COWS-L2H Corpus is provided in Table 6.3. Essays written by students in the first course levels are longer in terms of number of sentences but they contain shorter sentences compared with those written in the more advanced courses. As concerns the distribution of POS, essays written in the first years show a lower percentage of e.g. adpositions (*upos_ADP*) and subordinate conjunctions (*upos_SCONJ*) typically contained in longer and well-articulated sentences, while the use of main content words (e.g. *upos_NOUN*, *upos_VERB*) is almost comparable across years. The variation affecting morphosyntactic categories is reflected by the lexical density value, i.e. the ratio between content words over the total number of words, which is slightly higher in beginner essays. If we

Course Levels	Num Essays
SPA 1	205
SPA 2	409
SPA 3	413
SPA 21	160
SPA 22	130
SPA 23	140
SPA 24	120
SPA 31	65
SPA 32	83
SPA 33	85

Table 6.4. Number of essays included in the longitudinal section of the COWS-L2H corpus.

focus on differences concerning verbal morphology, a linguistic property particularly relevant in the development of Spanish curriculum, we can see how the use of more complex verb forms increases across course levels. Essays of the introductory courses contain a lower percentage of verbs in the past (*verbs_tense_Past*) and imperfect tenses (*verbs_tense_Imp*) (out of the total number of verb tenses) as well as a lower percentage of auxiliary verbs (*aux_**) typically used in more complex verb forms, such as copulative verbs or periphrastic moods and tenses. Interestingly, features related to verb inflectional morphology have the highest standard deviation, suggesting a quite wide variability among learners. A similar trend towards the acquisition of more complex verb structures can also be inferred by considering features extracted from the syntactic level of annotation: essays of the intermediate courses contain for example sentences with a higher average number of dependents of verbs (*verb_edges*) and in particular of verbs with a complex argument structures of 4 dependents (*verb_edges_4*).

As long as Spanish learners start mastering the second language, linguistic properties related to the construction of more complex sentences increase. This is for example the case of the depth of sentence tree (*parse_depth*) and of the length of syntactic relations (*max_links_len*) as well as of features concerning the use of subordination.

6.3. Experiments

We train a LinearSVM that takes as input pairs of essays written by the same students according to all the possible pairs of course levels (e.g. SPA 1 - SPA 2, SPA 2 - SPA 3, etc.). Specifically, we extract for each pair the linguistic features corresponding to the first and second essays and the difference between them. We standardize the input features by scaling each component in the range $[0, 1]$. To test the actual efficiency of the model, we perform the experiments with a 5-cross validation using different students during the training and testing phases. In order to provide our system with negative samples, we expand our datasets by adding reversed samples.

Since the students were asked to write essays responding to different prompts, we devise two set of experiments, pairing all the essays written by the same students that have: (i) the same prompt; (ii) both same and different prompts. Also, because of the small number of training samples for certain pairs of course levels we also decide to perform the experiments on a sentence-level, extracting the linguistic features for each sentence in the longitudinal subset of the COWS-L2H corpus and pairing them on the basis of the previously defined criteria. In order to obtain reliable results both on the document and sentence configurations, we consider only datasets at different pairs of course levels that contain at least 50 and 20 samples (including negative pairs) respectively. All the classification experiments are performed using the majority class classifier as baseline and accuracy as the evaluation metric.

6.4. Tracking Writing Skills' Evolution

Table 6.5 reports the results obtained at both the document and sentence levels, pairing essays that have the same prompt (*Same* columns) and both the same and different prompts (*All* columns). As a general remark, we observe that best results are those obtained with the document-level experiments. This is quite expected, since sentence-level classification is a more complex task that often requires a higher number of features to gain comparable accuracy [Dell'Orletta et al., 2014]. If we focus instead on the distinction between *Same* and *All* results, we notice that higher scores are mainly achieved considering pairs of essays that also have different prompts. Again, this result is not surprising because adding pairs of essays with different prompts within each datasets increases the number of training samples, thus leading to better scores. Despite this, the results obtained according to the *Same* and *All* configurations are quite similar and this allows us to confirm that classification accuracy is not significantly harmed if the two essay's prompts are the same, thus showing that our system is actually focusing on written language competence evolution properties rather than prompt-dependent characteristics.

Course Levels	Documents				Sentences			
	Same		All		Same		All	
	Score	Samples	Score	Samples	Score	Samples	Score	Samples
All Levels	0.68	2,208	0.7	5,536	0.59	1,047,156	0.61	2,570,366
SPA 1 - SPA 2	0.88	280	0.9	624	0.7	143,660	0.71	316,264
SPA 1 - SPA 3	0.97	178	0.95	440	0.75	85,032	0.75	209,048
SPA 1 - SPA 21	#	#	0.91	116	0.61	14,298	0.7	46,738
SPA 2 - SPA 3	0.62	528	0.62	1,192	0.56	323,332	0.56	724,400
SPA 2 - SPA 21	0.61	62	0.61	188	0.57	35,754	0.58	104,442
SPA 2 - SPA 22	#	#	0.59	68	0.55	8,048	0.63	29,670
SPA 2 - SPA 23	#	#	0.77	52	#	#	0.58	27,420
SPA 3 - SPA 21	0.59	158	0.55	364	0.53	82,104	0.54	190,596
SPA 3 - SPA 22	0.61	64	0.58	186	0.54	31,886	0.6	93,486
SPA 3 - SPA 23	#	#	0.89	106	0.59	13,404	0.59	45,804
SPA 3 - SPA 24	#	#	#	#	#	#	0.68	11,276
SPA 21 - SPA 22	0.59	132	0.62	302	0.52	57,326	0.54	132,454
SPA 21 - SPA 23	0.52	58	0.74	154	0.54	27,038	0.57	67,634
SPA 21 - SPA 24	#	#	0.7	92	0.47	9,268	0.56	35,384
SPA 22 - SPA 23	0.71	76	0.69	186	0.55	35,272	0.56	79,168
SPA 22 - SPA 24	0.69	158	0.73	164	0.5	23,446	0.56	66,184
SPA 23 - SPA 24	0.45	168	0.49	386	0.48	61,654	0.49	137,786
SPA 31 - SPA 32	0.8	100	0.63	212	0.55	27,608	0.55	57,790
SPA 31 - SPA 33	0.52	100	0.53	198	0.51	24,830	0.48	48,990
SPA 32 - SPA 33	0.54	96	0.59	256	0.5	24,154	0.55	66,466

Table 6.5. Classification results in terms of accuracy obtained both at document and sentence levels along with number of samples for each dataset. **Same** and **All** columns report the results obtained by pairing essays that have same prompt and both same and different prompts respectively. Since the labels within each dataset has been balanced, baseline accuracy is 0.50.

More interestingly, we notice that considering all the possible course level pairs at the same time our system is able to achieve quite good results, especially at document level classification (0.68 and 0.70 of accuracy for *Same* and *All* configurations respectively), thus showing that it is possible to automatically predict the chronological order of two essays written by the same student by using a wide spectrum of linguistic properties.

In general, our best scores are obtained by considering all the experiments that include essays written by students in the Beginner category (SPA 1, 2 and 3). This is particularly evident for the experiments that compare essays written during SPA 1 as one of the two considered course levels, most likely because the evolution from knowing nothing at all of a specific L2 to knowing enough to start writing is actually bigger that the difference between knowing a little and then learning a little more. Additionally, students at this beginning stage of L2 acquisition tend to use markedly fewer words per sentence, and the words they user are shorter; these features are

SPA 1 - SPA 2	SPA 1 - SPA 3	SPA 2 - SPA 3	SPA 3 - SPA 21	SPA 22 - SPA 23	SPA 31 - SPA 32
aux_mood_Ind	lexical_density *	aux_tense_dist_Pres *	lexical_density	upos_PUNCT	upos_ADP *
aux_tense_Pres *	upos_ADP *	aux_mood_Ind	upos_DET	dep_punct	dep_case *
aux_tense_Imp *	upos_VERB *	aux_tense_Imp *	dep_punct	upos_ADV	verbal_head_sent
aux_tense_Past *	upos_NOUN *	aux_tense_Past	upos_VERB	dep_advmod	upos_PUNCT
upos_ADP *	upos_ADJ	dep_punct *	aux_tense_Pres	upos_CCONJ	upos_PRON
verbs_tense_Past *	upos_PRON	upos_PUNCT *	upos_ADJ	dep_cc *	dep_mark
upos_VERB *	dep_det	dep_nsubj *	upos_NOUN	upos_VERB	dep_punct
upos_INTJ *	upos_PUNCT *	dep_iobj	dep_nsubj *	dep_case	aux_tense_Imp
verbal_head_sent *	upos_PROPN	upos_PRON	upos_PRON	aux_form_Part	verbs_tense_Pres
verbs_tense_Imp *	dep_case *	verbal_head_sent *	upos_SCONJ	upos_ADP	subord_dist
upos_ADJ *	upos_SCONJ *	dep_cop	upos_ADV *	dep_mark	dep_cop
ttr_form	upos_AUX	subj_post *	upos_PUNCT	dep_compound	dep_cc
upos_PRON *	dep_punct *	aux_form_Fin	aux_form_Fin	upos_INTJ *	lexical_density
upos_PROPN *	subord_dist *	verbs_tense_Imp *	dep_cc *	dep_nsubj *	upos_AUX
upos_PUNCT *	upos_CCONJ *	upos_AUX	aux_tense_Imp	upos_AUX	upos_ADV

TABLE 6.6. Feature rankings obtained with sentence-level (*Same*) classification results for six different course level pairs. Features that vary in a statistically significant way with Wilcoxon Rank-Sum test are marked with *.

particularly salient for the classifier. Observing instead the results obtained pairing student essays belonging to the other three course level categories (Intermediate, Composition and Heritage), we notice a considerable drop in classifier performance. For instance, if we compare essays written by students in the Composition category (SPA 23 - SPA 24) we can see that all the classification results are below the majority class baseline classifier. A possible reason might be that these two courses are specifically aimed at improving learners' writing skills, with an emphasis on academic writing in Spanish, thus involving specific properties, such as discourse-level characteristics, which are possibly not covered by our set of features.

6.5. Understanding Linguistic Predictors

Beyond classification results, we were interested in understanding which typologies of linguistic phenomena are more important for solving the classification task and whether such properties correlate to the teaching curriculum. To better explore this second research question, we perform a feature ranking analysis along with the classification experiments, which allows us to establish a ranking of the most important features according to the different classification scenarios. That is, we evaluate the importance of each linguistic property by extracting and ranking the feature weights assigned by the LinearSVM. Table 6.6 reports the feature rankings obtained with sentence-level classification results, including pairs of essays that have the same prompt (*Same* configuration). We considered in particular six different course level pairs which are mostly representative of different stages of writing development. The focus on sentence-level results rather than document-level allows capturing more fine-grained linguistic phenomena.

Because the COWS-L2H corpus was collected from a single university with set curriculum, we are able to compare the features utilized by the LinearSVM with the course curriculum. We find that the feature rankings as obtained from the LinearSVM can in many cases be explained by differences in curriculum at each level. For example, from SPA 1 to SPA 2 the most important features used by the model are all related to verbal morphology, particularly morphology of auxiliary verbs. This can be explained by the fact that SPA 1 and 2 are the courses where students are introduced for the first time to the notions of verb tense and person. SPA 1 is focused on managing the idea of person and number in a tense that is not particularly difficult to understand for a speaker of English: the present tense. SPA 2, however, introduces the difficult difference between the three tenses in the past: imperfect, preterite and plus-perfect. This fact explains why distribution of past tense main verbs (*verbs_tense_Past*) differs between essays written during SPA 1 and SPA 2. Additionally, SPA 2 introduces composed verb tenses that require an auxiliary. Specifically, the auxiliary verbs "haber", "estar", and "ser" are introduced in SPA 2 as part of the past tense forms. Thus, it is not surprising that the top four features used by our classifier for distinguishing between essays written in SPA 1 and SPA 2 are related to the use of auxiliary verbs.

Classification of essays written by students while enrolled in SPA 2 and SPA 3 also relies largely on differences in verbal morphology. While the distribution of present tense auxiliary verbs is the most important distinguishing feature, other compound verb tenses play a role at these levels. For example, differences in the distribution of imperfect auxiliary verbs (*aux_tense_Imp*) may be explained by the use of the pluperfect tense.

Between SPA 1 and SPA 3, the most important discriminating feature is lexical density. While there is no specific focus on lexical density in the course curriculum, this feature is a natural extension of increasing sentence complexity. Davidson et al. [2019] shows that as students progress through the Spanish course sequence, lexical density tends to decrease due to the increased use of function words in more complex sentences. Additionally, one of the final items covered in the SPA 1 curriculum is the use of the prepositions "por" and "para". Also, at all three beginning levels students are taught to use prepositions in constructing more complex sentence structures. This may explain why preposition usage (*upos_ADP*) is a key discriminating feature between essays written in SPA 1 and SPA2, as well as between SPA 1 and SPA 3. The prominence of this feature indicates that students are learning to more confidently use prepositions as their writing skills develop. The fact that (*upos_ADP*) is not a key discriminating feature between SPA 2 and SPA3 indicates that these changes are occurring primarily at the SPA 2 level, which accords with the course curriculum.

In spite of the still reasonable accuracy in discriminating more advanced levels, making a direct connection between the features used by the SVM and the course curriculum becomes more difficult. At these more advanced levels students have developed an individual writing style which results in a more complex relationship between the curriculum and the syntax used by students. At the SPA 3 - SPA 21 interval, the only three features which vary in a statistically significant way are the distributions of nominal subjects (*dep_nsubj*), adverbs (*upos_ADV*), and coordinating conjunctions (*dep_cc*). While the increased use of adverbs may be seen as a general sign of increased writing complexity, coordinating conjunctions are taught explicitly during SPA 3. Conjunctions are also practiced intensively during both SPA 21 and SPA 22 explaining their importance as a discriminating feature between these levels.

One of the clearest connections between curriculum and the features used by the LinearSVM occurs at the Heritage levels SPA 31 and SPA 32. Heritage learners of Spanish raised in an English-dominant country are known to use "English-like" prepositions in Spanish. For example, Pascual y Cabo and Soler [2015] report on preposition stranding (which is grammatical in English but ungrammatical in Spanish) among Heritage speakers of Spanish in the United States. We find that distributional differences in the use of prepositions, represented by the features *upos_ADP* and *dep_case*, is the key distinguishing feature between essays written by the same student during SPA 31 and SPA 32. This difference indicates that students are learning to use prepositions in a more "Spanish-like" manner, which is one of the major areas of feedback which instructors provide to Heritage students.

6.6. Discussion

This chapter presents a first study aimed at modeling the evolution of written language competence in Spanish as a Second and Heritage Language, using data from the COWS-L2H Corpus. We have described a rich set of linguistic features automatically extracted from student writing, and have demonstrated that it is possible to automatically predict the relative order of two essays written by the same student at different course levels using these features, especially when considering students enrolled in beginner-level Spanish courses. Finally, we have shown that the linguistic features most important in predicting essay order often reflect the explicit instruction that students receive during each course.

This work can help instructors and language researchers better understand the specific linguistic factors which contribute to improved writing proficiency. Additionally, the appearance of features in the LinearSVM ranking helps clarify the effect of instruction on writing performance, specifically on effects such as the known delay between students being taught a concept and that concept appearing in the students' writing. We also believe that this work may contribute to the development of better language assessment and placement tools. Finally, the findings in this work contribute to an understanding of the longitudinal development of Spanish learner syntax, thereby informing the types of errors that an AWCF system for learners of Spanish should focus on a different proficiency levels.

CHAPTER 7

COWS-L2H Study III: Errors by demographic group

Having presented studies related to both lexical and syntactic development, I now present a study which examines variation among Heritage, L3, and L2 learners of Spanish using error rates calculated using annotated and corrected parallel text.

7.1. Error annotation

One of the goals of the COWS-L2H project is to annotate grammatical errors in the corpus in a way that writing patterns typical of Spanish as a foreign language produced by student participants can be identified, catalogued, and easily utilized by researchers who use the corpus. To that end, we have begun the process of error-tagging the corpus based on specific error types; the first two error types for which we have completed annotation are gender and number agreement, and usage of the Spanish *a personal*. We chose to annotate these specific error types based on research questions we wished to explore, but we intend to expand our error annotations in the future, as our annotation scheme can be readily adapted to additional error types we choose to annotate. Further, we encourage other researchers to adapt the annotation scheme to the annotation of other error types and contribute their work to the COWS-L2H project.

Our current team of annotators consists of four graduate-level Spanish instructors who have native or near-native fluency in Spanish. As previously mentioned, we are expanding our error annotation project through a collaboration with a Spanish university, which will allow us to significantly expand both the number of annotators and the scope of our error annotation project in the near future.

Our in-text error-tagging scheme is as follows:

[error]{edit}<annotation>.

Consider the example error in (1), and its annotation in (2):

Error type	α	$F_{0.5}$
Gender-Number	0.780	0.784
“a personal”	0.741	0.730
Average	0.761	0.757

TABLE 7.1. Inter-annotator agreement: error annotations

(1) Yo vivo en el ciudad.

“I live in the city.”

(2) Yo vivo en [el]{la}<ga:fm:art> ciudad.

In (2), the first set of brackets encloses the words in the error in question, the curly brackets that follow give the corrected edit, and the angle brackets house the error tags. In this case, the tags indicate that the error was a gender agreement error (ga), that masculine gender was erroneously produced in place of the correct feminine gender (fm), and that the error occurred on the article (art). A full description of the error annotation scheme is provided with the dataset in the corpus GitHub repository.

Each essay is annotated by at least two of our four annotators to ensure the accuracy of our annotations and the suitability of our annotation scheme. Due to the open-ended nature of the annotation task (any token can be considered a possible position of annotation), determining the best measurement for inter-annotator agreement is challenging. In Table 7.1, we report Krippendorf's α [Krippendorff, 2011] considering every token as an annotation position. Thus, if both annotators choose to not annotate a token, indicating that the token is correct, we treat this lack of annotation as agreement. This choice makes sense because, by not making an explicit annotation on a given token, the annotators are implicitly labeling the token as correct. An alternative method of calculating agreement would be to consider only positions where at least one annotator indicated an error; however, this choice would ignore all positions at which both annotators agreed that no error exists, which is itself a form of agreement. To put our agreement values in more familiar context, we also report the $F_{0.5}$ score, commonly used in GEC, using one annotator as ground-truth. In terms of both Krippendorf's α and $F_{0.5}$, our annotators show strong agreement.

7.2. Parallel corrected text

Unfortunately, manual error annotation, while accurate and reproducible, is an extremely time consuming process. We therefore, supplement our error annotation with parallel corrected text, from

Course Level	Essays	Tokens	Errors
Beginner	1,566	35,980	432,290
Intermediate	356	7,537	110,490
Composition	438	8,598	141,353
Heritage	337	5,487	110,313
Other	225	4,112	68,393
Total	2,922	61,714	862,839

Table 7.2. Summary of corrected essays in the updated corpus. Note that these numbers do not count both of the double-corrected essays, hence the discrepancy between the previously mentioned 70,397 sentence count and the numbers in this table.

which we are able to extract a much larger set of proposed errors based on instructor corrections of student essays. Currently, the compositions collected in this project are corrected by two doctoral student associate instructors of Spanish. Both have native or near-native command of Spanish, have previously taught the Spanish courses from which the students have been recruited to participate in this project, and thus are accustomed to recognizing, interpreting, and correcting errors made by students of L2 Spanish. Of course, one must note that the goal of instructors is to teach students to write a standardized dialect of Spanish, thus all instructor corrections may not be identifying actual "errors", but rather dialectical usage which diverges from the proscribed standard; this is especially true for Heritage learners whose prior knowledge of Spanish may not conform to the academic standard used in classroom instruction. Thus, throughout this section of the dissertation, the term "error" is used to refer to corrections made by instructors to students' submitted writing; whether or not these are all true errors in production is not investigated in detail.

To date, instructors have corrected approximately one-fifth of the essays in the COWS-L2H corpus, for 61,714 sentences (862,839 tokens) of corrected text. The distribution of corrected essays is shown in Table 7.2. Unlike the error annotations, which target specific errors, the corrections made to this set of essays are more holistic in the manner of an instructor correcting a student's work. The result of the correction process is a parallel version of the text, from which corrections can be extracted using NLP tools such as ERRANT [Bryant et al., 2017]. This aligned set of original and corrected sentences can be used for training NLP systems such as grammatical error correction. To our knowledge, our corpus represents the first parallel dataset of holistically corrected Spanish text available to researchers.

As with our error annotations, we are in the process of completing additional corrections and anonymization, and will make more data publicly available as soon as practical. As can be seen in Table 7.2, the largest portion of our currently annotated corpus comes from beginning students; completing additional corrections will allow us to present a larger number of errors from students at more advanced levels. Given the wide variety of ways a sentence can be corrected, our goal is to have each essay corrected by three individuals. Multiple corrections will increase error coverage in our training data and will provide additional test references for NLP researchers who are trying to build automated error identification and correction models.

7.3. ERRANT

To determine how the error patterns of students vary by level and L1, I conduct a detailed analysis of error patterns in the COWS-L2H corpus. An effective analysis of error distribution in language learner writing requires error annotations which include the location and type of each error. The manual annotation of errors is a time-consuming and labor-intensive process which requires extensive training on the annotation scheme to ensure consistent and accurate annotations. Currently, the COWS-L2H corpus contains manual error annotations for two specific error types, gender and number agreement and usage of the personal "a". However, because I wish to use error data from the corpus to conduct an analysis across a wide variety of grammatical and stylistic errors, I must conduct a broader evaluation of the errors contained in student writing.

Given that the COWS-L2H corpus contains 3,516 essays which have been fully corrected by graduate instructors of Spanish, I am able to automatically extract and tag a diverse set of errors for analysis. While this method tends to be less accurate than manual error annotation by skilled annotators, it allows for the rapid analysis of multiple types of errors. The corrections made by instructors include both strictly grammatical corrections, such as replacing an erroneously selected article with its correct counterpart, spelling and orthographic corrections, and stylistic corrections such as word choice. See Table 7.3 for examples of the types of errors identified and corrected by instructors in the COWS-L2H corpus, as well as the resulting automatic annotations.

To automatically annotate the errors identified by the instructors, I align sentences from the original and error-corrected subset of the COWS-L2H corpus to create a parallel sentence dataset containing approximately 12,000 sentence pairs. Because the corrections can include both

Original:	Stephen King escribió muchos libros .
Corrected:	Stephen King ha escrito muchos libros .
	A 2 2 ||| M:VERB:TENSE ||| ha ||| REQUIRED ||| -NONE- ||| 0 A 2 3 ||| R:VERB |||escrito ||| REQUIRED ||| -NONE- ||| 0

TABLE 7.3. Example of ERRANT automated annotation

splitting run-on sentences and merging sentence fragments, I include both individual sentences and concatenated consecutive sentences in the search space for sentence alignment. Additionally, due to sentence reordering and merger during correction, sentences in the parallel original and corrected essays cannot be aligned based on sentence order. Rather, I calculate the Levenshtein edit distance between each original sentence and each sentence and concatenated sentence pair in the corrected text. I then align the sentences with the lowest Levenshtein distance. While this method does not account for merging sentence fragments, I found that splitting of run-on sentences was far more common in the correction process. This method also does not account for the well-known issue of linguistically nonsensical word-level alignments which result from the Levenshtein algorithm [Xue and Hwa, 2014]. For example, as shown by Xue and Hwa [2014], because the Levenshtein algorithm seeks only to minimize the number of edits, it is likely to align words like "repair" and "reparations". However, the sentence-level alignment at this stage is meant only to identify which two sentences correspond to one another in the parallel texts; word-level alignment to extract specific errors is completed after sentence-level alignment.

Once aligned, extracting error corrections from sentence pairs is a matter of aligning words and identifying all edits made to transform the original sentence into its corrected version. This process is achieved using the ERRor ANnotation Toolkit (ERRANT) [Bryant et al., 2017], which locates, categorizes and annotates corrections in parallel original and corrected sentences. ERRANT uses a modified version of Damerau-Levenshtein distance [Damerau, 1964] developed by Felice et al. [2016] to align words in parallel texts. Specifically, Felice et al. [2016]'s word alignment method seeks to introduce linguistic information into the alignment algorithm by creating a substitution cost function which considers differences in lemma form and part-of-speech, in addition to the character-level differences used by the original Damerau-Levenshtein algorithm. Substitutions in which the aligned words share the same lemma form and/or part of speech (as determined by SpaCy [Honnibal and Montani, 2017]) cost less than do linguistically unrelated substitutions. As a result, words which have similar spelling and linguistic function are more likely to be aligned. Felice et al. [2016] argues that the resulting alignments are more natural and human-like than alignments generated by simple character-level alignment used by the Damerau-Levenshtein algorithm.

Once word-level alignment is completed, errors can be readily identified by comparing the differences between the original and corrected versions of the parallel text. ERRANT uses a set of approximately fifty ordered rules to classify each identified error into one of three operations and one of seventeen general error classes based on the dependency label and part of speech of both the original and corrected word form. The full errant tagset, including the seventeen basic categories and subdivisions thereof, is shown in Figure 7.1. Words removed during correction are tagged with the operation label "U" for "unnecessary". For those words inserted during correction, the operation label is "M" for "missing", while the operation label for replacements is "R". Most errors identified by ERRANT correspond to part of speech tags, such as Noun and Verb. However, the system also includes specific tags for word order, morphological, spelling, and orthographic errors. For example, two words with the same lemma are tagged as morphological variants ("MORPH"). Aligned words which are not identical but which have at least half of their characters in common are tagged as "SPELL", while words which differ only in capitalization are tagged "ORTH". Additionally, two adjacent words whose order is swapped by correction are tagged "WO" for "word order". See Table 7.3 for examples of these error types. A detailed explanation of the ERRANT system's rule-based error tagging method can be found in Bryant et al. [2017].

Although ERRANT was originally designed for error analysis of English texts, I have modified the system to process and tag Spanish parallel texts. To achieve this modification, I set the SpaCy library settings used by ERRANT to the "es_core_news_sm" Spanish model, allowing the system to utilize SpaCy's Spanish dependency parser and part-of-speech tagger. Because ERRANT's rules are largely based on part of speech and dependency tag analysis, many of the required changes can be achieved simply by switching to an effective Spanish dependency and POS tagger. I also provide the system with a Spanish word list generated from the Spanish HunSpell dictionary so that it may identify spelling errors in Spanish text. Finally, I remove and replace English-specific error-tagging rules, such as the use of "to" to indicate infinitive verbs. My modified ERRANT code is available on GitHub[6]. It should be noted that the Spanish version of Errant is not as fine-grained

[6]https://github.com/ucdaviscl/cowsl2h

Code	Meaning	Description / Example
ADJ	Adjective	*big → wide*
ADJ:FORM	Adjective Form	Comparative or superlative adjective errors. *goodest → best, bigger → biggest, more easy → easier*
ADV	Adverb	*speedily → quickly*
CONJ	Conjunction	*and → but*
CONTR	Contraction	*n't → not*
DET	Determiner	*the → a*
MORPH	Morphology	Tokens have the same lemma but nothing else in common. *quick (adj) → quickly (adv)*
NOUN	Noun	*person → people*
NOUN:INFL	Noun Inflection	Count-mass noun errors. *informations → information*
NOUN:NUM	Noun Number	*cat → cats*
NOUN:POSS	Noun Possessive	*friends → friend's*
ORTH	Orthography	Case and/or whitespace errors. *Bestfriend → best friend*
OTHER	Other	Errors that do not fall into any other category (e.g. paraphrasing). *at his best → well, job → professional*
PART	Particle	*(look) in → (look) at*
PREP	Preposition	*of → at*
PRON	Pronoun	*ours → ourselves*
PUNCT	Punctuation	*! → .*
SPELL	Spelling	*genectic → genetic, color → colour*
UNK	Unknown	The annotator detected an error but was unable to correct it.
VERB	Verb	*ambulate → walk*
VERB:FORM	Verb Form	Infinitives (with or without "to"), gerunds (-ing) and participles. *to eat → eating, dancing → danced*
VERB:INFL	Verb Inflection	Misapplication of tense morphology. *getted → got, fliped → flipped*
VERB:SVA	Subject-Verb Agreement	*(He) have → (He) has*
VERB:TENSE	Verb Tense	Includes inflectional and periphrastic tense, modal verbs and passivization. *eats → ate, eats → has eaten, eats → can eat, eats → was eaten*
WO	Word Order	*only can → can only*

FIGURE 7.1. ERRANT tagset, including both basic part-of-speech categories and subdivisions thereof. Taken from Bryant et al. [2017].

in its classification of verb errors due to need to write many additional rules to account for the relatively complex verbal morphology of Spanish; rather, the system classifies the large majority of verbal morphology errors simply as "MORPH" indicating that the two tokens in question share the same lemma but different surface forms. Where possible, the system is more specific, for example, it correctly identifies substitutions of a verb with an infinitive as a "VERB:FORM" error. For the purposes of this study, I condense all related verb errors into a single "VERB" category; future work in developing more fine-grained automated analysis of Spanish verbal morphology is warranted to improve future iterations of the proposed method and to obtain a more detailed understanding of the development of verbal morphology in Spanish learner writing.

7.4. Method and implementation

Given the amount of data and demographic information collected from various students sub-populations in the COWS-L2H corpus, it is possible to conduct a fine-grained analysis of student errors across a wide-range of variables. However, in order to limit the scope of this study, I investigate the error rates across two paired groups of students.

The first pair consists of Heritage learners and advanced English-dominant L2 learners of Spanish. Various studies (Montrul [2005], Lynch [2008], Montrul [2012]) argue that, in many aspects, Heritage learners are more similar intermediate-to-advanced L2 learners than they are to native-dominant speakers of the target language. For example, past work has observed non-native-like performance among Heritage learners in inflectional morphology (such as verb and gender agreement) [Polinsky, 2008a, Montrul, 2012], determiner usage [Montrul and Ionin, 2012], and lexical access [Hulsen et al., 2002]. Therefore, I wish to investigate error rate variation between these two sub-populations of students.

The second pair consists of lower-level English-dominant L2 learners and Mandarin-dominant L3 learners of Spanish. As previously mentioned, L3 learners of Spanish in the context of American universities are, by necessity, also L2 learners of English. As all such students can be assumed to have passed the TOEFL exam prior to admission at an American university, their level of English proficiency can be assumed to be rather high. This raises important questions about how these students linguistic background affects their acquisition of Spanish, if at all, compared to the acquisition patters of English-dominant L2 learners. Several previous studies have argued that the learning of an L2, especially one which is typologically more similar to the target L3 than the learner's L1, effectively blocks syntactic and lexical transfer in L3 acquisition [Rothman, 2011, Foote, 2009]. If one were to accept Rothman's "typological primacy model", one would expect to find little difference between the error rates of L2 and L3 learners of Spanish. To further investigate this hypothesis, I conduct a statistical analysis of the error rates, as extracted from Errant, of English-dominant L2 learners of Spanish and their Mandarin-dominant L3 leaner peers.

In order to identify meaningful differences in error rates across student populations, I first extract all errors (as identified by instructor corrections) from the entire sub-corpus of parallel corrected data using the modified ERRANT tool. I then iterate through the identified errors in each essay and group these errors by both part of speech and operation type (for example, determiner replacements) to obtain a set of error counts for each student essay. I simultaneously track the count of each part of speech and the overall token count for later use in calculating error rates. Once these counts are obtained, I group the essays by L1 and course level so I can segregate the target groups for analysis. Finally, calculate the mean and standard deviation for each L1-course grouping. For example, I identify all students who are L1 speakers of English enrolled in Spanish 1, and calculate the mean and standard deviation for this group of students across 14 of the error classes used by ERRANT, with three classes omitted due to low error counts.

7.5. Statistical analysis

For my introductory level English and Mandarin students, I aggregate the students based on demographic data indicates that indicates that they are enrolled in either SPA1, SPA2, or SPA3, and who report either English or Mandarin as their primary language. I also include students who report L1 proficiency in English and another language, such as Hindi, in the English group.

Similarly, I aggregate all students enrolled in either SPA21, SPA22, SPA23, or SPA24 and who report English as one of their L1s into the English group (there are a handful of students enrolled in these courses who report both English and Spanish as co-dominant L1s; I exclude these students from my analysis as their experience with Spanish would likely classify them as Heritage learners). Finally, I group all students enrolled in SPA31, SPA32, and SPA33 into the Heritage learner category.

Because the student groups I am comparing are independent (there is no overlap between the members of the populations), I am able to conduct a straightforward analysis for statistical significance using the two-sided t-test for each of the error categories, and for each of the sub-population pairings. In each case, the research hypothesis is that there is a difference between the error rate of the two populations, while the null hypothesis is simply that there is no difference in error rate between the target populations. I implement the statistical tests in Python using the SciPy package. Prior to running the t-tests, I remove outliers from the data using the "IQR method", wherein any values which fall 1.5 times the interquartile range above the third quartile or below the first quartile (that is, 2.7 standard deviations above or below the mean) are removed from the dataset. Because Student's t-test [Student, 1908] assumes homogeneity of variance, I must first use Levene's test [Levene, 1961] to ensure that the variance between my two populations for

Category	1-3 English		1-3 Mandarin		p-val	21-24 English		Heritage		p-val
	mean	std	mean	std		mean	std	mean	std	
ADJ	0.078	0.078	0.068	0.078	0.235	0.098	0.072	0.077	0.084	0.062
ADP	0.195	0.135	0.147	0.102	<0.001	0.183	0.112	0.107	0.074	<0.001
ADV	0.035	0.057	-	-	-	0.085	0.093	0.027	0.045	<0.001
AUX	0.104	0.111	0.035	0.049	<0.001	0.108	0.086	0.083	0.100	0.049
CCONJ	0.054	0.066	0.067	0.094	0.462	0.057	0.061	0.063	0.077	0.553
DET	0.170	0.125	0.139	0.100	0.004	0.175	0.126	0.100	0.073	<0.001
NOUN	0.155	0.105	0.119	0.066	<0.001	0.114	0.077	0.099	0.073	0.131
ORTH	0.012	0.011	0.008	0.008	<0.001	0.007	0.006	0.009	0.007	0.015
OTHER	0.020	0.015	0.015	0.010	<0.001	0.016	0.010	0.017	0.011	0.475
PRON	0.163	0.151	0.131	0.137	0.035	0.219	0.155	0.089	0.083	<0.001
SCONJ	0.078	0.122	0.018	0.047	<0.001	0.060	0.075	0.076	0.101	0.228
SPELL	0.012	0.014	0.002	0.003	<0.001	0.003	0.004	0.003	0.004	0.222
VERB	0.162	0.109	0.163	0.125	0.970	0.210	0.122	0.097	0.068	<0.001

TABLE 7.4. Mean error rate, standard deviation, and p-value for comparison between 1) lower-proficiency English- and Mandarin-dominant students, and 2) Intermediate-to-advanced proficiency English-dominant and Heritage students. Numerator is count of each error type. Denominator is token count for ORTH, OTHER, and SPELL; otherwise, count of target POS. NOTE: ADV figures are not included for Mandarin students due to insufficient data.

the target error type is homogeneous. If the p-value for Levene's test shows no significant difference between the variance of the two populations (that is, a p-value greater than or equal to 0.05) I use Student's t-test; otherwise, I use Welch's t-test [Welch, 1947], a generalization of Student's t-test which does not assume homogeneity of variance. I report the mean, standard deviation, and p-value for each error category across my two comparison groups in Table 7.4.

7.6. Key differences and findings

At this stage, I am not seeking to draw any specific conclusions regarding the differences observed between my defined learner groups. Rather, I conduct an exploratory analysis to understand what differences in error rates between these groups may warrant further investigation using both more fine-grained classification and more powerful statistical analysis techniques. To that end, I describe below my preliminary findings and discuss next steps in clarifying the extent of and possible reasons for the differences observed.

7.7. Lower proficiency learners

As previously discussed, the two largest populations represented in introductory level (that is, SPA 1-3) of the COWS-L2H corpus are English-dominant L2 learners and Mandarin-dominant L3 learners of Spanish. As shown in Table 7.4, the error rates seen in texts written by these two groups are similar in many respects. For example, the two groups are nearly identical in their error rates related to verb choice and morphology. This is an interesting finding given that both Spanish and Mandarin have verbal systems which clearly mark aspect in a manner that is more complex than English. Thus, I would expect Mandarin students to more readily acquire Spanish verbal morphology; however, this does not appear to be the case, at least from the error rate data presented above. Given the fact that the the present analysis is based on data from students enrolled in courses designed for lower-proficiency students, it is possible that neither the English nor the Mandarin students are using more complex forms of verbal aspect marking at this early stage of Spanish acquisition. Additionally, because of English's relatively simply verbal morphology, the Mandarin-dominant students in the corpus, who are all proficient in English, are receiving no transfer benefit in acquiring Spanish verbal morphology from their knowledge of English. Conversely, L2 English-dominant learners appear to make more errors in noun choice and morphology than do L3 Mandarin-dominant students, despite the fact that neither language uses gender agreement on nouns, which accounts for a large portion of the observed errors. Additionally, English modifies nouns for number agreement, like Spanish, but unlike Mandarin; so if typological similarity of their dominant language were the only factor involved in determining the difficulty for students in acquiring the target form, English-dominant students should have an advantage over Mandarin-dominant learners. This discrepancy highlights the possibility of transfer from the Mandarin-dominant students' L2, English, as hypothesized by several of the transfer models discussed in Section 2.1.3.

One of the most salient error rate distinctions between my two lower-proficiency groups is in spelling and orthographic errors. ERRANT defines orthographic errors as simply errors in whitespace usage and capitalization; I expand this definition slightly in my ERRANT-SP implementation to include errors in use of accent marks (which are quite common in learner text). Spelling errors are identified as those aligned tokens which share at least half of their characters, and which do not fall into another category (such *NOUN* for cases of noun inflection). As can be seen in Table 7.4, Mandarin-dominant students enrolled in SPA 1-3 make far fewer spelling and orthographic errors than do their English-dominant peers. Without much more investigation, I cannot identify the underlying cause for this discrepancy, but it may arise from pedagogical differences in writing instruction in China and the United States. For example, Mohan and Lo [1985] reports that his Chinese-speaking students (most of whom are from Hong Kong and therefore Cantonese speakers) indicate that their previous study of writing in English focused on accuracy at the sentence level, rather than larger organizational features of their writing. A more directly applicable study is Hsiang and Graham [2016], who report on a large-scale survey of writing teachers in Chinese primary schools. These surveys indicate that, on the whole, Chinese teachers use a "product-based model" of writing instruction, which focuses on "correct production, with relatively little emphasis on critical writing processes such as planning and revising" [Hsiang and Graham, 2016]. On the other hand, surveys of teachers in the United States indicate that more time and attention is paid to the development of writing planning and revision skills, with relatively less focus on accuracy at word and sentence level [Gilbert and Graham, 2010, Graham, 2019]. Of course, these are broad statements about trends which certainly vary depending on both regional, school, and individual instructor practices. However, it is worth noting that, in all measures in which a statistically significant difference was found between English- and Mandarin- dominant students in SPA 1-3, the English-dominant learners show a higher error rate than their Mandarin-dominant counterparts.

Another interesting comparison which warrants further investigation is the difference in error rates between Mandarin- and English- dominant students in the production of determiners. Unlike English, which has the definite article "the" and indefinite article "a/an", there is no definite or indefinite article in Mandarin [Li and Thompson, 1989]. According to White [2008] "It has long been observed that L2ers have problems with article acquisition, particularly if the L1 lacks articles." Thus, in the absence of transfer from L2 English, one would expect Mandarin-dominant students to produce determiner errors at a higher rate than English-dominant students. However, in the data examined for this study, I observe the opposite: Mandarin-dominant students make fewer determiner errors than their English-dominant peers. This observation indicates that, in the context of L3 acquisition, L1 transfer may not be as large a factor as some research, such as Schwartz and Sprouse [1996], has previously argued. Thus, our findings do not support the simple L1 transfer hypothesis, which states that language learners transfer the syntactic structures of their L1 to all subsequently acquired languages. The fact that native English speakers do not perform markedly better than native Mandarin speakers in measures of grammatical determiner usage (which includes gender and number agreement of articles) indicates that the Mandarin speakers, whose language does not use articles and differs significantly in its number agreement system from both English and Spanish, are not transferring these aspects of their L1 when learning Spanish as an L3. These findings could support several of the transfer hypothesis discussed in Section 2.1.3: L2 transfer [Williams and Hammarberg, 1998], the Cumulative Enhancement Model [Flynn et al., 2004], or the Typological Primacy model [Rothman, 2011], among others. Further research is warranted to determine which model best explains the data in COWS-L2H and other Spanish learner corpora, and to garner a more complete understanding of the impact of L2 transfer in the acquisition of L3 Spanish.

7.8. Advanced L2 and Heritage learners

According to Montrul [2012], "identifying how L2 learners and heritage speakers differ in their linguistic competence and processing abilities is a critical step towards developing efficient pedagogical strategies in language teaching." Previous research [Lynch, 2008, Rothman, 2009, Silva-Corvalán, 1994] has pointed out that the language proficiency of Heritage learners is not comparable to that of speakers living in a country where the target language is dominant. Further, Montrul [2012] states that Heritage learners are "'interrupted' native speakers who retain a great deal of native abilities but whose competence in the heritage language is comparable to the linguistic abilities achieved by adult second language learners." Additional studies of the language competence of L2 and Heritage learners have found that while Heritage learners show more native-like performance in their phonetics and phonology, Heritage and L2 learners perform similarly in measures of morphosyntactic development [Au et al., 2002, Knightly et al., 2003]. However, the present study found that the error rates of Heritage learners are significantly lower than those of advanced L2 learners in several grammatical categories including pronoun, preposition, and determiner usage, as well as verb choice and morphology errors. While the present study does not attempt to explain or account for these differences in detail, the marked differences observed between advanced L2 students and the Heritage learners indicates that the relationship between these two groups may be more subtle than previously thought, and that further research is needed to understand the language competence of Heritage learners, and how their abilities compare with those of both L2 and target-dominant individuals.

For example, Montrul [2002] states that Heritage learners, whose experience with Spanish is likely limited to speaking, produce more errors in written elicitation tasks relative to L2 learners. Given the fact that the elicitation tasks in COWS-L2H are all written, we would expect to see higher error rates among our Heritage population relative to the advanced L2 learners; however, we find the opposite for all but one of the categories we investigated (orthographic errors), as shown in Table 7.4. That said, one must also consider the fact that the Heritage learners in the Heritage course series of COWS-L2H likely represent more proficient Heritage learners, as discussed in Section 2.1.1. Thus, the statements of researchers like Montrul may not be entirely applicable to the group represented in the present study.

The discrepancy in errors on verbs, which includes both verb choice and verbal morphology errors, is a clear distinction between our Heritage learners and the advanced L2 population. As can be seen in Table 7.4 and Table 7.5, advanced L2 learners make roughly twice the number of verb errors as to their Heritage learner counterparts. This finding does not fully accord with reports in previous work, such as Montrul [2002] and Potowski et al. [2009], who both report that Heritage learners show similar distribution of usage of specific verbal morphology (in the case of Montrul [2002], preterit-imperfect distinctions), but that Heritage learners produce more errors. Additional investigation of specific aspects of verbal morphology is necessary to better understand which aspects of verbal morphology are more sensitive to attrition and incomplete acquisition in Heritage learners.

Additional observations include the fact that Heritage learners produce fewer errors in the production of determiners, prepositions and pronouns. These findings again do not fully accord with previous studies comparing Heritage and L2 learners. For example, Montrul [2011] found that L2 learners significantly outperformed Heritage learners in a written task which required participants to select the correctly inflected determiner form preceding a noun. However, the data related to pronouns is not unexpected given previous research; Montrul [2010b] found on a task to correctly assess the use of clitic pronouns that Heritage learners performance approximates that of native speakers in several sub-categories. One possible difference between the previous studies and the present research is the fact that most previous work has been based on acceptability judgment tasks, short elicitation tasks, and comprehension tasks, which may not adequately simulate natural language production. The present study, on the other hand, uses data drawn from a corpus of relatively long (250+ token) written productions. Further investigation is needed to both confirm the preliminary observations presented here, as well as to further sub-categorize the broad part-of-speech classes used in the present study. For example, are Heritage learners producing fewer errors than L2 learners on all types of pronoun or all types of determiners, or are specific types less error prone in Heritage learners?

An additional interesting feature of this data is that the observed differences between advanced L2 learners and Heritage learners appear stable across the Heritage series. That is, if we compare the most advanced learners (those in Spanish 24 and upper division courses) with either the beginning Heritage course (Spanish 31) or the advanced Heritage course (Spanish 33), the results do not change drastically. This comparison is detailed in Table 7.5. Thus, it appears that Heritage learners begin their study of Heritage Spanish with specific advantages over their L2 peers, rather than developing these advantages over the course of their Heritage classroom experience. This finding again casts doubt on much previous research which has highlighted the similarity between L2 and Heritage learners. Of course, one must consider the possibility that the students who choose to enroll in the Spanish 31-33 series are the more proficient among the Heritage learner population, with less proficient students enrolling in the non-Heritage Spanish series. If the larger Heritage learners population is self-segregating based on proficiency, one could expect those higher proficiency students who choose to enroll in Spanish 31-33 to show more native-like performance than their advanced L2 peers. Additional investigation of student demographics is needed to determine how the Heritage learner population is distributed across the various course offerings represented in the COWS-L2H data.

7.9. Discussion and implications

One consideration which should be addressed is differences in the correction methods used by the graduate student instructors who corrected the essays in COWS-L2H from which the presented errors rates are extracted. Clearly, "correcting" an essay is, at least partially, a subjective task, and it is likely that two instructors would correct the same essay in different ways. A necessary next step in the analysis of the data in COWS-L2H is to extract error rates by corrector, to ensure that no significant differences in error rate can be attributed to corrector bias. The same issue arises when attempting to make comparisons across different parallel corrected corpora. For example, in an examination of the CATE corpus [Lu, 2010], I found that error rates were, in general, far higher

Category	24+ English		31 Heritage		p-val	33 Heritage		p-val
	mean	std	mean	std		mean	std	
ADJ	0.160	0.092	0.084	0.086	<0.001	0.074	0.083	<0.001
ADP	0.251	0.042	0.154	0.111	<0.001	0.183	0.112	0.107
ADV	0.075	0.081	0.041	0.053	0.043	0.020	0.039	0.001
AUX	0.129	0.090	0.059	0.074	<0.001	0.084	0.099	0.024
CCONJ	0.074	0.044	0.055	0.065	0.165	0.075	0.093	0.933
DET	0.195	0.090	0.104	0.066	<0.001	0.098	0.075	<0.001
NOUN	0.170	0.092	0.113	0.096	0.007	0.097	0.069	<0.001
ORTH	0.012	0.009	0.011	0.006	0.513	0.009	0.007	0.012
OTHER	0.018	0.010	0.022	0.018	0.203	0.016	0.011	0.362
PRON	0.268	0.075	0.165	0.160	<0.001	0.070	0.060	<0.001
SCONJ	0.067	0.069	0.090	0.097	0.277	0.069	0.102	0.921
SPELL	0.001	0.001	0.004	0.004	<0.001	0.003	0.004	<0.001
VERB	0.231	0.109	0.130	0.083	<0.001	0.091	0.071	<0.001

Table 7.5. Mean error rate, standard deviation, and p-value for comparison between 1) the study's most advanced English-dominant L2 learners (Spanish 24 and upper division), 2) Introductory Heritage learners (Spanish 31), and 3) Advanced Heritage learners (Spanish 33). Numerator is count of each error type. Denominator is token count for ORTH, OTHER, and SPELL; otherwise, count of target POS.

than those seen in even the lowest-proficiency students in COWS-L2H. In CATE, the mean error rate on determiners was 35.6%, and the mean rate of verb errors was 40.7%. Thus, it appears that the correction methods used in CATE are more expansive, though a more detailed analysis of the CATE corpus is needed to confirm this observation.

As presented in Table 7.4 and Table 7.5, the student sub-populations represented in the COWS-L2H make errors at different rates beyond those expected based on proficiency alone, indicating potential effects of language transfer, writing methodology, and Heritage language experience. As mentioned previously, the present study does not seek to draw specific conclusions about the linguistic competence of one group relative to another. Rather, I present these findings in an effort to understand where potential differences between learner sub-populations may exist which warrant further investigation. Although preliminary, this study represents one of few large-scale corpus-based investigations of discrepancies between Spanish learner groups. Additionally, if these preliminary findings can be confirmed and expanded upon further analysis, they will contribute substantially to the understanding of L1 and L2 transfer in L2 and L3 learners, and to language attrition and (re)acquisition in Heritage learners.

CHAPTER 8

GEC Approach & Implementation

As previously discussed, one of the key goals of the present dissertation is the development and implementation of an AI-powered automated corrective feedback system for learners of Spanish that is adaptable to learner proficiency, first language, and other individual learner attributes, and that is able to deliver feedback in a manner that is supported by research in second language pedagogy. The implementation of this system requires the development and testing of several components that will be the topic of the following chapters, namely:

(1) A method for creation of realistic synthetic training data to supplement the real learner data from the COWS-L2H corpus.
(2) Grammatical error correction (GEC) models trained on this data to identify and correct errors in learner text.
(3) A system to classify identified and corrected errors into meaningful categories so that appropriate feedback can be provided to students.
(4) Templates and generative language models to craft feedback to be presented to students.
(5) An application that pulls these components together so that students can input their writing and receive automatically generated feedback in a near-real-time manner.

The studies presented in the preceding chapters demonstrate the large amount of variation among L2 learners of Spanish with regard to lexical development, syntax acquisition, and frequency and types of errors made. Given the demonstrated variation between student sub-populations, the foregoing studies help motivate my hypothesis that an error correction and feedback model that is tuned to specific learner attributes will be more effective in both correctly identifying student grammatical and stylistic errors, and that such a model will facilitate presenting those errors in ways that are most beneficial to individual students. The utility of learner data to train automated feedback models is well established in the literature on GEC [Napoles and Callison-Burch, 2017, Leacock et al., 2010, Zhao et al., 2019, Chollampatt and Ng, 2018, Junczys-Dowmunt et al., 2018].

Additionally, the efficacy of augmenting sparse learner data with artificially generated parallel error data has been clearly demonstrated [Junczys-Dowmunt et al., 2018, Grundkiewicz and Junczys-Dowmunt, 2019, Stahlberg and Kumar, 2021]. But, in order to augment data in a way that more accurately reflects the distributions of errors, syntactic structures, and lexical items seen in learner data, we must first identify those structures and patterns. For example, as the error analysis presented in Chapter 7 indicates, Heritage learners are less likely than their L2 learner peers to make errors in production of pronouns; therefore, the error correction model should predict pronoun errors less frequently for Heritage learners to more accurately reflect the distribution seen in student writing and thereby reduce model error. To facilitate a model that will be able to effectively make such predictions, I explore methods of generating synthetic training data that replicates error rates observed in different student groups.

All models tested in the following experiments are evaluated using precision, recall, and the $F_{0.5}$ metric. Precision is the number of total model predictions that were correctly predicted. Recall is the number of total members of the target class that were predicted by the model, and $F_{0.5}$ is the modified harmonic mean of precision and recall, which gives more weight to precision than to recall. $F_{0.5}$ is the standard evaluation metric used in GEC, as many researchers argue that reducing false positives is more important than avoiding false negatives in the GEC task. That is, it is more important that corrections presented to students be accurate rather than that all possible corrections are caught by the model. Although previously discussed in Chapter 3, I provide the definition of $F_{0.5}$ in Formula 8.1 for context.

$$F_{0.5} = ((1 + 0.5^2) * Precision * Recall)/(0.5^2 * Precision + Recall) \tag{8.1}$$

8.1. Initial tests of error correction using COWS-L2H

Before I begin exploring methods of synthetic data augmentation for Spanish GEC, I first must establish the baseline performance of models trained only on real learner data from COWS-L2H. To demonstrate the efficacy of the current version of the COWS-L2H corpus in the GEC domain, and to further demonstrate the efficacy of using error corrected learner data in training GEC models, I trained a baseline GEC model by fine-tuning mT5 [Xue et al., 2021], a multilingual text-to-text transformer model, on parallel corrected corpus data from COWS-L2H. I also implement a BiLSTM

GEC model as described in Davidson et al. [2020] as a baseline for comparison to previous Spanish GEC projects. All of the above models cast GEC as a monolingual translation problem, an approach which has proven fruitful in GEC applications (for example, Napoles and Callison-Burch [2017]). For these baseline models, we do not augment our training data with synthetically generated error data.

All experiments, with the exception of the Llama2 model described below, are run using sentence-level correction; that is, each sentence from a source essay is paired with its counterpart from the corrected essay, giving an input/output sentence training example. I choose to correct learner writing at the sentence level rather than training a model to correct the entire essay at once in order to make the real learner data compatible with the source data used for synthetic data generation. Source data from Cañete et al. [2020] is provided as sentences rather than longer texts, so I need my models to be trained to handle single sentence input and output. To preprocess the parallel original and corrected essay pairs provided in COWS-L2H into suitable training data for a sentence correction model, I first align sentences in the original and corrected essay texts. Essays are first split into sentences using NLTK [Loper and Bird, 2002] and then aligned to create parallel-corrected sentence pairs for training. Given the fact that the correction process may result in the removal or reordering of sentences, I must use string matching to ensure that sentences are correctly aligned. I use The Fuzz[7] string matching package in Python and align each sentence in the original essay with the most similar sentence in its corrected counterpart essay. We train and evaluate our models using a 70/15/15 train/test/validation split. I implement the mT5 model in Python using Huggingface Transfomers[8] and PyTorch Lightning[9], using the 1.2 billion parameter *mt5-large* variant of the model. During fine-tuning we use a batch size of 16 and a maximum sequence length of 64, and fine-tune for 2 epochs.

In order to compare to results reported in previous Spanish GEC work [Yadav, 2022, Davidson et al., 2020], I train the BiLSTM and fine-tune the mT5 models on a previously released version of COWS-L2H, which contains 12,678 parallel corrected sentence pairs, as well as on the updated, larger version of corpus that contains approximately 70,397 sentence pairs (giving a 70% training set of approximately 49,000 sentences). As expected, I find that the increased size of the updated version

[7]https://github.com/seatgeek/thefuzz
[8]https://huggingface.co/docs/transformers
[9]https://lightning.ai/docs/pytorch

Model	Recall	Precision	$F_{0.5}$
Previous corpus w/o data augmentation			
BiLSTM	0.094	0.139	0.101
mT5-large	0.30	0.102	0.216
Updated corpus w/o data augmentation			
BiLSTM	0.254	0.153	0.224
mT5-large	**0.619**	**0.326**	**0.525**

TABLE 8.1. GEC results demonstrating the marked improvement in model performance when trained using the updated version of the corpus.

of the COWS-L2H corpus markedly improves Spanish GEC performance, even with no synthetic data augmentation. Training mT5-large on our current 49,277 sentence pair training set **more than doubles** the $F_{0.5}$ score as compared to training the same model on our previously released set of 12,678 sentence pairs. Our results when training on the previously released version of our dataset are comparable to those achieved by Davidson et al. [2020] and Yadav [2022], who used a BiLSTM and mT5, respectively. Results are shown in Table 8.1. These results clearly demonstrate the utility of the parallel original-corrected essays pairs provided in COWS-L2H for the GEC domain, and show that the recently released expanded COWS-L2H greatly increases potential value of the corpus to GEC researchers. These results also serve as a useful starting point for the further research on training GEC models using data from COWS-L2H that will follow in this chapter.

8.2. Synthetic data generation

In an effort to further improve the GEC results reported above, I now explore the use of additional synthetically generated parallel error data for initial finetuning of a GEC model for Spanish learners. Much work in GEC in recent years has focused on developing methods to generate synthetic training data to augment the learner writing used for training large neural GEC models. For example, Zhao et al. [2019] uses random insertion and shuffling operations to add noise to error-free text, while Grundkiewicz et al. [2019] uses spell-checkers to create confusion sets of similar words to corrupt source data. More recently, Ye et al. [2023] propose a method of generating synthetic data without the need of a source monolingual corpus. The synthetic datasets proposed in these works are used to pretrain neural encoder-decoder models (modeled after those used for machine translation), that are then further fine-tuned on real annotated learner data. Recently, much work in data augmentation for GEC has focused on developing methods to replicate the distribution of learner errors in synthetic training data, resulting in improved end-model performance [Takahashi et al., 2020, Lichtarge et al., 2020, Stahlberg and Kumar, 2021]. The data augmentation method I propose for my dissertation draws heavily from that of Stahlberg and Kumar [2021], who proposes the use of sentence corruption models that condition the generation of an "errorful" sentence on a "correct" version of the sentence along with one or more error tags. Because the error tags are explicitly passed to the generation model, replicating an error distribution seen in real learner data becomes relatively straightforward. In their work, Stahlberg and Kumar [2021] demonstrate that the use of synthetic training data that replicates the error distribution extracted from a large corpus of learner text significantly improves model performance. Further, Stahlberg and Kumar [2021] propose a method of adapting GEC models to writer proficiency by modifying the error distribution used to create their synthetic training data. While Stahlberg and Kumar [2021] do not find this method effective for building models adapted to three different CEFR levels, their work is limited to English learner data from the BEA 2019 corpus [Bryant et al., 2019] that was collected from numerous educational settings and not controlled for topic. Further, Stahlberg and Kumar [2021] does not account for learner L1 or language education experience when developing their CEFR-level targeted models. Building on their work, I developed a data augmentation technique that as-closely-as-possible replicates the error distribution seen in Spanish learner data from the COWS-L2H corpus across various levels of proficiency, L1, and linguistic experience.

In my approach, the generation of synthetic data for model training begins with collecting a large quantity of error-free text that can be corrupted to provide additional training data without the need for costly and time-consuming human annotation. However, identifying and corrupting error-free text from a large Spanish corpus, such as the Corpus del Español [Davies, 2016], is not sufficient to create data that "looks like" learner data. While the method proposed by Stahlberg and Kumar [2021] seeks to replicate the error distribution seen in real-world data when generating synthetic training data, they make no effort to ensure that the underlying sentence structure and lexicon is similar to that of language learner writing. To put this in simpler terms, inserting errors into *Moby Dick* will not result in data that resembles learner writing, even if the distribution of inserted errors is roughly the same. I hypothesize that, like the benefits seen from replicating error rates, finding data with similar lexical density and diversity, as well with similar indicators of syntactic complexity such as parse depth, will result in synthetic training data that more closely resembles real learner data, thus resulting in better performing GEC models. Alternatively, given recent improvements in the field of machine translation, automated translation of learner data written by students learning other languages (for example, learners of English) to Spanish can potentially be used a source of data that replicates the syntactic complexity and lexical diversity of Spanish learner writing.

To generate the necessary synthetic training data, I collected a set of well-formed sentences that I then corrupted to create my synthetic training data for a given proficiency and L1 pair. I insert errors into the data using the method described in Stahlberg and Kumar [2021]. Specifically, I fine-tuned an mT5 encoder-decoder model [Stahlberg and Kumar, 2020] conditioned on an input sentence and an error tag; the fine-tuned model takes as input a well-formed sentence and a target error tag, and outputs a rewritten version of the same sentence that includes an error of the target type.

8.2.1. Data sources. While there is a lot of Spanish text available online, such as texts from Project Gutenberg and Spanish Wikipedia, that could potentially be used as sources for synthetic data generation, we run into a potential issue of the linguistic misalignment between these well-formed texts, most likely written by native or highly proficient speakers of Spanish, with the linguistic characteristics of L2 learner writing. Particularly, when considering data drawn from Wikipedia and news articles, the texts likely contain a much more formal style than our target student writing. Additionally, such data sources are known to contain far fewer examples of certain grammatical forms; for example, Wikipedia and news data is likely to contain fewer instances of first- and second-person construction than we see in student texts. However, the availability of L2 student-written Spanish language texts is far more limited than general Spanish language text. As such, I develop two different synthetic errorful datasets using my proposed approach that I use to fine-tune language models using several different setups.

First, I extract a large (5 million sentences) dataset of general Spanish text, drawn at random from the Spanish BERT training dataset of approximately 300 million sentences made available by Cañete et al. [2020]. This data comes from fifteen different sources, including Spanish Wikis, ParaCrawl [Bañón et al., 2020], and OpenSubtitles, to name a few. Details of the dataset sources can be found at `https://github.com/josecannete/spanish-corpora`. I preprocessed the data to remove URLs and multi-spaces, but retain casing. I also removed sentences that begin with symbols or numbers, as I found that these tend not to contain grammatical text, as well as those sentences containing non-Latin characters and those identified as containing non-Spanish text using the Python langdetect package. This preprocessing and filtering results in an initial dataset of approximately 4.3 million sentences.

Second, I source L2 Spanish learner essays from two primary sources: the uncorrected portion of the COWS-L2H corpus (2461 essays containing 51,976 sentences), and the L2 Spanish subcorpora of CEDEL2 [Lozano, 2022] (4,399 essays containing 92,906 sentences). The resulting dataset, containing approximately 145,000 sentences, is far smaller than the general Spanish text dataset mentioned above. However, given that these are actual learner-written essays elicited using prompts, these texts allow me to generate additional synthetic training data that aligns closely with the linguistic properties of the L2 learners targeted by my proposed error feedback system. Given that these texts are uncorrected, they may contain errors that will remain once synthetic errors are added. Therefore, it is important to test whether or not including this data as part of the fine-tuning of my GEC model actually improves the error correction capabilities of the model or not.

8.2.2. Alignment with linguistic properties of L2 Spanish writing. Creating a synthetic training corpus that as closely as possible reflects the linguistic properties of L2 Spanish is an important aspect of the present project. Measuring these properties, such a lexical density, lexical complexity, and syntactic complexity, is difficult when faced with single sentences, as found in the Cañete et al. [2020] dataset, rather than full texts. However, given the well known correlation between syntactic complexity and sentence length [Ortega, 2003, Kyle and Crossley, 2018], I chose to test applying an additional filter to my augmentation source data. To that end, I calculated the average number of tokens per sentence and the average number of characters per non-stop word token for each level in the COWS-L2H corpus, as shown in Table 8.2. Stop words are commonly used function words (such as 'de', 'la', 'que', 'el', 'en', 'y', 'a', and 'los') and auxiliary verbs that are considered to contribute relatively little to the semantic content of a given sentence, and that, in text processing, are considered to provide minimal value for purposes of sentence classification [Wilbur and Sirotkin, 1992]. In the current context, removing stop words helps me get a better view of the length, and hence perceived complexity, of words used by students in their writing. I use the stop word list defined for Spanish in the NLTK corpus package [Bird, 2006].

Proficiency Level	Sentence count	Avg. sentence length (in tokens)	Avg. token length (in characters)
Beginner	44,921	12.134	5.759
Intermediate	7,404	14.787	5.926
Composition	8,491	16.684	5.980
Heritage	6,934	19.401	6.165
Weighted total	67,750	13.834	5.848

TABLE 8.2. Average sentence length and average token length by clustered level from a 67k sample of COWS-L2H. This data shows a clear increase in the sentence and token length as one would expect as students progress in their acquisition of Spanish.

As an initial effort to ensure that highly complex sentences are excluded from the source data for synthetic augmentation, I calculate the outlier sentence lengths for the student cohort with the longest average sentence length (the Heritage cohort), using the interquartile range (IQR) method. This approach identifies as an outlier any value that is more than 1.5 times the IQR greater than the Q3 value, or 1.5 times the IQR less than the Q1 value. This method gives me a maximum sentence length of 40.5 tokens in length; the calculated lower bound is negative, so I do not set a lower bound. Thus, when filtering the data drawn from the Cañete et al. [2020] corpus, I remove any sentences whose token count exceeds 41 tokens, as tokenized by NLTK. I use the student cohort with the highest average token count so that the overall length of the selected source sentences is not more than the longest sentences in the corrected COWS-L2H dataset.

As for the average token length of non-stop words in each sentence, I identified an outlier range of average token length greater than 8.6 tokens, again calculated using the IQR method. To test the necessity of removing sentences whose average token length exceeds this threshold, I sampled 10,000 sentences from my dataset and checked the average token length. I found only 3 sentences in this set of 10,000 that have an average token length greater than 8.6 tokens. This finding aligns with the known Zipfian distribution of word length relative to frequency [Sigurd et al., 2004], indicating that it would be quite unlikely to have a large number of sentences with such a high average token length. Therefore, I chose to forego further filtering of the dataset based on this metric, given that it would result in very little change to the overall dataset.

8.2.3. Synthetic errorful sentence generation. Generation of synthetic training data for grammatical error correction models can be achieved using a number of approaches, ranging from random noise insertion [Xie et al., 2018] or using a reverse spell-checker to inject errors [Grundkiewicz and Junczys-Dowmunt, 2019] to injecting specific error types (such as verb conjugation errors) based on a distribution drawn from real data [Takahashi et al., 2020]. As shown in Takahashi et al. [2020] and Stahlberg and Kumar [2020], injecting errors that align with the error rates observed in real-world learner data results in better downstream system performance, potentially allowing one to generate sufficient training data to fine-tune an LLM model given a relatively small number of learner texts from which to extract error rate data and examples. Given this observation, I modify the *Tagged Corruption Models* approach proposed by Stahlberg and Kumar [2021], in which they train a sequence-to-sequence model to output a corrupted sentence given an original sentence and a target error tag, with the goal that "the tag distribution in the synthetic data can be made to match the distribution of a specific target domain" [Stahlberg and Kumar, 2021]. That is, given the known error rates observed in a dataset, one can select conditioning error tags at similar rates to create synthetic data that more closely resembles the real-world data in terms of error distribution. Training this type of model is relatively straightforward, as one need only generate a set of error tags of a size equal to the number of sentences to be corrupted with the same error tag rates as observed in real-world data, and then randomly select an error tag to condition the generation of each synthetic errorful sentence from a source sentence. This approach is similar to the "online" synthetic data generation approach described in Stahlberg and Kumar [2021]. That is, for every example sentence x_n in a source data set N, I draw an error tag t_n^* from $P*$, a known distribution of Errant error tags (extracted from parallel real data). I then condition the generation of an errorful sentence y_n on the original sentence x_n and the error tag t_n^*. In order to create a synthetic dataset that more closely resembles learner data, I maintain a separate error tag distribution, $P*_{level,L1}$ for each level and L1 combination observed in the leaner data from which I source the error tag distribution. I retain the L1 and level information so that, during GEC model training, I can condition generation on the errorful sentence along with the L1 and level, resulting in a model that considers L1 and level when making its correction predictions. Thus, the final generated synthetic dataset contains the following items for each synthetic errorful sentence: original sentence, level, L1, and the generated synthetic errorful sentence.

Given the relatively small size of the available error-corrected Spanish leaner data in the COWS-L2H corpus relative to the English data used by Stahlberg and Kumar [2021] to train their error generation model, as well as advances in large language models since their original paper was published, I do not attempt to train a transformer model from scratch on learner data as done by Stahlberg and Kumar [2021]. Rather, I fine-tune an mT5 model [Xue et al., 2021] to generate errorful sentences from my source data. Specifically, I fine-tune the model using real-world instructor-corrected data drawn from the COWS-L2H corpus, with the corrected sentence as the input and the original error-containing sentence as the output. To conduct this training, I first filter sentences from corrected sub-corpus of COWS-L2H that do not contain errors, such that I have a set of sentence that contain errors and a set that does not contain errors. Given that I want my final model to be conditioned on only one error tag per input, and the fact that many sentences in COWS-L2H contain multiple errors, I convert sentences in COWS-L2H that contain more than one error into multiple single-error examples. This is done by running ERRANT on the original-corrected sentence pair in reverse order (that is, with the corrected sentence as the original and the errorful sentence at the target in ERRANT); the resulting M2 file contains the individual edits necessary to convert the corrected sentence back into the original errorful sentence one edit at a time. I then apply these edits individually, creating a training sentence pair for each tagged edit, thus giving me a single error tag and a single error edit per example sentence pair.

Using this approach, I generate a synthetic training dataset of 3 million synthetic parallel corrected sentences from the source data drawn from the Cañete et al. [2020] dataset, as discussed in Section 8.2.1. Additionally, I use the same approach to generate 144,882 synthetic sentence pairs using the data drawn from CEDEL2 and the uncorrected portion of COWS-L2H. I use this synthetic training data for Stage 1 finetuning of my GEC model, as discussed in Section 8.4.1 below.

8.3. Synthetic data generation with LLM models

It should be noted that recent developments in generative large language models (LLMs) may make the separate training of a synthetic error generation model, as described in this section, unnecessary. Initial experiments with GPT-4 reveal that, with relatively simple prompting, the model is capable of effectively rewriting sentences to include specific error types, given a correct sentence and a description of the target error type. Future work in domain is likely to focus on the use of generative LLMs to generate synthetic training data for GEC models. Unfortunately, as shown by limited testing on data from COWS-L2H, as well as related work [Fang et al., 2023, Bryant et al., 2023], these same models have shown strong tendency to over-correct learner writing, resulting in major rewrites when only targeted corrections are desired. An example of this type of over-correction is shown below:

Model: GPT-4 (w/ 10-shot in-context learning).

Input Sentence: Él es un ex estudiante de Davis.

GPT-4 Correction: Fue estudiante en Davis.

GPT-4 Explanation: The correct verb is "fue" (was) as it's talking about a past condition.

Instructor: No change needed.

Thus, until more controllable models or better prompting strategies are developed, it is likely that LLMs will be useful in dataset generation, but that their role in actual GEC tasks will remain limited. Of course, given the rapid progress being made in this domain, my predictions may be shown incorrect in short order. As an initial test of the efficacy of using GPT-4 to generate synthetic data for the GEC task, I use the model to create a small dataset of 500,000 synthetic sentence pairs, with source sentences taken from the Cañete et al. [2020] corpus as described in Section 8.2.1 above. Errorful sentences are generated using a simple prompting strategy as shown in Figure 8.1. As with the synthetic data generation method outlined above, my approach for generating data with GPT-4 again involves drawing error tags from the known error tag distribution for the various student proficiency levels and L1s studied in Chapter 7. These error tags are then converted into plain text descriptions that I wrote or this purpose; for example, a **R:VERB:SVA** tag from ERRANT is translated to plain English as "subject-verb agreement error". Finally, I select a sentence from the 500,000 sentence set, pair the sentence with an error type, and use this information to prompt GPT-4. Similar to the synthetic data generated following the method outlined in Section 8.2.3, I test the utility of the data generated by GPT-4 by using it as stage 1 fine-tuning data for an mT5-based GEC model.

8.4. Model development and adaptation

In order to build an error feedback tool that can be used by students learning Spanish, I need a GEC system that is effectively able to correct a large cross-section of errors in Spanish learners' writing. Given that the probability of different types of errors varies significantly by L1 and level, as

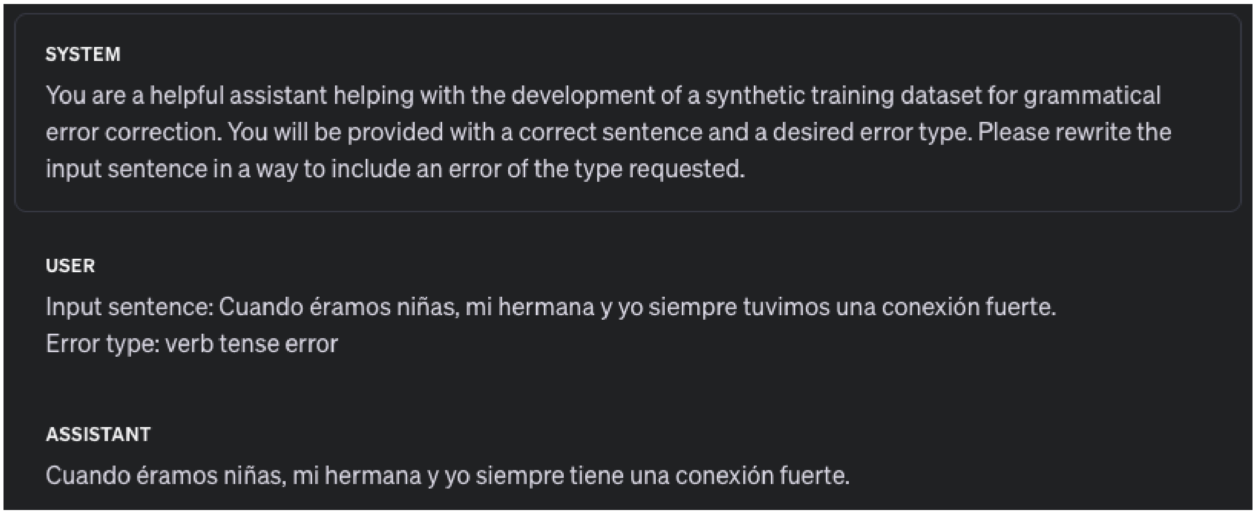

FIGURE 8.1. Example of simple LLM prompt used to generate synthetic error data using GPT-4.

discussed in Chapter 7, I hypothesize that a model trained using synthetic data that is aligned with the error distributions observed in learner text and that is able to consider L1 and level when making its predictions will be more effective in correcting student writing. To investigate this hypothesis, as well as to test the effectiveness of the synthetic data generation approach discussed in Section 8.2, I train several models using variations of synthetic data and L1 and level information, using two different pretrained LLMs as the starting point. The primary model I use in this research is mT5 [Xue et al., 2021] with additional testing done using Meta's Llama 2 [Touvron et al., 2023].

8.4.1. Primary model training. mT5 [Xue et al., 2021] is a multilingual version of the open-source T5 language model presented in the paper "Exploring the limits of transfer learning with a unified text-to-text transformer" [Raffel et al., 2020b]. As the title implies, the T5 and mT5 models take a sequence as input and generate an output sequence using an encoder-decoder architecture [Sutskever et al., 2014]. As is common with a large majority of current state-of-the-art models in NLP, T5 is built using the Transformer model architecture [Vaswani et al., 2017], with one transformer serving as the encoder and a second as the decoder. This architecture itself is not novel, as the encoder-decoder architecture using transformers is outlined in the original transformers paper [Vaswani et al., 2017]. While T5, and subsequently mT5, makes relatively minor changes in the architecture introduced by Vaswani et al. [2017], such as modifying the method of adding positional embeddings, the overall architecture remains largely unchanged. The key innovation of the T5 family of models is its use of large-scale, well cleaned pretraining data and an increased number of trainable parameters. The mT5-large model (the primary model used in the experiments reported below) has 1.2 billion parameters that have been pretrained using a masked language modelling task on a approximately 27TB of cleaned data (drawn from web crawl, books, and other openly available sources) in 101 source languages. T5 and mT5 are both designed as tools for transfer learning; that is, the model is intended to be first fine-tuned on large amounts of textual data using a language modelling objective, to create a general model capable of "understanding" language. To be applied to a specific task (in the present case, GEC), one must fine-tune the pretrained model on paired input and output sequences that represent the expected input and desired output of the target task. This task could be, for example, machine translation in which case the model would be fine-tuned on parallel translation data between languages. Considering that GEC is often framed as a monolingual machine translation task, as discussed in Chapter 3, the T5 family of models is well suited to the GEC task, and has been used widely for that purpose (see, for example, Rothe et al. [2021], Fang et al. [2023], and Qorib et al. [2022]). This choice further follows recent research in using mT5 fine-tuning for GEC in lower-resourced languages [Gomez et al., 2023, Pajkak and Pajkak, 2022, Korre and Pavlopoulos, 2022].

To fine-tune a pretrained mT5 model for task of correcting grammatical errors for second language learners of Spanish, I start with the mT5-large model, available on the Huggingface hub [9]. Our fine-tuning training data consists of:

(1) 3 million sentences of synthetic parallel error data generated using source data from the Cañete et al. [2020] corpus.

(2) 145,000 synthetic parallel sentences generated using data taken from CEDEL2 and uncorrected portion of COWS-L2H.

(3) 2,922 instructor-corrected essays from the COWS-L2H project, containing approximately 71,000 parallel training sentences (of which 70% are used for training).

To first test the efficacy of using synthetic data to fine-tune the proposed language model, I trained multiple versions of the fine-tuned model using various combinations of synthetic and real data. I first divided the real learner data into a 70/15/15 train/test/validation split. I then tested fine-tuning the model on 70% train split of real COWS-L2H learner data only - approximately 49,000

[9] https://huggingface.co/google/mt5-large

parallel sentences from COWS-L2H, as discussed in Section 8.1. This resulted in a surprisingly effective GEC model that was able to achieve an $F_{0.5}$ score of 0.525 on the held-out test set, a not unimpressive score given the limited amount of training data and the complexity of the GEC task. I then proceeded to train a series of fine-tuned models to test the importance of including synthetic training data in the LLM finetuning process, and various methods of incorporating that data. I fine-tune mT5 using the following combinations of training data and approaches; all fine-tuning uses a batch size of 16 and a maximum sequence length of 64, and I fine-tune each model version for 2 epochs on the stated dataset:

- Fine-tuning on various amounts of synthetic data only. Beginning with 500,000 randomly selected sentence pairs of synthetically generated error data, I fine-tune multiple models with progressively increasing amounts of synthetic data, increasing by 500,000 sentence pairs for each successive model (1 million, 1.5 million, 2 million, 2.5 million, 3 million). This experiment allows me to investigate how increasingly large amounts of synthetic training data affects fine-tuned model performance. Additionally, I train a baseline model for comparison that uses the reverse spell-checker technique [Grundkiewicz and Junczys-Dowmunt, 2019] to insert errors.
- Starting with the models trained in the previous step, I further fine-tune each model using the 49k training set of real learner data from COWS-L2H. This experiment investigates the effect of a two-stage fine-tuning process that first fine-tunes on synthetic data, and then fine-tunes on real learner data as a separate step. This approach facilitates the retraining of new GEC models with different small datasets of real learner data, effectively allowing end-users to train a GEC model better suited to their specific learner community.
- Combining synthetic and learner data for a single fine-tuning stage. In this experiment, for each of the dataset sizes outlined in step 1 above, I add the full 49k COWS-L2H training data and randomly sort the combined data for training. This experiment is designed to ascertain if the two-stage fine-tuning approach is better at aligning the resultant model with the target learner group, as I hypothesize it will be.
- Finally, as an initial test of the efficacy of generating training data with GPT-4, I first fine-tune an mT5 model on the 500,000 sentence-pair synthetic dataset generated using GPT-4. I then proceed to stage-2 fine-tuning on the 49k COWS-L2H training dataset.

8.4.2. L1 and proficiency level informed model training. In the next model training experiment, I test if providing the model with explicit information about proficiency level and native language of the learner writing the input sentence will improve model performance as hypothesized by allowing the model to better align its predictions with the error rates of various learner sub-populations. To conduct this experiment, I retrain an mT5-large model using the dataset configuration of my best-performing model from the dataset experiments above, namely the two-stage fine-tuning model first fine-tuned on 2 million synthetic sentence pairs and then fine-tuned on 49k sentence pairs of real learner data. However, during training and inference, I add explicit information about learner L1 and level by appending two additional special tokens to each input sentence that provide the model with information about L1 and level (this includes the artificial L1 and level information inserted into inputs during synthetic data generation, as discussed in Section 8.2.3. The method of adding these special tokens to the tokenizer vocabulary in PyTorch and HuggingFace Transformers is shown in Figure 8.2. I hypothesize that providing this information to the model at training time will allow the model to more accurately align its correction predictions with the error distributions seen in learner data for students with different L1s and proficiency levels.

```
#add the special tokens for level and L1
self.tokenizer.add_tokens(['<_beginner>', '<_intermediate>', '<_advanced>', '<_upper>'])
self.tokenizer.add_tokens(['<_english>', '<_spanish>', '<_mandarin>', '<_other>'])
```

FIGURE 8.2. PyTorch code for adding special tokens to tokenizer vocabulary

8.4.3. Alternate models tested. I addition to the primary GEC model testing using mT5 and synthetic data augmentation, I conducted limited experiments using a fine-tuned version of Meta's Llama2 [Touvron et al., 2023]. Specifically, I fine-tune the Huggingface's pretrained version of the Llama2-7b variant that contains 7 billion parameters and is pretrained using a language modeling objective, but is not fine-tuned for chat-type interactions. Unlike the sentence-by-sentence corrections used in training the mT5 model (necessitated by the format of available source data used for synthetic data generation), I fine-tuned the Llama2 model to correct entire essays, hoping to capitalize on the additional context to improve output corrections. I fine-tuned the model using a 2-shot prompting approach; that is, for each fine-tuning example, I provided the model with two example original-corrected essays pairs, a source original essay, and an output target. A full example prompt is shown in Appendix A. I fine-tune the model using LORA [Hu et al., 2021] to reduce the number of trainable parameters and 4-bit quantization to reduce the model's memory footprint. I fine-tune the model with 2045 corrected essays from COWS-L2H. I then test the model's performance in the GEC task by using the model to correct a held-out test set of 438 corrected essays, splitting and aligning the input and output sentences into sentence pairs as required for ERRANT, and evaluating the $F_{0.5}$ using ERRANT.

8.5. Results

I report overall results from my primary model variants in Table 8.3. As can be seen, my best performing model is the two-stage fine-tuned mT5-large, first fine-tuned on 2 million sentences of synthetic data generated using the approach described in Section 8.2.3, using error rate data extracted from error-corrected COWS-L2H data to create a synthetic dataset that reflects the error distribution seen in our real learner data. These results confirm my hypothesis that an error-informed method of generating synthetic data is a well-motivated approach that significantly improves model performance relative to simpler data augmentation approaches. A very similar approach is used in Stahlberg and Kumar [2021] to generate a massive (500 millions sentence) synthetic dataset for English GEC; they demonstrate that at such a large scale, synthetic data may even be able to replace real error-corrected training data entirely for languages where error corrected data is unavailable.

Additionally, providing the model information about L1 and level, as described in Section 8.4.2, results in a small, but consistent, improvement in model performance across all models tested, as compared to the corresponding model trained with no L1 and level information. This improvement, while modest, still supports the provision of L1 and Level information to the GEC model at training time, especially since providing this information is straightforward and does not increase required resources for model training. Of course, further testing with additional languages and possibly larger datasets would help clarify if the improvement seen with this particular dataset can be generalized to other languages and datasets.

Next, to understand the impact of synthetic training data size on model performance, I report the F1 score for the "Both (ERB, two-stage) w/ L1 and Level" condition trained with increasingly large amounts of synthetic data. This is the variant in which I first fine tune on the error-rate based

Training data	**Precision**	**Recall**	$\mathbf{F_{0.5}}$
Artificial only (3 mil) - reverse ASpell	0.047	0.024	0.039
Artificial only (3 mil) - error rate based (ERB)	0.205	0.263	0.214
Artificial only (500k) - GPT-4 generated ERB	0.394	0.293	0.369
COWS-L2H only (12k sents)	0.216	0.131	0.191
COWS-L2H only (49k sents) w/o L1 and Level	0.608	0.311	0.510
COWS-L2H only (49k sents) w/ L1 and Level	0.619	0.326	0.525
Both (ASpell, single-stage) w/o L1 and Level	0.572	0.313	0.491
Both (ERB, single-stage) w/o L1 and Level	0.671	0.381	0.582
Both (ERB, single-stage) w/ L1 and Level	0.692	0.397	0.602
Both (GPT-4, single-stage) w/o L1 and Level	0.695	0.478	0.637
Both (GPT-4, single-stage) w/ L1 and Level	0.715	0.491	0.655
Both (ASpell, two-stage) w/o L1 and Level	0.596	0.339	0.518
Both (ERB, two-stage) w/o L1 and Level	0.684	0.372	0.586
Both (ERB, two-stage) w/ L1 and Level	0.696	0.394	0.603
Both (GPT-4, two-stage) w/o L1 and Level	0.699	**0.514**	0.652
Both (GPT-4, two-stage) w/ L1 and Level	**0.737**	0.505	**0.675**
Llama2 finetuned on COWS-L2H	0.663	0.502	0.623

TABLE 8.3. Precision, Recall and $F_{0.5}$ scores for primary mT5 model variants, along with limited results from fine-tuned Llama2. ERB refers to error rate based synthetic data generation. GPT-4 refers to synthetic data generated using a simple prompting approach with GPT-4. Single-stage refers to a single training stage with both synthetic and real learner data combined. Two-stage refers to first fine-tuning on synthetic data, then fine-tuning on learner data. All models (other than Llama2) are fine-tuned mT5-large.

synthetically generated data, then fine-tune on COWS-L2H data in a two-stage fine-tuning process. These results, shown in Figure 8.3, demonstrate that model improvement garnered from increasingly large amounts of synthetic training data begins to level-off after 2 million synthetic sentences used. All of my reported models in Table 8.3 are trained on 3 million synthetic sentence pairs (with the exception of GPT-4 generated synthetic data, which I only tested at 500,000 sentence pairs), as that was the best performing quantity in my experiments. However, it seems that, given computational or time constraints, one could likely limit the quantity of synthetic data to 2 million sentences or less with little effect on overall model performance, as additional synthetic data provides diminishing returns after this point.

To further compare the GEC performance of these models, particularly when comparing synthetic training data approaches, it is important to consider the precision and recall by error type. Specifically, Table 8.4 and Table 8.5 show the model precision, recall and $F_{0.5}$ for the *COWS-L2H only (49k sents) w/ L1 and Level* and *Both (GPT-4, two-stage) w/ L1 and Level* model conditions,

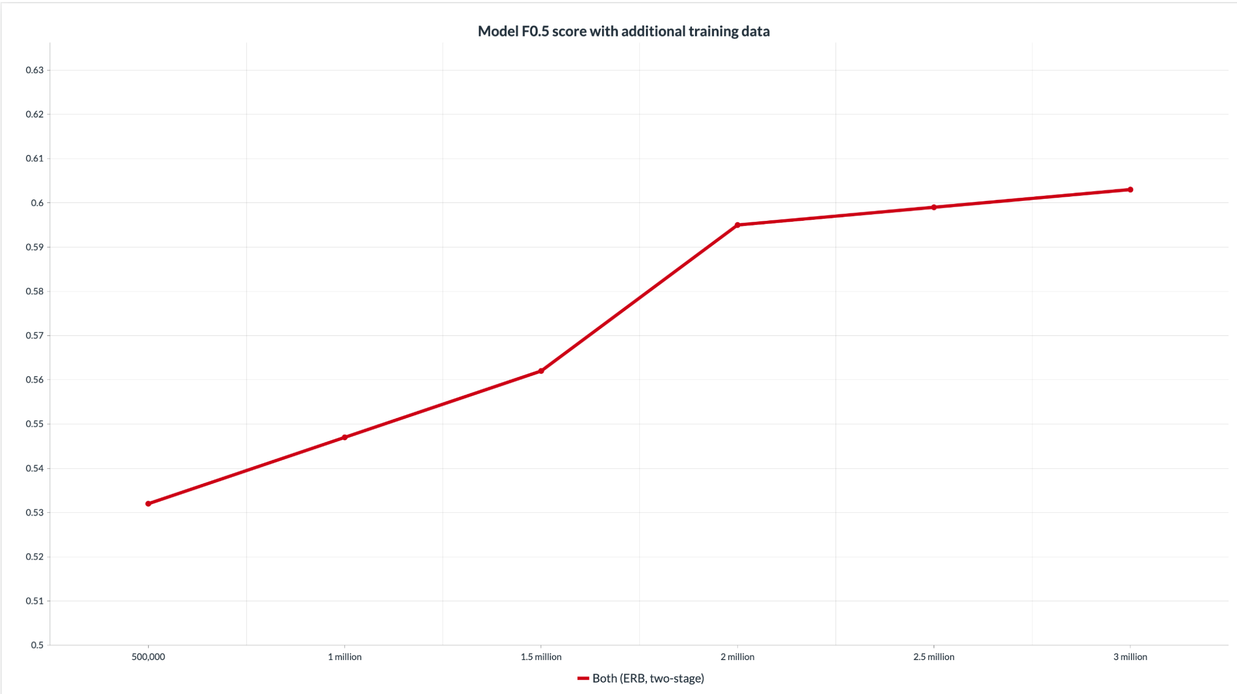

FIGURE 8.3. Model performance changes with increasing quantities of synthetic data. The Y-axis is $F_{0.5}$ score, while the X-axis is the number of synthetic sentence pairs used in stage 1 of the 2-stage model training process.

respectively. As can be seen in these tables, while the latter model (the best performing model in the experiments I conducted) outperforms the former model trained on real learner data only (as expected given the additional training data), improvements are not even. Although I do observe improvement in nearly every category, some show much more improvement than others. Specifically, the most improvement is observed in the ADJ (adjective), ADV (adverb), PREP (preposition) and VERB (verb) categories. Likewise, the least improvement (among the categories with more than an handful of examples in the test set) is seen in the NOUN, PRON (pronoun) and SPELL (spelling error) categories. The fact that nouns are among the least improved categories is not terribly surprising given that lexical (word-choice) errors are notoriously difficult to correct, given the large number of potential replacement targets that exist. That is, since there are a large number of other nouns that can be valid substitutes for an erroneous noun, it is challenging for the language model to select the same replacement as the human corrector in many cases. Spelling errors are likewise difficult because, given a greater edit distance between the misspelled word and the intended target word, it may be difficult for the language model to select the intended target depending on

Category	TP	FP	FN	P	R	F0.5
ADJ	26	18	71	0.591	0.268	0.476
ADV	5	3	38	0.625	0.116	0.333
CONJ	1	0	1	1	0.5	0.833
DET	313	161	412	0.660	0.432	0.597
MORPH	42	22	58	0.656	0.42	0.59
NOUN	2917	1922	6135	0.603	0.322	0.513
NOUN:INFL	5	7	21	0.417	0.192	0.338
NOUN:NUM	3	4	5	0.429	0.375	0.417
NOUN:POSS	0	0	1	1	0	0
ORTH	242	143	235	0.629	0.507	0.6
OTHER	471	427	2249	0.525	0.173	0.373
PART	3	0	8	1	0.273	0.652
PREP	27	25	113	0.519	0.193	0.388
PRON	125	79	105	0.613	0.544	0.598
PUNCT	14	15	48	0.483	0.226	0.393
SPELL	1805	868	2803	0.675	0.392	0.59
VERB	68	43	237	0.613	0.223	0.454
VERB:INFL	0	3	0	0	1	0
VERB:SVA	1	1	1	0.5	0.5	0.5
VERB:TENSE	0	1	1	0	0	0
WO	31	13	84	0.705	0.27	0.533
TOTAL	**6099**	**3755**	**12626**	**0.62**	**0.326**	**0.525**

TABLE 8.4. Precision, Recall and $F_{0.5}$ scores by error type for the *COWS-L2H only (49k sents) w/ L1 and Level* model condition.

context. Additionally, the classification rules used in ERRANT can result in some changes being classified as spelling errors that may more appropriately fall into another category, such as verbal morphology errors, potentially resulting in an over-representation of this particular class in the evaluation scoring. Regardless of these imbalances in performance between different error classes, the key takeaway of these tables is that we observe improvement in nearly every category, thus further demonstrating the efficacy of data augmentation for this task.

One further observation worth noting from the experiments reported in Table 8.3 is the fact that the fine-tuned Llama2 model, which I fine-tuned on COWS-L2H parallel learner data only, significantly outperforms the mT5 model trained on the same dataset (*COWS-L2H only (49k sents) w/ L1 and Level*). Thus, it is quite possible that fine-tuning Llama2 on the combined synthetic and learner data may result in a model that outperforms its mT5 counterpart. However, given that my primary goal with the present project is to implement an effective GEC model that can be used at scale as part of an automated written corrective feedback application for learners of Spanish, I chose

Category	TP	FP	FN	P	R	F0.5
ADJ	54	12	43	0.818	0.557	0.748
ADV	22	7	22	0.759	0.5	0.688
CONJ	1	0	1	1.000	0.5	0.833
DET	336	128	389	0.724	0.463	0.651
MORPH	47	20	53	0.701	0.47	0.639
NOUN	4629	1795	4423	0.721	0.511	0.666
NOUN:INFL	12	9	14	0.571	0.462	0.545
NOUN:NUM	6	5	2	0.545	0.75	0.577
NOUN:POSS	0	0	1	0	0	0
ORTH	276	121	201	0.695	0.579	0.668
OTHER	1076	424	1644	0.717	0.396	0.617
PART	5	0	6	1	0.455	0.806
PREP	67	17	73	0.798	0.479	0.704
PRON	177	64	53	0.734	0.77	0.741
PUNCT	28	12	34	0.7	0.452	0.631
SPELL	2526	713	2082	0.78	0.548	0.719
VERB	139	31	166	0.818	0.456	0.706
VERB:INFL	0	2	0	0	0	0
VERB:SVA	1	1	1	0.5	0.5	0.5
VERB:TENSE	0	1	1	0	0	0
WO	53	12	62	0.815	0.461	0.707
TOTAL	**9450**	**3376**	**9276**	**0.737**	**0.505**	**0.675**

TABLE 8.5. Precision, Recall and $F_{0.5}$ scores by error type for the *Both (GPT-4, two-stage) w/ L1 and Level* model condition.

to focus my attention on improving the mT5 models rather than further exploring the capabilities of Llama2. Given Llama2's much larger size (7 billion parameters) compared to mT5-large (1.2 billion parameters), training and running the model is much more computationally demanding than mT5. As a result, deploying the Llama2 model is markedly more expensive, especially when used at scale, such as when deployed for use by a language class. Thus, all further system development reported in the following chapters utilizes the best-performing mT5 model, (*Both (GPT-4, two-stage) w/ L1 and Level*), as the method of correcting errors in student input text.

CHAPTER 9

Feedback Generation

The next step in developing CorreGram, my proposed AWCF tool for learners of Spanish, is to develop an effective method of turning automatically rewritten sentences containing error corrections into pedagogically motivated feedback for learners. As mentioned in Chapter 2, direct error feedback (feedback that simply tells a learner what their error is and how to fix it) is viewed by many researchers as a less-than-ideal manner of providing writing feedback to language learners. While the benefits of direct versus indirect feedback remain a matter of debate in language pedagogy research, many L2 writing instructors argue for the primary use of indirect feedback as "a means to engage student writers in guided problem-solving and to encourage them to take more responsibility for their own progress" [Ferris, 2010]. According to [Heift and Hegelheimer, 2017], the "usefulness of computer-generated corrective feedback largely lies in enabling learner self-study and practice of the target language by identifying and explaining error sources". Previous work had shown that providing metalinguistic explanations without explicit corrections, which I term guided self-correction, tends to result in better student engagement and immediate gains in target-form usage [Sauro, 2021] and may improve long-term learning outcomes in writing tasks [Gao and Ma, 2019, Barrot, 2021]. Given this evidence for the benefits of self-guided correction, this dissertation will focus on a means of providing indirect feedback to language learners based on corrections made by the grammatical error correction model discussed in Chapter 8. Specifically, I modify an approach to error feedback presentation proposed in Liang et al. [2023]. In this approach, corrections made by the GEC model to student input are extracted and tagged for error type using ERRANT. Errors to be presented to students are then prioritized based on a set of priority rules that can be modified by instructors, and feedback is generated in according to one of two primary strategies. If the error is of a type that has a hand-written feedback template available, the feedback template is populated with specific information from the input text [Woodworth and Barkaoui, 2020]. If the error type does not have a corresponding template defined in the feedback generation script, then the application uses GPT-4 to generate an appropriate feedback message to be presented to the student. These strategies and methods of adapting them based on instructor and student needs are outlined in detail below.

9.1. Adaptability and Error Prioritization

One of the key features of the AWCF system I propose for students learning Spanish is its adaptability to student proficiency level. Not only does the GEC model consider L1 and proficiency when generating its corrections, but how and which of those errors are presented to students can be readily adapted to learner proficiency by instructors whose students are using the system. Currently, the system simply has a list of error presentation priorities that are hard-coded and are used for all students regardless of proficiency level. For example, the current system design first presents students with feedback on subject-verb agreement errors, if any are found, followed by other verb inflection errors, followed by preposition choice errors, and so forth. However, this "priority list" is easily modified by system users, particularly instructors whose students are using the system, to prioritize different errors, or to prioritize them in a different order, according to the pedagogical goals of the course and/or student proficiency level. For example, if one of the course objectives is inculcating correct use of prepositions, an instructor may want to prioritize preposition omission and choice errors over, for example, adjective agreement errors. Setting an appropriate priority for error presentation is important, as the system also allows instructors to set a maximum number of errors to be presented to students. The purpose of the maximum number of errors to be presented, and the ranking of errors based on perceived pedagogical impact, is based on the "less is more" approach discussed in Lee [2019]; while this study focused on feedback provided by instructors, rather than automated feedback, the underlying principle is that focused feedback, designed to target the specific forms that are appropriate to a student's proficiency level and course objectives, is both more engaging to students and results in better learning outcomes than comprehensive feedback that attempts to correct every error in a student submission [Lee, 2019]. This idea aligns with studies such as Ranalli [2018] and Bitchener and Ferris [2012], who discuss the fact that presenting students with corrections that exceed their expected knowledge of the language may do more harm than good. For example, it may be detrimental to present a Spanish 1 student with a correction indicating that a verb should be in the subjunctive mood, given that this correction is likely to cause confusion rather than helping the student to understand and internalize the underlying grammar.

As previously mentioned, instructors are unlikely to correct every error in a student text, either because they wish to target particular learning objectives or to avoid overwhelming students with corrections [Hendrickson, 1980]. Similarly, the proposed error feedback system should be able to adjust the error output based on learning objectives and number of errors, as providing feedback that is too advanced or proposing too many corrections may overwhelm or confuse the student user. While it does not make sense to set aside all errors except those which are the target of the current curriculum, it is possible to prioritize those errors that are most relevant to the current course objectives. Additionally, if the system detects a number of errors above a set threshold (that can be adjusted depending on student proficiency and preferences), the system should begin removing errors from its output starting with those of the lowest priority, again to avoid overwhelming students with an excessive number of presented suggestions. The tolerable number of corrections offered in a single task is a pedagogical question that needs to be decided in consultation with language educators and through feedback from student users. Currently, CorreGram defaults to presenting a maximum of three errors per student essay submitted, and does not vary this number based on proficiency. This default system setting is arbitrary and may not reflect a reasonable number of errors that students can effectively engage with. Also it will likely be useful to vary maximum correction counts by proficiency level; this modification may be useful in avoiding overwhelming lower-proficiency students with too many corrections while providing more advanced learners more comprehensive feedback.

9.2. Template-based feedback generation

As mentioned in Section 3.4, many existing AWCF systems use pre-written templates to provide feedback to learners. Template-generated feedback is often grouped into two basic types: specific and generic [Woodworth and Barkaoui, 2020]. Generic feedback, the easier of the two to generate, provides the same generic text whenever an error of a specific class is detected. Woodworth and Barkaoui [2020] use the sentence fragment template from *Criterion* [Ramineni and Deane, 2016] as an example; when *Criterion* detects a sentence fragment, it informs the user that "This sentence may be a fragment. Proofread it to be sure that it has at least one independent clause with a complete subject and predicate." Specific feedback, on the other hand, may be easier to students to understand and internalize, as it puts the proposed correction in context Woodworth and Barkaoui

[2020]. In CorreGram, I use specific feedback for all template-based feedback provided to students. For example, if a student wrote “Fui en la tienda”, and the system identified a **R:PREP:WC** error (that is, replacement, preposition, word choice), CorreGram’s first feedback message to the student would be “In this sentence ’Fui en la tienda.’ you made a mistake on the preposition ’en’, which doesn’t sound natural. What other preposition should you have used? Please rewrite the sentence with the correct preposition.” Thus, the student is provided with context and a brief explanation of the error type, but is then asked to correct the error themselves.

Another important feature of the template-based feedback provided by CorreGram is the fact that it provides the student with at least two opportunities to self-correct before revealing the full text of the proposed correction. Returning to the example above, if the student is unable to identify the replacement preposition needed to correct the error, the student will be presented with the following follow-up feedback “That still seems a bit off. Think about common prepositions and what might sound better here. Try one more time”, and asked once again to rewrite the sentence with the correct preposition. Only if the student is still unable to correct the error, will then be provided with the full text of the corrected sentence: “I still don’t think that’s right. I’d recommend using ’a’ in this case. Here’s the corrected sentence: ’Fui a la tienda’.” As can be seen from the progression of feedback, the goal of CorreGram’s template-based feedback is designed to encourage student self-correction as a means of improving student learning outcomes. However, given the debate in the second language teaching community surrounding the use and benefits of direct versus indirect feedback, CorreGram also generates a single feedback example that can be utilized instead of the self-correction templates. In the example above, the system would also generate the following feedback message, which is stored but not shown to students under default settings: “In this sentence ’Fui en la tienda’ you made a mistake on the preposition ’en’, which doesn’t sound natural. I’d recommend using ’a’ in this case. Here’s the corrected sentence: ’Fui a la tienda’.”. While the system defaults to using the “self-correction” setting for those error types for which feedback templates have been written (currently 10 common error types that were chosen in consultation with instructors from the UC Davis Department of Spanish & Portuguese), instructors can set a flag in the server script for CorreGram to turn the system on “direct feedback” mode. In “direct feedback” mode, a single feedback text is provided to students for each error. All of the templates (the three templates comprosing the “self-correction” mode, as well as the single "direct feedback" text) are generated simultaneously by the feedback generation script and returned to the application. CorreGram then presents the individual feedback texts to the user as needed depending on the user's ability to self-correct, or in the case of "direct feedback", the single direct feedback text is presented to the user. This ability to choose between "self-correction" and "direct feedback" modes provides flexibility to instructors to adapt the proposed system to their pedagogical approaches, as well as for conducting experiments comparing direct and indirect feedback approaches in AWCF systems. An example of the Python code for generating feedback for a single error type is shown in Figure 9.1 below.

```
        if "R:VERB:SVA" in target:
            #Don't know if I should include the full sentence
            response_short = "In this sentence '{orig_sent}' you made a mistake on the verb '{orig_tok}'. The correct verb form
here is '{cor_tok}'. Remember to make your verbs agree with their subjects. Here's the corrected sentence:
{cor_sent}".format(orig_sent=orig_sentence.text, orig_tok=edit_item.o_str, cor_tok=edit_item.c_str, cor_sent=cor_sentence.text)
            llm_explanation = chain2.invoke({"l1": l1, "level": level, "original": orig_sentence, "corrected": cor_sentence})

            line_1 = "In this sentence '{orig_sent}' you made a mistake on the verb '{orig_tok}'. What verb form should you have
used?".format(orig_sent=orig_sentence.text, orig_tok=edit_item.o_str)
            response_1_correct = "Good job. Remember to make your verbs agree with their subjects."
            response_1_incorrect = "Not quite. Think about subject-verb agreement. How should your verb be changed to agree with
the subject '{subject}'?".format(subject=helper_functions.get_subject_phrase(orig_sentence))
            response_2_correct = "Good job. Remember to make your verbs agree with their subjects."
            response_2_incorrect = "Good try, but not quite. It's tricky, I know. The correct verb form here is '{cor_tok}'.
Remember to make your verbs agree with their subjects. Here's the corrected sentence: {cor_sent}".format(cor_tok=edit_item.c_str,
cor_sent=cor_sentence.text)

            out_dict['edit_' + str(error_count)] = {"response_short": response_short, "llm_explanation": llm_explanation,
"line_1":line_1, "response_1":{'correct':response_1_correct, 'incorrect':response_1_incorrect}, 'response_2':
{'correct':response_2_correct, 'incorrect':response_2_incorrect}}
```

FIGURE 9.1. Python code defining the set of template texts for presenting subject-verb agreement errors to CorreGram users.

Although template-based feedback is brittle when faced with unexpected student input or more complex errors, using templates offers several specific advantages over using generative LLMs to generate feedback presented to students. First, templates are controllable, allowing an instructor or system designer to know exactly what feedback will be presented to a student given a specific error type. This control makes it easier for instructors to ensure that the feedback being provided to students is level-appropriate and aligned to course learning objectives. Additionally, because templates are clearly defined in the CorreGram code, adding new templates or modifying existing templates to better suit student needs and course goals is a straightforward process.

9.3. LLM generated feedback

Given the relatively large number of possible error tags in ERRANT (see Figure 7.1, which are multiplied given that many can be either substitution, insertion, or deletion operations, writing a template for every possible error type is a challenging proposition. Rather than simply ignoring errors for which I have not written a feedback template, as is done by many template-based AWCF systems such as *Criterion* [Ramineni and Deane, 2016], I wish to provide feedback for as many errors as possible. Given recent advances in the language generation capabilities of LLMs such as GPT-4, I conducted extensive tests with GPT-3.5, GPT-4, and, most recently, GPT-4o, to test their ability to provide meaningful feedback given an original sentence and its error corrected counterpart (generated by the GEC model). As mentioned previously, GPT-4 is not an effective tool for the GEC task, as it tends to significantly rewrite student text to make it more "natural sounding" rather than providing only those corrections deemed necessary to make the text grammatical and comprehensible. However, my tests with the GPT family of LLMs revealed that these models are quite effective at providing explanations of *why* sentences have been rewritten the way they were, either by instructors or by a GEC model. That is, given an original errorful sentence written by a student and a corrected version of the same sentence, GPT-3.5 and GPT-4 can usually provide a reasonable explanation of the reasons for the differences between the two sentences. In this procedure, I crafted a prompt for the GPT models as shown in Figure 9.2. This figure shows the zero-shot setting, in which no explicit examples of the desired output type are provided to the model for in-context learning. I also tested using a three-shot setup, in which the model was provided with three examples of original and corrected sentences and their respective feedback prior to the target original and corrected sentence. While I have not conducted an extensive review of the feedback generated by GPT-4 for this task, I did conduct a qualitative review of 100 generated feedback examples for each setting. I found that 88% of the zero-shot and 90% of the three-shot feedback generation provided a reasonable explanation of the target error. Thus, while the three-shot setting does perform slightly better, at least in my limited test, its improvement may not justify the additional cost incurred by prompting the model with a significantly longer context.

Below is an example of the type of feedback generated by GPT-4o in the zero-shot setting using the prompt shown in Figure 9.2.

Input: Ella es una feminista y habla mucho de las problemas en la sociedad.

Output: Ella es feminista y habla mucho sobre los problemas en la sociedad.

Explanation from GPT-4 (w/ 3-shot in-context learning):

This sentence has a couple of errors. The first is that “feminista” is typically used without an indefinite article, so it should be “Ella es feminista.” Additionally, there’s a prepositional error. Instead of “de las problemas,” it should be “sobre los problemas” to properly convey “talking about” or “discussing” the issues.

```python
llm = ChatOpenAI(openai_api_key=OPENAI_API_KEY)

prompt2 = ChatPromptTemplate.from_messages([
    ("system", "You are the teacher of a Spanish {level} course. You are writing corrections for a student whose native
language is {l1}. You want to provide your students with feedback about mistakes in their writing. Given an original sentence
written by a student, and a corrected version of the sentence written by you, explain to the student why you made the corrections
you made. Don't change either sentence. Just explain the differences between them in terms of grammar in a way a student can
understand."),
    ("user", "Original sentence: {original}\nCorrected Sentence: {corrected}")
])

chain2 = (    prompt2
            | llm
            | StrOutputParser()
        )
```

FIGURE 9.2. Python code defining the zero-shot GPT-4 prompt for generating feedback to present to CorreGram users. This setting is used when a hand-written template is not available for the error type in question.

Given the quite reasonable performance by GPT-4 and GPT-4o in generating feedback given an original text and its corrected counterpart as context, I adopt this approach for use in CorreGram to provide feedback to learners for error types for which a hand-written template has not been created. As previously discussed, the LLM-generated feedback allows me to provide feedback to students on errors beyond the relatively constrained set for which I have written custom templates. As the CorreGram system matures and is implemented in the classroom setting, and as generative LLM technology improves and becomes less costly, I will need to continue to examine the balance between the controllability and predictability provided by hand-written templates and the flexibility, specificity, and novelty provided by LLM-generated feedback explanations. Given a good LLM and a well-crafted prompt, one could potentially prompt the model to generate multi-stage feedback in a manner similar to the feedback provided by my hand-crafted feedback templates. As an example, see the prompt and feedback generated by Claude shown in Appendix B. Clearly generating this type of complex, multi-turn feedback is possible with an LLM; however, in a small-scale test of 20 sentences containing various errors from COWS-L2H, this approach proved unreliable and the output alignment with the target errors was inconsistent. While the LLM would often generate correct feedback for one error, it would incorrectly classify others, or leave them our entirely. This lack of control of output and potential for hallucination when generating text with LLMs is a known issue (see, for example, Lee et al. [2024]. Clearly, using an LLM to generate the type of multi-stage feedback described in this chapter is worth further exploration, as careful prompt design, improved LLM training, and the use of methods to better control output (such as constrained decoding) may improve the quality of LLM output for the present task. However, I leave further exploration of this subject to future work, as the required experiments are outside the scope of the present work. It is also worth noting that running an LLM at scale can become quite costly, especially when requesting relatively lengthy outputs as required by multi-stage feedback, which could severely hamper the utility of the proposed AWCF system to Spanish language educators. As such, although I explore the use of LLMs for generation of feedback to students, CorreGram primarily relies on templates (supplemented with LLM generation when needed) to generate feedback at this time.

CHAPTER 10

AWCF Implementation and Use

Finally, I turn to integrating the previously discussed components - the grammatical error correction model outlined in Chapter 8, fine-tuned with both synthetic and real leaner data and trained to consider student proficiency level and L1 when generating error corrections, and the feedback-generation methods outlined in Chapter 9 - into a single automated written corrective feedback system that can be used by students learning Spanish to receive near-real-time feedback on grammatical and major stylistic errors in their writing, with the goal of approximating the types of corrections and feedback that an instructor might provide. The resulting web-based system is **CorreGram**, a tool that can be used on-demand by learners of Spanish, or that can be modified and hosted by individual instructors or language programs to provide error feedback to their students. A prototype of the system is currently available as a web app for use by students in target courses at UC Davis. Of course, we are still some ways from replacing the error correction, feedback provision, and pedagogical skills of a trained educator. I argue, however, that the proposed system, CorreGram, has potential application in reducing instructor workload and providing students with useful feedback on their writing in a near-real-time manner that is simply impossible for an instructor to achieve. The current iteration of the system is offered as a preliminary demonstration of how the various components discussed in previous chapters can be combined into a working interface usable by students. More work is needed to ensure that the feedback provided by the system is useful to students at scale; even if the current feedback approach is not deemed effective by language educators, the error correction and feedback methods proposed should be sufficiently adaptable to allow future researchers and language educators to modify the system to better suit the needs to their students.

10.1. Web App Design and Implementation

The web app for CorreGram is implemented in using the Streamlit open source web application development tool for Python (github.com/streamlit/streamlit). Using Streamlit to build the web application allows for rapid prototyping and deployment of CorreGram, an important feature given that the app is designed primarily as a proof-of-concept at this stage of development. The use of Streamlit also makes it as straightforward as possible for instructors and language programs to modify and update CorreGram to fit their specific needs, including using alternate language models to handle different languages and to keep up with the rapid development of LLM technology.

The initial design of CorreGram is quite simple. Students input their first language; currently, the underlying model is trained using "English", "Spanish", "Mandarin Chinese", and "Other" as the available languages, so these are the current language selections available in CorreGram. Next, students input their proficiency level; again, the underlying model is trained using three proficiency classes, "Beginner", "Intermediate", and "Advanced", so these are the choices available in CorreGram. The choices related to proficiency and L1 were made based on the most common characteristics of the student population represented in COWS-L2H. However, given the modular nature of CorreGram's design, these categories can be changed by retraining the GEC model to accept different categories and by updating the CorreGram web app script. The front page of the CorreGram web application is shown in Figure 10.1.

Once the student has entered their L1 and proficiency level, the student may proceed to inputting text to be analyzed by CorreGram, as shown in Figure 10.2. Once received, the input text is split into sentences, which are then sent one-at-a-time to the fine-tuned T5 GEC model, the goal of which is to rewrite sentences if needed to correct any grammatical or major stylistic errors present in the text, as described in Chapter 8. The output of the GEC model (the "corrected" sentence) and the input (the "original" sentence) are then aligned and compared using a modified version of the ERRANT [Bryant et al., 2017] error annotation toolkit, ERRANT-SP [Davidson et al., 2020]. The set of corrections to be presented to the student are then selected based on a set of heuristic rules; which errors will be presented is determined by a priority list that selects errors of a given type first. For example, if the priority list begins with `['R:VERB:SVA', 'R:VERB:INFL', 'R:PREP:WC', ...]`, the system will first search the tagged errors from the student essay to determine if there are any subject-verb agreement errors (`'R:VERB:SVA'`); if it finds such an error, it will generate feedback for that error and add it to the JSON file (generated by the `feedback_generation.py` script) containing the feedback to be presented to the student, and then proceed to searching for the next-highest priority error type. If the system finds more than one example of any type of error,

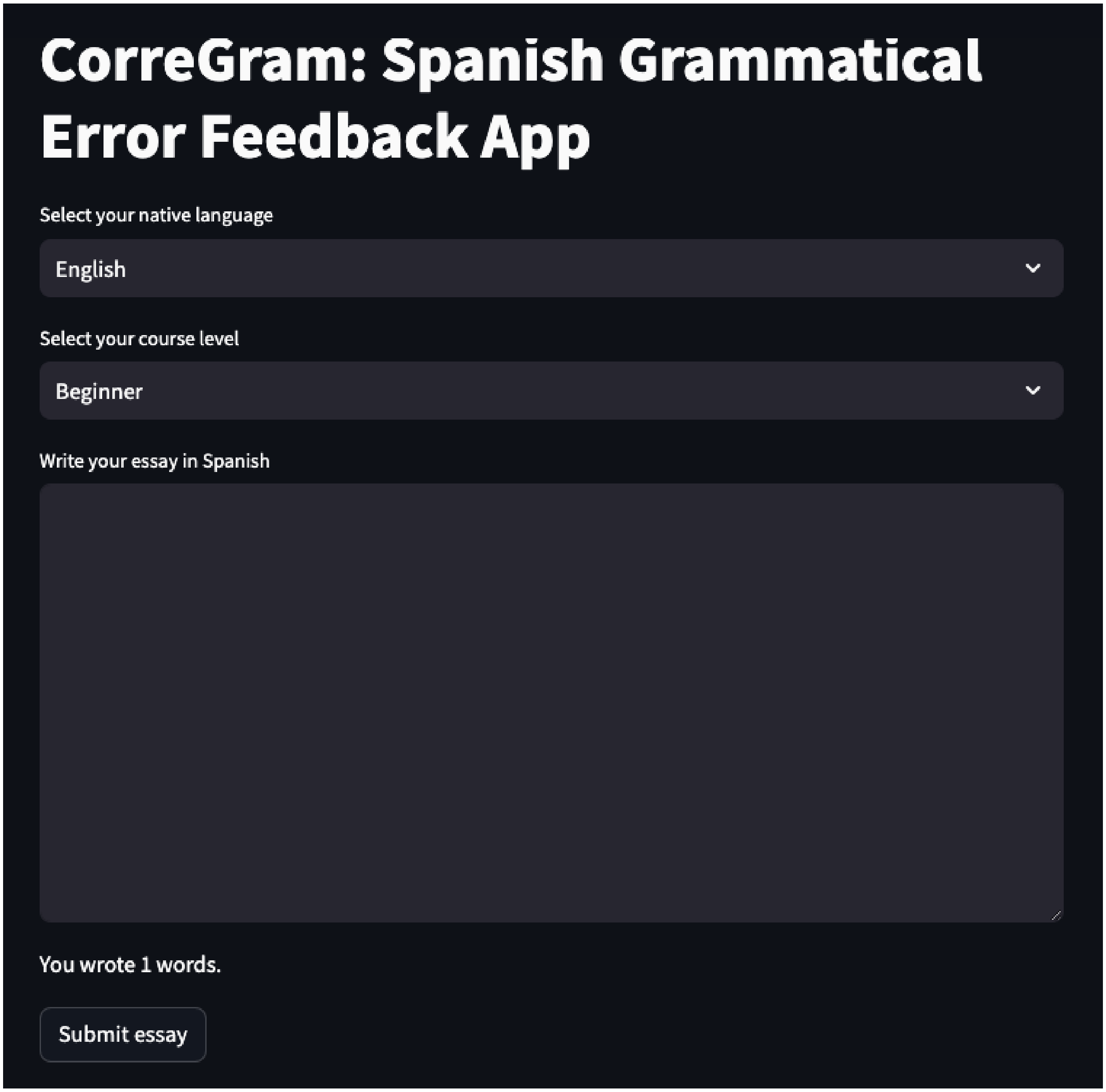

FIGURE 10.1. Web app layout as shown when accessing CorreGram online. Note the two drop downs and text box allowing students to provide demographic information and to submit text to CorreGram. The system provides a word count to students for reference, but no minimum word count is requried. CorreGram operates on a sentence-by-sentence basis, so students can provide as little or as much text as desired.

it will generate feedback for only the first such error identified. If there are no errors of a given type present in the student's essay, the system will proceed to search for the next error type in the priority list, in this case verb inflection errors ('R:VERB:INFL'). Only error types present in the priority list are presented to students; thus the priority list also serves as an exclusion list with which instructors can effectively prevent specific error types from being presented to students. The error priority list is defined by default in the current implementation of CorreGram based on an consultation with Spanish language instructors, and is currently targeted primarily to less advanced learners (that is the error types prioritized are ones that a less proficient learner is more likely to make, such as subject-verb agreement errors). However, as mentioned previously, the priority list is readily modifiable by instructors or others using CorreGram for language instruction. This feature makes it easy for instructors to define what types of errors they want the system to focus on during its feedback generation process; they can also instruct the system to completely ignore specific error types as they see fit to meet their pedagogical goals. For example, it would likely not be useful for a student in Spanish 1 to be presented with feedback instructing them to use the subjunctive mood, as that aspect of the language has not yet been taught. This control over the types of errors shown to students is pedagogically useful as it allows instructors to decide what types of errors to prioritize based on current course objectives, and which types to ignore to avoid confusing students. Additionally, while only one priority list is currently defined in the present prototype version of CorreGram, I intend to create separate priority lists for different proficiency levels, allowing the system to be deployed for use by students of varying levels of proficiency simultaneously (for example, when deployed for use by an instructor teaching multiple different classes, or for a language program as a whole).

An additional parameter that is configurable by system administrators is the number of errors in a given submission for which the system will present feedback. The purpose of this error cap is to avoid overwhelming students with too many edits; presenting an excessive number of edits can lead to confusion and feelings of discouragement in second language learners [Hendrickson, 1980], both in teacher provided and automatically generated feedback settings. By default, the system is configured to provide at most three error corrections on a given text input. However, in testing with students this limit seems to be a too-low error cap, especially for longer essays, resulting in many errors not being presented to students. In future testing, I intend to configure the system to have a flexible error presentation cap, for example allowing at most two errors to be presented per input sentence, rather than having a global maximum. As previously mentioned, this error cap is configurable by instructors and system administrators using CorreGram; they can modify the maximum error cap as they see fit based on their students' pedagogical needs. Additionally, I am planning to implement proficiency level based error caps in the next version of CorreGram, allowing administrators using the system to set different numbers of errors to be presented for students with different proficiency levels, given that students with lower proficiency are more likely to suffer from the type of discouragement previously mentioned.

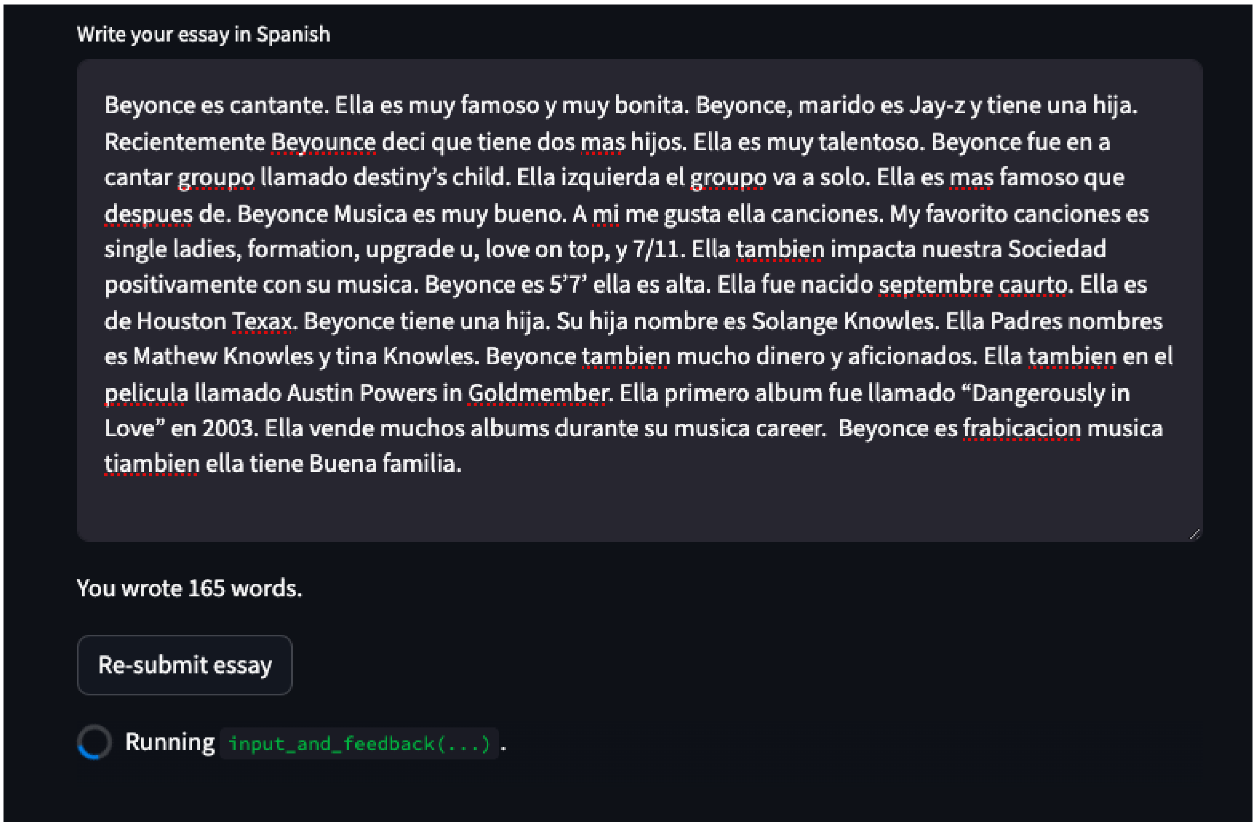

FIGURE 10.2. Processing of short essay written by a student in CorreGram.

10.2. Provision of feedback to students

Once feedback has been generated as described in Chapter 9 and Section 10.1 above, the generated feedback is presented to students in a multi-step process that presents metalinguistic clues about the error and allows for student self-correction. This style of feedback generation, termed "implicit feedback" is based on work such as Ferris [2012], Ellis et al. [2006] and Barrot [2021], who argue that providing students with direct explicit feedback, in which they are simply shown the correction and possibly an explanation, is sub-optimal. As discussed in Chapter 2 and Chapter 9, CorreGram primarily relies of "implicit" feedback, wherein students are told they made an error and provided clues about the error, such as pointing out where in the sentence the error is located. Students are then asked to correct the error, allowing students to self-correct, which is arguably more pedagogically valuable and results in better retention of the information provided [Barrot, 2021]. The ability to provide this type of multi-stage feedback in a near-real-time manner is one of the primary advantages of an AWCF system like CorreGram. Instructors simply do not have the bandwidth to provide this type of one-on-one, multi-step feedback to students, as doing so would require an instructor to be present while the student is writing, to quickly make corrections, and then to walk the student through each error and its correction. This may be possible in an individual tutoring setting, but is certainly not possible for the average language instructor in a classroom setting.

The feedback to be presented to students is generated using a combination of hand-crafted templates and generative LLM technology, as discussed in detail in Chapter 9. Once feedback has been generated and prioritized, errors are presented to the student one-at-a-time using a simple text interface, as shown in Figure 10.3. Errors for which a custom feedback template has been written are presented using the implicit feedback approach discussed above. First, the student is simply told that they made an error on the target work type, for example, preposition, verb, etc. For example, a student might be told "In this sentence, you made an error on the verb. What verb form should you have used?" The student is then given an opportunity to correct the error. If the student is able to self-correct, the system commends the student (as shown in Figure 10.4, and moves on to the next error to be presented. If the student is not able to correct their mistake, then the system provides a bit more information to help the student to self-correct; for example, "That's not quite right. Remember, in Spanish verbs have to agree with the subject of the sentence." Once again, the student is asked to self-correct. Only if the student is unable to self-correct after two attempts does the system provide the actual correction to the student. This process is shown in Figure 10.5.

For error types that do not yet have a custom feedback template written, I use GPT-4 on the OpenAI API to generate an explanation of the error and its correction that can be presented to the student, as shown in Figure 10.6. However, as discussed in Chapter 9, I have recently experimented with using LLMs to generate multi-stage feedback of the type generated using templates to date, with impressive results (see Appendix B. Thus, in the next iteration of CorreGram, it is likely

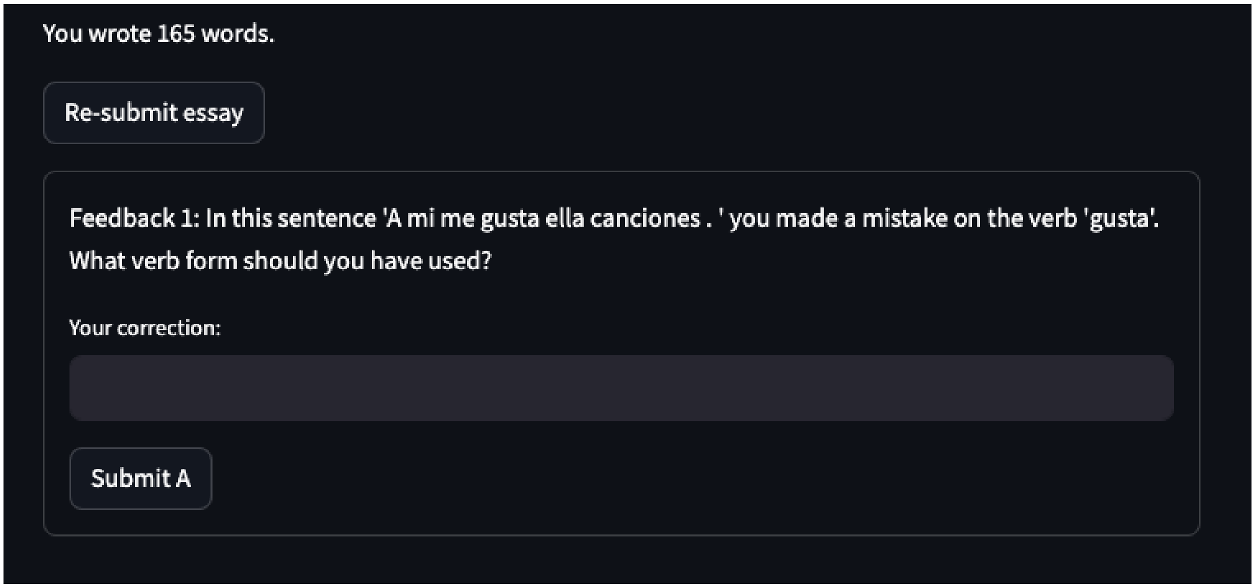

FIGURE 10.3. Presentation of an error using template-based feedback. Note that, although this sentence contains multiple errors, CorreGram only addresses one error at a time in order to avoid confusion.

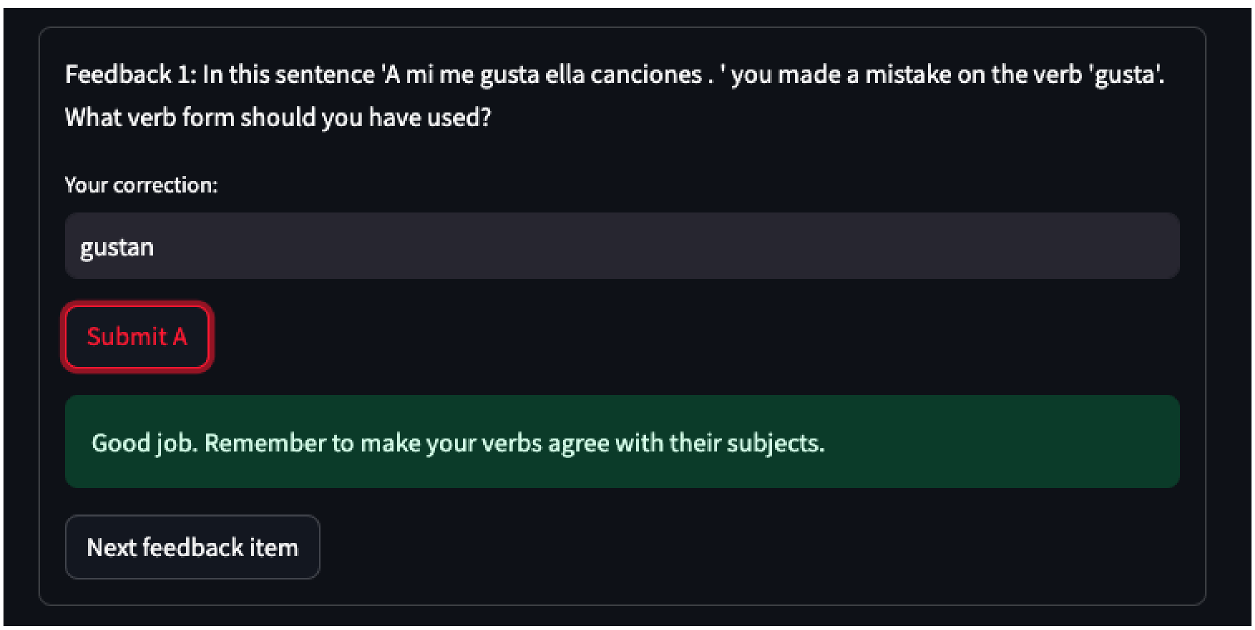

FIGURE 10.4. Successful self-correction results in a simple Good job! message to the student.

that these LLM-generated feedbacks will also take the multi-step implicit form preferred by many language pedagogy experts.

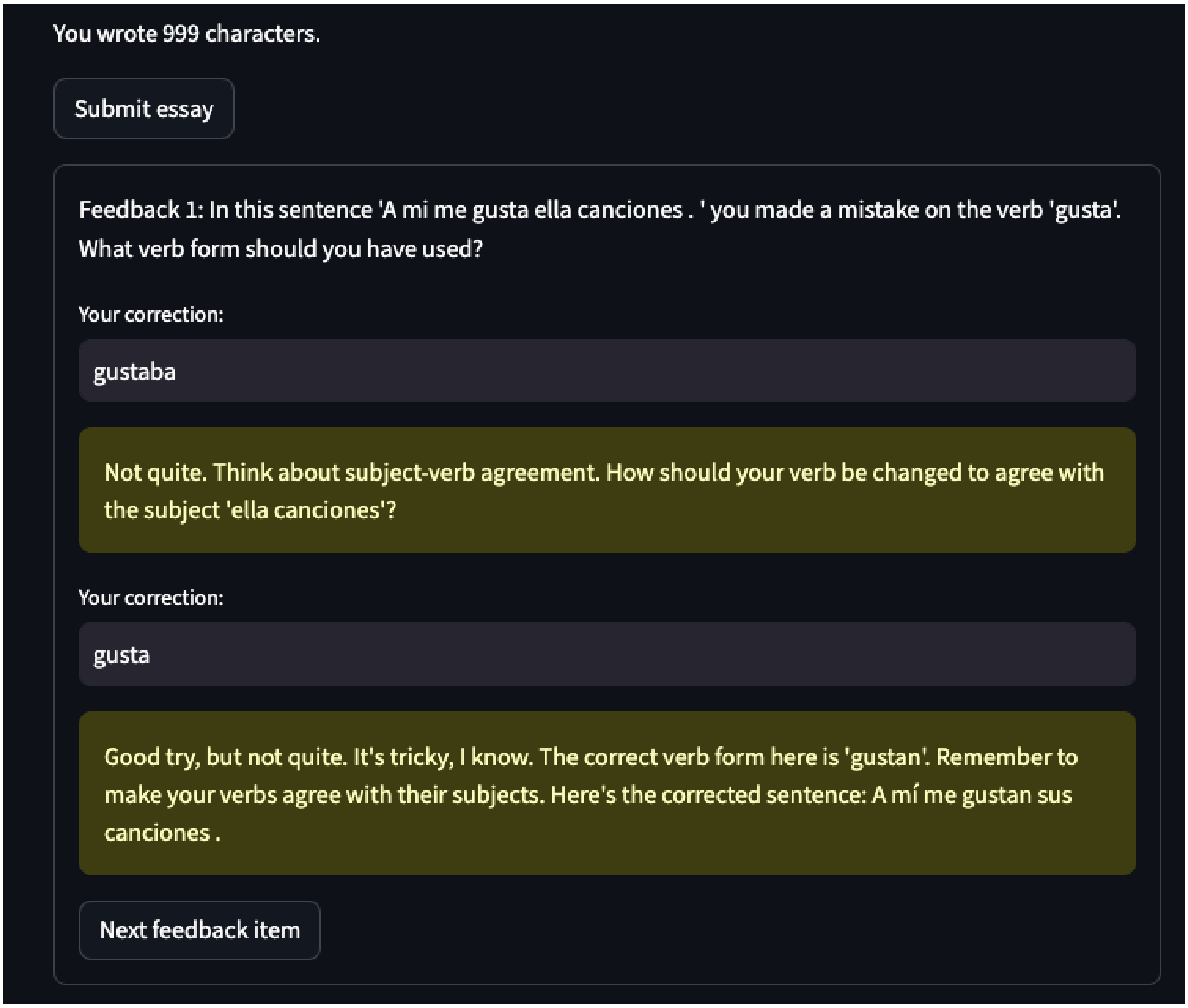

Figure 10.5. The student is provided with two opportunities to self-correct their mistake, with slightly more information provided at each opportunity. Only if the student is unable to self-correct is the correction and its explanation provided to the student.

10.3. Initial Testing

As previously mentioned, the current implementation of CorreGram is intended as a preliminary demonstration of a GEC-driven AWCF system for second language learners. As such, it has multiple possible improvements, such as expanding the amount of feedback generated using LLMs, as well as potential bugs that need to be thoroughly tested and resolved. Additionally, I am working to reduce model size and improve inference speed to facilitate lower-cost usage of the system and improve feedback response times. However, CorreGram is available for use, and it has been tested

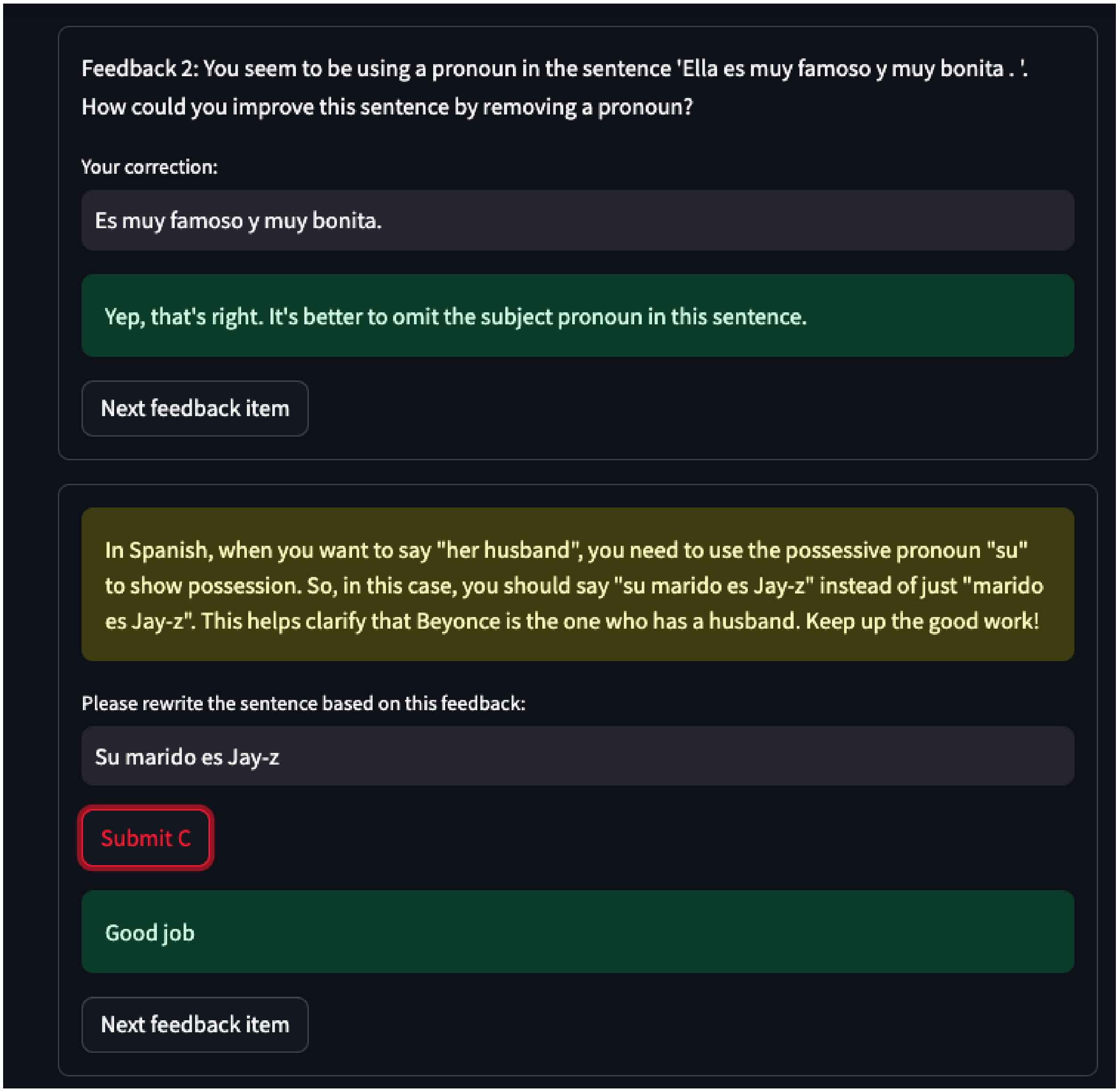

FIGURE 10.6. An example of further feedback provided to a student. In the first box, the student successful self-corrects. The student is commended for self-correcting their mistake and provided with a brief explanation. The second box is an example of presentation of an error for which a custom template has not been written. In this case, the feedback provided is generated using GPT-4.

by a handful of Spanish students, as well as instructors and others interested in the project. Initial feedback has been quite positive; overall, users have been impressed with the system's ability to correctly identify many errors and craft meaningful feedback to guide self-correction. Of course, the next step in the project is deploying the system at a larger scale; to that end, I have teamed up with instructors in the Department of Spanish and Portuguese at UC Davis to test the system in an online introductory Spanish course, which will put the system in front of many students as a means of stress-testing the system and exploring its pedagogical applications. These tests are planned for Fall 2024. Our goal with these tests is to not only receive feedback from students in terms of their perceived benefit from the system, but to also conduct pre- and post- usage writing tests to determine if using CorreGram over the course of an academic term actually results in demonstrable reductions in student error rates compared to students who receive traditional feedback from instructors only. This test will also allow us to more thoroughly examine the pedagogical assumptions that I made in implementing the system. For example, the choice to use implicit, multi-stage feedback was made based on previous corrective feedback research. However, there has been little research related to feedback modality in automated corrective feedback; thus the ability to put the system in front of students affords the opportunity to test student preferences with regard to feedback modality (implicit vs. explicit) as well as the impact of modality on the student error rates, if any. I also expect that the use of the system by students in a real-world setting will expose many bugs and other unforeseen issues, thus allowing me to improve the system in future iterations. Ultimately, the current implementation of CorreGram is a prototype, and should be treated as such, but with additional testing and refinement, most importantly, use by students in a classroom setting, I look to create a writing feedback resource for Spanish language instructors and students that is, at present, simply not available. Additionally, given that the methods used to create CorreGram and its underlying GEC model rely on a reasonably-sized corpus of error-annotated Spanish text (COWS-L2H), this approach should be adaptable to any language for which such a collection of error-corrected text is available.

CHAPTER 11

Conclusion

Much previous research argues that Automated Corrective Feedback (ACF) can serve an important role in second language acquisition [Li et al., 2015, Tatawy, 2002, Ranalli, 2018]. However, while several ACF systems are available for English learners, such as Grammarly, little work has been done to develop similar systems for Spanish learners. This lack of research into developing automated feedback tools for L2 Spanish is somewhat surprising given the fact that, as of 2016, 51% of students enrolled in university language courses in the United States were studying Spanish [of Arts & Sciences, 2016]. Additionally, as of 2013, there were over 21 million students of L2 Spanish worldwide [de la Concha et al., 2013]. In this dissertation, I present findings on the linguistic variation in the population of students enrolled in Spanish courses at a large US university. I introduce a method for generating synthetic data for training and fine-tuning language models for correcting errors in learner text that considers variations in error rates by L1 and proficiency level when generating synthetic error data. I show that synthetic data generated using this approach results in consistent improvements to a GEC model based on the T5 architecture. In order to better understand the error rates present in real learner data, I also present three studies of student error rates and syntactic development across different demographic and proficiency cohorts. These studies further support the criticism of the "one-size-fits-all" approach to error correction and feedback. Based on these studies, I proceed to train a GEC model that considers L1 and proficiency level when generating error corrections, resulting in small by consistent improvements over a baseline "one-size-fits-all" model. Finally, I present CorreGram, a new learning resource for students studying Spanish as a second or Heritage language that integrates linguistic variation in it training and generation processes, and provides near-real-time automated corrective feedback to students learning Spanish. While the system is principally targeted at adults studying Spanish as a second or Heritage language in a university setting, the proposed methods could be applied to any population of language learners for which an error distribution can be established. The proposed system is innovative in several ways. It is the first system that explicitly conditions generation of corrected output on L1 and proficiency, thus allowing the model to incorporate differences between student groups into its learned representation. Second, my proposed system is, to my knowledge, the first data-driven AWCF model proposed for use with L2 and Heritage learners of Spanish. Finally, my proposed pipeline is the first AWCF system that allows instructors to modify system parameters and feedback design to better fit specific pedagogical goals. To date I have conducted limited testing with positive results in terms of perceived usability and usefulness of the system. However, in the near future I and colleagues in the Spanish and Portuguese Department at UC Davis will deploy the system for larger-scale use by students enrolled in Spanish language courses.

The primary contribution of this dissertation is establishing that "one-size-fits-all" systems for error correction and feedback generation for L2 learners do not take into account important attributes of learner linguistic variation. I show that by taking this variation into account when training models and building systems to provide automated feedback to language learners, we stand to create more effective systems that are better at correcting errors and presenting feedback, and that provide a better learning experience for students.

APPENDIX A

Llama 2 fine-tuning prompt

Instruction: Given an essay written by a student learning Spanish, rewrite the essay to correct any grammatical or stylistic errors in the text. Your goal is not to heavily edit the essay, but rather to lightly edit it in the manner of an instructor correcting a student's writing for a class assignment.

Example 1:

Original essay:

Había una chica que no le gusta la pandemia. Cada manaña ella no quiere dejar su cama. Todos los días son abburidos y van muy despacito. Ella solo juege con su telefóno, mira las peliculas de Netflix, y come muchas comidas no sanos. Ella sueñe de va a la playa o los montañas para esquiar. No va a los conciertos, no pasa tiempo con sus amigos. Ella no puede porque nada es abierto. Solo hizo la tarea, duerme, y habla con su familia. Ella quiere grita mucho. Después una mesa, ella tiene diez mas libras y ella quiere llorar. Ella llore para muchos horas cuando Netflix juege en el antecedente. Ella quiere su vida tipicamente. Pero, en la manaña, despues una noche muy mal, todos estaba differente. Ellla tiene un nueve mentalidad porque ella realiza la sola cosa que ella tiene controlar de es su mentalidad. Ella no pasa muchos horas en la cama y fue corrando a la parque. Despues, ella va al *PLACE* para pasa tiempo con su hermano y porque ellos le gustan *PLACE*. Ella decide que no quiere pedir mas tiempo de su vida. Ella mire el bueno en todo de su vida. Ella pasa mas tiempo con su familia, ella va a correr mas. Ella tiene nuevos habiladades de yoga y organizado su diario. Ella llama los amigos del telefóno y habla mucho. Los amigos son muy graciosos y quieren la vida tipicamente tambien. Ella hizo mas tarea y tiene mas motivación para tener éxito en la vida. Ella no pido mas tiempo. Ella estaba contenta para pasar tiempo durante la pandemia. Ella estaba mas paciente y solo tiene controlar de su mentalidad. En el fin ella pasa mucho tiempo con sus amigos y van al playa. Es no una autobiografía, pero es una historia de crecimiento personal y para mirar para el bueno en todo de su vida.

Corrected essay:

Había una chica a la que no le gustaba la pandemia. Cada mañana no quería salir de su cama. Todos los días eran aburridos y pasaban muy despacito. Ella solo jugaba con su teléfono, veía películas de Netflix y comía mucha comida poco sana. Soñaba con ir a la playa o a las montañas para esquiar. No iba a los conciertos, no pasaba tiempo con sus amigos. No podía porque nada estaba abierto. Solo hacía sus tareas, dormía y hablaba con su familia. Ella quería gritar mucho. Después de un mes, ella tenía diez libras más y quería llorar. lloraba durante muchas horas cuando Netflix dejaba de funcionar. Ella quería su vida normal. Pero, una mañana, después de una noche muy mala, todo estaba diferente. Ella tenía una nueva mentalidad porque se dio cuenta de que la única cosa que tenía que controlar era su mentalidad. No pasó muchas horas en la cama y fue a correr en el parque. Después, fue al Chick-fil-a para pasar el rato con su hermano y porque a ellos les gusta el Chick-fil-a. Ella decidió que no quería perder más tiempo de su vida. Miraba lo bueno de su vida. Pasaba más tiempo con su familia, iba a correr más. Tenía nuevas habilidades de yoga y había organizado su agenda. Llamaba a los amigos por teléfono y hablaba mucho. Los amigos eran muy graciosos y querían la vida normal también. Ella hacía más tareas y tenía más motivación para tener éxito en la vida. Ella no perdió más tiempo. Ella estaba contenta en pasar el rato durante la pandemia. Estaba más paciente y solo tenía que controlar su mentalidad. Al final pasó mucho tiempo con sus amigos y fueron a la playa. No es una autobiografía, sino una historia de crecimiento personal, para enseñar todo lo bueno de su vida.

Example 2:

Original essay:

Me llamo NAME y tiene AGE AGE años. Tengo el pelo rubia, los ojos azules y soy un poco alta para una mujer. Soy altética y me encantan todas las actividades en la naturaleza. Algunos de mis favoritos son la escalada en roca, mochilero y montar a caballo. Llevo 18 años montando caballos tanto en la competencia como entrenando caballos jóvenes. El viaje de mochilar más reciente que hice fue en el desierto de la desolación en Tahoe. Mi novio y yo pasamos cinco días y caminamos unas diez millas al día. Los lagos y las vistas eran espectaculares. Un vieje de escalada en roca más reciente también fue en Tahoe. Estoy un estudiante de UNIVERSITY UNIVERSITY en mi ultima año. Estoy estudiando psicología y desarrollo humano y quiero ser un maestro de escuela elemental. En este momento trabajo como una ayuda de educación especial en una escuela primaria y me encanta mucho. En nuestro aula tenemos cinco niños con autismo y trabajamos con ellos individualmente durante la mayor parte del día. Mi parte favorita es cuando nos reunimos por la mañana con ellos y bailan alrededor de las canciones que tocamos. Tengo tres hermanos, sus nombres son NAME y NAME. NAME tiene diece nueve años y NAME tiene viente dos años. Estamos muy cerca y nos gusta pasar tiempo juntos al menos una vez a la semana. NAME y yo montamos caballos juntos mucho y a veces le doy lecciones, ya que he estado montando más tiempo que su. Mi mamá creció montando caballos también, así que a veces cabalga con nosotros.

Corrected essay:
Me llamo NAME y tiene AGE años. Tengo el pelo rubio, los ojos azules y soy un poco alta para una mujer. Soy atlética y me encantan todas las actividades en la naturaleza. Algunas de mis favoritos son la escalada en roca, ir de mochilera y montar a caballo. Llevo 18 años montando a caballo, tanto en competiciones como entrenando caballos jóvenes. El viaje de mochilera más reciente que hice fue en el desierto de la desolación en Tahoe. Mi novio y yo pasamos cinco días y caminamos unas diez millas al día. Los lagos y las vistas eran espectaculares. El viaje de escalada en roca más reciente también fue en Tahoe. Soy una estudiante de UNIVERSITY UNIVERSITY en mi último año. Estoy estudiando psicología y desarrollo humano y quiero ser una maestra de escuela elemental. En este momento trabajo como ayudante de educación especial en una escuela primaria y me encanta. En nuestra aula tenemos cinco niños con autismo y trabajamos con ellos individualmente durante la mayor parte del día. Mi parte favorita es cuando nos reunimos por la mañana con ellos y bailan las canciones que tocamos. Tengo tres hermanos, sus nombres son NAME y NAME. NAME tiene diecinueve años y NAME tiene veintidós años. Estamos muy cerca y nos gusta pasar tiempo juntos al menos una vez a la semana. NAME y yo montamos a caballo juntas mucho y a veces le doy lecciones, ya que he estado montando más tiempo que ella. Mi mamá creció montando a caballo también, así que a veces cabalga con nosotras.

Now it’s your turn:
Original essay:
{data_point[“original_essay”]}

Corrected essay:

{data_point["cor_essay"]}

APPENDIX B

Claude multi-step feedback generation

Prompt:

```
You are a helpful assitant helping the teacher of a Spanish 1 course. You are writing
    corrections for a student whose native language is English. You want to provide your
    students with feedback about mistakes in their writing. Given an original sentence
    written by a student, and a corrected version of the sentence written by you, explain
     to the student why you made the corrections you made. Don't change either sentence.
    Just explain the differences between them in terms of grammar in a way a student can
    understand.

Your goal is to provide feedback in a way that can be automatically presented to students
    , and that gives the students an opportunity to correct their own mistakes. Please
    format your response as in JSON format containing the following basic elements. Don't
     include these descriptions in your output, just use them as examples of what you
    need to generate:

line_1 = The first text presented to the student. This should provide a hint as to what
    the error is and ask the student to fix it.
response_1_correct = The text to be presented to students if they were able to
    successfully correct the error. This should be encouraging and summarize the
    correction.
response_1_incorrect = To be presented to student if they are unable to correct their
    mistake the first time they try. This should provide a little more information about
    the error in question and ask the student to correct it.
```

response_2_correct = To be presented to student if they are able to correct the error on their second try. Again, this should be encouraging and summarize the error and correction.

response_2_incorrect = Used if the student still can’t correct the error themselves. This should explain the error and the correction in detail.

The JSON you generate will be used by a script that presents feedback to students one step at a time. The student will be asked to self-correct, then if unable to do so, provided a bit more info and asked again to self-correct. If still unable to do so, the full error and its correction will be explained to the student. So keep these goals in mind when generating your output.

If there is more than one correction in the sentence, make sure to create a JSON formatted dictionary for each correction you identify in the sentence pair.

Please provide your responses in English

Now, here are the sentences to be compared:

Original sentence: Ella es una feminista y habla mucho de las problemas en la sociedad.

Corrected sentence: Ella es feminista y habla mucho sobre los problemas en la sociedad.

Claude output:

‘‘‘json


```
[
  {
    "line_1": "Check the article used with ’feminista’. Remember that in Spanish,
        professions don’t need an indefinite article when describing someone’s profession
        .",
    "response_1_correct": "Well done! You correctly removed ’una’ before ’feminista’.",
    "response_1_incorrect": "Not quite. In Spanish, we don’t use an indefinite article
        before a profession. Try removing ’una’ before ’feminista’.",
```

```
    "response_2_correct": "Great job on the second try! ’Ella es feminista’ is the correct way to express someone’s profession.",
    "response_2_incorrect": "The correct sentence is ’Ella es feminista’. In Spanish, we do not use ’una’ before a profession when stating what someone is."
  },
  {
    "line_1": "Look at the preposition used with ’habla mucho’. Is ’de’ the correct preposition to talk about topics in Spanish?",
    "response_1_correct": "Excellent! ’Sobre’ is the correct preposition to use when talking about topics.",
    "response_1_incorrect": "Not quite there. When discussing topics in Spanish, we use ’sobre’. Try replacing ’de’ with ’sobre’.",
    "response_2_correct": "You’ve got it now! ’Habla mucho sobre’ is the correct phrase for discussing topics.",
    "response_2_incorrect": "The correct phrase is ’habla mucho sobre’. In Spanish, ’sobre’ is used to talk about topics, not ’de’."
  },
  {
    "line_1": "Review the gender of the noun ’problemas’. Is the article ’las’ the correct match for ’problemas’?",
    "response_1_correct": "Correct! ’Los problemas’ matches the masculine noun ’problemas’.",
    "response_1_incorrect": "Almost there. Remember that ’problemas’ is a masculine noun, so it needs a masculine article. Try ’los’ instead of ’las’.",
    "response_2_correct": "Nice work! ’Los problemas’ is the correct usage for the masculine noun ’problemas’.",
    "response_2_incorrect": "The correct phrase is ’los problemas’. ’Problemas’ is masculine, so we use ’los’ instead of ’las’."
  }
]
```

Bibliography

iie.org. Number of international students in the United States hits all-time high. URL https://www.iie.org/Why-IIE/Announcements/2019/11/Number-of-International-Students-in-the-United-States-Hits-All-Time-High.

US Census Bureau. Hispanic heritage month 2017, Aug 2018. URL https://www.census.gov/newsroom/facts-for-features/2017/hispanic-heritage.html.

Statisticalatlas.com. The Demographic Statistical Atlas of the United States, Languages in California (state). URL https://statisticalatlas.com/state/California/Languages.

Manel Lacorte and Jesús Suárez-García. Teaching Spanish at the university level in the United States. *Observatorio de la lengua española y las culturas hispánicas en los Estados Unidos. Instituto Cervantes at the Faculty of Arts and Sciences of Harvard University*, 2016.

Silvina Montrul and Tania Ionin. Dominant language transfer in Spanish heritage speakers and second language learners in the interpretation of definite articles. *The Modern Language Journal*, 96(1):70–94, 2012.

Jinrong Li, Stephanie Link, and Volker Hegelheimer. Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, 27: 1–18, 03 2015. doi: 10.1016/j.jslw.2014.10.004.

Mounira Tatawy. Corrective feedback in SLA Corrective Feedback in Second Language Acquisition. *Studies in Applied Linguistics TESOL*, 2, 01 2002. doi: 10.7916/salt.v2i2.1645.

Jim Ranalli. Automated written corrective feedback: how well can students make use of it? *Computer Assisted Language Learning*, 31(7):653–674, 2018.

Natsuko Shintani. The effects of computer-mediated synchronous and asynchronous direct corrective feedback on writing: a case study. *Computer Assisted Language Learning*, 29(3):517–538, 2016. doi: 10.1080/09588221.2014.993400. URL https://doi.org/10.1080/09588221.2014.993400.

Zia Tajeddin, Minoo Alemi, and Roya Pashmforoosh. Acquisition of Pragmatic Routines by Learners of L2 English: Investigating Common Errors and Sources of Pragmatic Fossilization. *Tesl-Ej*, 21

(2):n2, 2017.

Cristóbal Lozano et al. CEDEL2: Corpus escrito del español L2. 2009.

Christopher Bryant, Mariano Felice, Øistein E Andersen, and Ted Briscoe. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, 2019.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. Minimally-Augmented Grammatical Error Correction. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 357–363, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020a. URL http://jmlr.org/papers/v21/20-074.html.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. A simple recipe for multilingual grammatical error correction. *arXiv preprint arXiv:2106.03830*, 2021.

Mark Davies. Corpus del español: Two billion words, 21 countries. *Corpus del español: Two billion words*, 21, 2016.

Marie Stevenson and Aek Phakiti. The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, 19:51 – 65, 2014. ISSN 1075-2935. doi: https://doi.org/10.1016/j.asw.2013.11.007. URL http://www.sciencedirect.com/science/article/pii/S1075293513000512. Feedback in Writing: Issues and Challenges.

Maria Nadejde and Joel Tetreault. Personalizing Grammatical Error Correction: Adaptation to Proficiency Level and L1. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 27–33, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5504. URL https://www.aclweb.org/anthology/D19-5504.

Silvina Montrul. Current issues in heritage language acquisition. *Annual Review of Applied Linguistics*, 30:3, 2010a.

Audrey Paredes, Cynthia Estrada, Rikka Venturanza, and Robert Teranishi. La Lucha Sigue: The University of California's Role as a Hispanic-Serving Research Institution System, 05 2021.

Maria Polinsky and Olga Kagan. Heritage languages: In the 'wild' and in the classroom. *Language and Linguistics Compass*, 1(5):368–395, 2007.

Joshua A Fishman. *Can threatened languages be saved?: Reversing language shift, revisited: A 21st century perspective*, volume 116. Multilingual Matters, 2001.

Nelleke Van Deusen-Scholl. Toward a Definition of Heritage Language: Sociopolitical and Pedagogical Considerations. *Journal of Language, Identity, and Education*, 2(3):211–230, 2003.

Maria Polinsky. American russian: Language loss meets language acquisition. In *Formal Approaches to Slavic Linguisitics. Cornell Meeting*. Michigan Slavic Publications, 1995.

Jason Rothman. Understanding the nature and outcomes of early bilingualism: Romance languages as heritage languages. *International journal of bilingualism*, 13(2):155–163, 2009.

Andrew Lynch. The linguistic similarities of Spanish heritage and second language learners. *Foreign Language Annals*, 41(2):252–381, 2008.

Carmen Silva-Corvalán. *Language contact and change: Spanish in Los Angeles*. ERIC, 1994.

William O'Grady, On-Soon Lee, and Jin-Hwa Lee. Practical and theoretical issues in the study of heritage language acquisition. *Heritage Language Journal*, 8(3):315–332, 2011.

Alicia Pousada. Bilingual education in the us. *Journal of communication*, 29(2):84–92, 1979.

Silvina Montrul. Subject and object expression in Spanish heritage speakers: A case of morphosyntactic convergence. *Bilingualism: Language and cognition*, 7(2):125–142, 2004.

universityofcalifornia.edu. Fall enrollment at a glance, Sep 2023. URL https://www.universityofcalifornia.edu/about-us/information-center/fall-enrollment-glance.

Jim Cummins. Teaching for transfer: Challenging the two solitudes assumption in bilingual education. *Encyclopedia of language and education*, 5:65–75, 2008.

Neal Snape. Exploring Mandarin Chinese speakers' L2 article use. *Representational deficits in SLA: Studies in honor of Roger Hawkins*, pages 27–51, 2009.

Tania Ionin and Silvina Montrul. The role of L1 transfer in the interpretation of articles with definite plurals in L2 English. *Language learning*, 60(4):877–925, 2010.

Hansong Cai and Luna Jing Cai. An Exploratory Study on the Role of L1 Chinese and L2 English in the Cross-Linguistic Influence in L3 French. *Online Submission*, 9(3):1–30, 2015.

Jason Rothman and Jennifer Cabrelli Amaro. What variables condition syntactic transfer? A look at the L3 initial state. *Second Language Research*, 26(2):189–218, 2010.

Khaled Karim and Hossein Nassaji. First language transfer in second language writing: An examination of current research. *Iranian Journal of Language Teaching Research*, 1:117–134, 01

2013.

Camilla Bardel and Ylva Falk. The role of the second language in third language acquisition: The case of Germanic syntax. *Second Language Research*, 23(4):459–484, 2007.

Bonnie D Schwartz and Rex A Sprouse. L2 cognitive states and the full transfer/full access model. *Second language research*, 12(1):40–72, 1996.

Sarah Williams and Bjorn Hammarberg. Language switches in L3 production: Implications for a polyglot speaking model. *Applied linguistics*, 19(3):295–333, 1998.

Suzanne Flynn, Claire Foley, and Inna Vinnitskaya. The cumulative-enhancement model for language acquisition: Comparing adults' and children's patterns of development in first, second and third language acquisition of relative clauses. *International Journal of Multilingualism*, 1(1):3–16, 2004.

Jason Rothman. L3 syntactic transfer selectivity and typological determinacy: The typological primacy model. *Second Language Research*, 27(1):107–127, 2011.

John Bitchener and Dana R Ferris. *Written Corrective Feedback in Second Language Acquisition and Writing*. Routledge, 2012.

Svetlana Koltovskaia. Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing*, 44:100450, 2020. ISSN 1075-2935. doi: https://doi.org/10.1016/j.asw.2020.100450. URL http://www.sciencedirect.com/science/article/pii/S1075293520300118.

Dana R Ferris. Written corrective feedback in second language acquisition and writing studies. *Language Teaching*, 45(4):446–459, 2012.

John Truscott and Angela Yi-ping Hsu. Error correction, revision, and learning. *Journal of second language writing*, 17(4):292–305, 2008.

Susan M Gass. Grammar instruction, selective attention and learning processes. *Foreign/Second Language Pedagogy Research*, pages 134–141, 1991.

Rod Ellis. Grammar teaching: Practice or consciousness-raising. *Methodology in Language Teaching: An Anthology of Current Practice*, 167:174, 2002.

Eun Sung Park, Sunhee Song, and Yu Kyoung Shin. To what extent do learners benefit from indirect written corrective feedback? a study targeting learners of different proficiency and heritage language status. *Language Teaching Research*, 20(6):678–699, 2016.

John Truscott. The case against grammar correction in L2 writing classes. *Language learning*, 46 (2):327–369, 1996.

Jessie S Barrot. Using automated written corrective feedback in the writing classrooms: effects on L2 writing accuracy. *Computer Assisted Language Learning*, pages 1–24, 2021.

Trude Heift and Volker Hegelheimer. Computer-assisted corrective feedback and language learning. *Corrective feedback in second language teaching and learning*, pages 51–65, 2017.

K James Hartshorn, Norman W Evans, Paul F Merrill, Richard R Sudweeks, DIANE Strong-Krause, and Neil J Anderson. Effects of dynamic corrective feedback on esl writing accuracy. *Tesol Quarterly*, 44(1):84–109, 2010.

Julie Cheville. Automated Scoring Technologies and the Rising Influence of Error. *The English Journal*, 93(4):47–52, 2004. ISSN 00138274. URL http://www.jstor.org/stable/4128980.

NTCE. *NCTE*, 2014. URL https://ncte.org/statement/writingassessment/.

Dana R Ferris and John Hedgcock. *Teaching L2 composition: Purpose, process, and practice*. Routledge, 2013.

Semire Dikli. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1), 2006.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. GECToR–Grammatical Error Correction: Tag, Not Rewrite. *arXiv preprint arXiv:2005.12592*, 2020.

James M Hendrickson. The treatment of error in written work. *The Modern Language Journal*, 64 (2):216–221, 1980.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Christopher Bryant and Hwee Tou Ng. How far are we from fully automatic high quality grammatical error correction? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 697–707, 2015.

Ian S Fraser and Lynda M Hodson. Twenty-One Kicks at the Grammar Horse: Close-Up: Grammar and Composition. *English Journal*, 67(9):49–54, 1978.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, 2014.

Muhammad Reza Qorib and Hwee Tou Ng. System combination via quality estimation for grammatical error correction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12746–12759, 2023.

Kathleen F McCoy, Christopher A Pennington, and Linda Z Suri. English error correction: A syntactic user model based on principled mal-rule scoring.

Grigori Sidorov, Anubhav Gupta, Martin Tozer, Dolors Catala, Angels Catena, and Sandrine Fuentes. Rule-based system for automatic grammar correction using syntactic n-grams for English language learning (l2). In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 96–101, 2013.

Christopher Bryant, Mariano Felice, and Edward John Briscoe. Automatic annotation and evaluation of error types for grammatical error correction. Association for Computational Linguistics, 2017.

Christopher Jack Bryant. *Automatic annotation of error types for grammatical error correction.* PhD thesis, University of Cambridge, 2019.

Alla Rozovskaya and Dan Roth. Building a state-of-the-art grammatical error correction system. *Transactions of the Association for Computational Linguistics*, 2:419–434, 2014.

Christopher Bryant and Ted Briscoe. Language model based grammatical error correction without annotated training data. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 247–253, 2018.

Daniel Dahlmeier and Hwee Tou Ng. A beam-search decoder for grammatical error correction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 568–578, 2012.

Kyusong Lee and Gary Geunbae Lee. Postech grammatical error correction system in the CoNLL-2014 shared task. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 65–73, 2014.

Kenneth Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, 2011.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Masahiro Kaneko, Kengo Hotate, Satoru Katsumata, and Mamoru Komachi. TMU Transformer System Using BERT for Re-ranking at BEA 2019 Grammatical Error Correction on Restricted Track. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 207–212, 2019.

Yiyuan Li, Antonios Anastasopoulos, and Alan W Black. Towards Minimal Supervision BERT-Based Grammar Error Correction (Student Abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13859–13860, 2020.

Simon Flachs, Ophélie Lacroix, and Anders Søgaard. Noisy channel for low resource grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 191–196, 2019.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3):643–701, 2023.

Zheng Yuan and Mariano Felice. Constrained grammatical error correction using statistical machine translation. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 52–61, 2013.

Mariano Felice, Zheng Yuan, Øistein E Andersen, Helen Yannakoudakis, and Ekaterina Kochmar. Grammatical error correction using hybrid systems and type filtering. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 15–24, 2014.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. The amu system in the conll-2014 shared task: Grammatical error correction by data-intensive and feature-rich statistical machine translation. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 25–33, 2014.

Zheng Yuan and Ted Briscoe. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–386, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 595–606, 2018.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Kenneth Heafield. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 252–263, 2019.

Marek Rei, Mariano Felice, Zheng Yuan, and Ted Briscoe. Artificial error generation with machine translation and syntactic patterns. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 287–292, 2017.

Sudhanshu Kasewa, Pontus Stenetorp, and Sebastian Riedel. Wronging a Right: Generating Better Errors to Improve Grammatical Error Detection. *arXiv preprint arXiv:1810.00668*, 2018.

Ziang Xie, Guillaume Genthial, Stanley Xie, Andrew Ng, and Dan Jurafsky. Noising and denoising natural language: Diverse backtranslation for grammar correction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 619–628, 2018.

Felix Stahlberg and Shankar Kumar. Synthetic data generation for grammatical error correction with tagged corruption models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, 2021.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. Jfleg: A fluency corpus and benchmark for grammatical error correction. *arXiv preprint arXiv:1702.04066*, 2017.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, 2021.

Sam Davidson, Aaron Yamada, Paloma Fernandez-Mira, Agustina Carando, Claudia H Sanchez-Gutierrez, and Kenji Sagae. Developing NLP tools with a new corpus of learner Spanish. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7238–7243, 2020.

Helen Yannakoudakis, Øistein E Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31(3):251–267, 2018.

Sylviane Granger. *The Computer Learner Corpus: a versatile new source of data for SLA research.* na, 1998.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

Shamil Chollampatt, Duc Tam Hoang, and Hwee Tou Ng. Adapting grammatical error correction based on the native language of writers with neural network joint models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1901–1911, 2016.

Diane Nicholls. The cambridge learner corpus: Error coding and analysis for lexicography and elt. In *Proceedings of the Corpus Linguistics 2003 conference*, volume 16, pages 572–581, 2003.

Yujin Takahashi, Satoru Katsumata, and Mamoru Komachi. Grammatical error correction using pseudo learner corpus considering learner's error tendency. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 27–32, 2020.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Min Zeng, Jiexin Kuang, Mengyang Qiu, Jayoung Song, and Jungyeul Park. Evaluating prompting strategies for grammatical error correction based on language proficiency. *arXiv preprint arXiv:2402.15930*, 2024.

Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. Analyzing the performance of gpt-3.5 and gpt-4 in grammatical error correction. *arXiv preprint arXiv:2303.14342*, 2023.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Lm-critic: Language models for unsupervised grammatical error correction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7763, 2021.

Tao Fang, Shu Yang, Kaixin Lan, Derek F Wong, Jinpeng Hu, Lidia S Chao, and Yue Zhang. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *arXiv preprint arXiv:2304.01746*, 2023.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. Large language models are state-of-the-art evaluator for grammatical error correction. *arXiv preprint arXiv:2403.17540*, 2024a.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. Revisiting meta-evaluation for grammatical error correction. *arXiv preprint arXiv:2403.02674*, 2024b.

Rod Ellis, Shawn Loewen, and Rosemary Erlam. Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in second language acquisition*, 28(2):339–368, 2006.

Kai-Hui Liang, Sam Davidson, Xun Yuan, Shehan Panditharatne, Chun-Yen Chen, Ryan Shea, Derek Pham, Yinghua Tan, Erik Voss, and Luke Fryer. Chatback: Investigating methods of providing grammatical error feedback in a gui-based language learning chatbot. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 83–99, 2023.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. Gee! grammar error explanation with large language models. *arXiv preprint arXiv:2311.09517*, 2023.

Masahiro Kaneko and Naoaki Okazaki. Controlled generation with prompt insertion for natural language explanations in grammatical error correction. *arXiv preprint arXiv:2309.11439*, 2023.

AAAS. The State of Languages in the U.S.: A Statistical Portrait, Dec 2016. URL https://www.amacad.org/publication/state-languages-us-statistical-portrait.

Instituto Cervantes. El español una lengua viva-Informe 2019. 2019.

Mark Davies. Corpus del Español: 100 million words, 1200s-1900s, 2002.

Leonardo Campillos Llanos. A Spanish learner oral corpus for computer-aided error analysis. *Corpora*, 9(2):207–238, 2014.

Rosamond Mitchell, Laura Domínguez, María J Arche, Florence Myles, and Emma Marsden. SPLLOC: A new database for Spanish second language acquisition research. *EuroSLA yearbook*, 8(1):287–304, 2008.

Guillermo Rojo and Ignacio M Palacios. Learner Spanish on Computer. The CAES 'Corpus de Aprendices de Español' Project.

Kris Buyse and Eva González Melón. El corpus de aprendices Aprescrilov y su utilidad para la didáctica de ELE en la Bélgica multilingüe. *Multilingüismo y enseñanza de ELE en contextos multiculturales*, pages 247–261, 2012.

Aaron Yamada, Sam Davidson, Paloma Fernández-Mira, Agustina Carando, Kenji Sagae, and Claudia Sánchez-Gutiérrez. COWS-L2H: A corpus of Spanish learner writing. *Research in Corpus Linguistics*, 8(1):17–32, 2020.

Nicole Tracy-Ventura, Rosamond Mitchell, and Kevin McManus. The LANGSNAP longitudinal learner corpus. *Spanish learner corpus research: Current trends and future perspectives*, 78: 117–142, 2016.

Silvina A Montrul. Is the heritage language like a second language? *Eurosla Yearbook*, 12(1):1–29, 2012.

Nick C Ellis, Rita Simpson-Vlach, Ute Römer, M O'Donnell, and Stefanie Wulff. Learner corpora and formulaic language in SLA. Cambridge University Press, 2015.

Scott A Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S McNamara, and Kristopher Kyle. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6):340–359, 2017.

Magali Paquot and Sylviane Granger. Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32(32):130–149, 2012.

Carmen Gregori-Signes and Begoña Clavel-Arroitia. Analysing lexical density and lexical diversity in university students' written discourse. *Procedia-Social and Behavioral Sciences*, 198:546–556, 2015.

Victoria Johansson. Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics*, 53:61–79, 2008.

Philip M McCarthy and Scott Jarvis. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392, 2010.

Elizabeth Bates, Inge Bretherton, and Lynn Sebestyen Snyder. *From first words to grammar: Individual differences and dissociable mechanisms*, volume 20. Cambridge University Press, 1991.

Mildred C Templin. Certain language skills in children; their development and interrelationships. 1957.

David D Malvern and Brian J Richards. A new measure of lexical diversity. *British Studies in Applied Linguistics*, 12:58–71, 1997.

Peter Broeder, Guus Extra, and Roeland van Hout. *Acquiring the linguistic devices for pronominal reference to persons: A crosslinguistic perspective on complex tasks with small words*. De Gruyter Mouton, 1986.

Jean Ure. Lexical density and register differentiation. *Applications of linguistics*, 443452, 1971.

Michael Alexander Kirkwood Halliday Halliday and Michael Alexander Kirkwood Halliday. *Spoken and written language*. Oxford University Press, USA, 1989.

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

Kate Wolfe-Quintero, Shunji Inagaki, and Hae-Young Kim. *Second language development in writing: Measures of fluency, accuracy, & complexity*. Number 17. University of Hawaii Press, 1998.

Averil Coxhead. A new academic word list. *TESOL quarterly*, 34(2):213–238, 2000.

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

HM Bell. *Using frequency lists to assess L2 texts*. PhD thesis, University of Wales Swansea, 2003.

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

Joan Bybee. *Usage-based grammar and second language acquisition*. Routledge, 2008.

Michael Tomasello and Hannes Rakoczy. What makes human cognition unique? from individual to shared to collective intentionality. *Mind & language*, 18(2):121–147, 2003.

Timothy Merrick Hall. *Learner chunks in second language acquisition*. PhD thesis, Teachers College, Columbia University, 2017.

Alison Wray. Identifying formulaic language: Persistent challenges and new opportunities. *Formulaic language*, 1:27–51, 2009.

John Sinclair and Les Sinclair. *Corpus, concordance, collocation*. Oxford University Press, USA, 1991.

Kathy Conklin and Norbert Schmitt. The processing of formulaic language. *Annual Review of Applied Linguistics*, 32:45–61, 2012.

Helen B Sorhus. To hear ourselves—implications for teaching english as a second language. *ELT Journal*, 31(3):211–221, 1977.

Britt Erman and Beatrice Warren. The idiom principle and the open choice principle. *Text & Talk*, 20(1):29–62, 2000.

Pauline Foster. Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In *Researching pedagogic tasks*, pages 85–104. Routledge, 2013.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. Longman grammar of written and spoken english. *Harlow: Longman*, 1999.

Peter Howarth. Phraseology and second language proficiency. *Applied linguistics*, 19(1):24–44, 1998.

Paul Rayson. From key words to key semantic domains. *International journal of corpus linguistics*, 13(4):519–549, 2008.

Alex Housen and Folkert Kuiken. Complexity, accuracy, and fluency in second language acquisition. *Applied linguistics*, 30(4):461–473, 2009.

Antoine Tremblay, Bruce Derwing, Gary Libben, and Chris Westbury. Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language learning*, 61 (2):569–613, 2011.

Nick C Ellis, RITA Simpson-Vlach, and Carson Maynard. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and tesol. *TESOL quarterly*, 42(3): 375–396, 2008.

Douglas Biber, Susan Conrad, and Viviana Cortes. *Lexical bundles in speech and writing: An initial taxonomy.* na, 2003.

Yves Bestgen. Comparing lexical bundles across corpora of different sizes: The zipfian problem. *Journal of Quantitative Linguistics*, 27(3):272–290, 2020.

Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang. A corpus-based approach to automatic compound extraction. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages

242–247, 1994.

Guadalupe Valdés. The teaching of heritage languages: An introduction for Slavic-teaching professionals. *The Learning and Teaching of Slavic Languages and Cultures*, 375:403, 2000.

Maria Polinsky. Gender under incomplete acquisition: Heritage speakers' knowledge of noun categorization. *Heritage language journal*, 2008a.

Silvina Montrul. Incomplete acquisition and attrition of Spanish tense/aspect distinctions in adult bilinguals. *Bilingualism: Language and cognition*, 5(1):39–68, 2002.

Andrew Lynch. The relationship between second and heritage language acquisition: Notes on research and theory building. *Heritage Language Journal*, 1(1):26–43, 2003.

Carmen Silva-Corvalán. Analyzing linguistic variation: Statistical models and methods. *Journal of Linguistic Anthropology*, 16(2):295, 2006.

Maria Polinsky. Without aspect. *Case and grammatical relations*, pages 263–282, 2008b.

Minsun Song, William O'Grady, Sookeun Cho, Miseon Lee, and Yung-Hee Kim. Korean language in america 2. 1997.

William O'Grady, Miseon Lee, and Miho Choo. The acquisition of relative clauses by heritage and non-heritage learners of korean as a second language: A comparative study. *Journal of Korean Language Education*, 12(2):283–294, 2001.

Hi-Sun Helen Kim. *Processing strategies and transfer of heritage and non-heritage learners of Korean*. PhD thesis, 2005.

Agnes Bolonyai. (in) vulnerable agreement in incomplete bilingual L1 learners. *International Journal of Bilingualism*, 11(1):3–23, 2007.

Ji-Hye Kim, Silvina Montrul, and James Yoon. Binding interpretations of anaphors by korean heritage speakers. *Language Acquisition*, 16(1):3–35, 2009.

Silvina Montrul and Melissa Bowles. Back to basics: Differential object marking under incomplete acquisition in Spanish heritage speakers. *Bilingualism*, 12(3):363–383, 2009.

Silvina Montrul. On knowledge and development of unaccusativity in Spanish L2 acquisition. 2005.

Silvina Montrul. Multiple interfaces and incomplete acquisition. *Lingua*, 121(4):591–604, 2011.

Annabelle David. A developmental perspective on productive lexical knowledge in L2 oral interlanguage1. *Journal of French Language Studies*, 18(3):315–331, 2008.

Nadja Nesselhauf. The use of collocations by advanced learners of english and some implications for teaching. *Applied linguistics*, 24(2):223–242, 2003.

Jennifer Thewissen. The phraseological errors of French-, German-, and Spanish speaking EFL learners: Evidence from an error-tagged learner corpus. In *Proceedings from the 8th Teaching and Language Corpora Conference (TaLC8)*, pages 300–306. Associação de Estudos e de Investigação Científica do ISLA-Lisboa Lisbon, 2008.

Pang Ping. A study on the use of four-word lexical bundles in argumentative essays by chinese english: A comparative study based on weccl and locness. *CELEA journal*, 32(3):25–45, 2009.

Anna Vogel Sosa and James MacFarlane. Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word of. *Brain and language*, 83(2):227–236, 2002.

Anna Siyanova-Chanturia, Kathy Conklin, and Norbert Schmitt. Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, 27(2):251–272, 2011.

Inbal Arnon and Neal Snider. More than words: Frequency effects for multi-word phrases. *Journal of memory and language*, 62(1):67–82, 2010.

Yuly Asención-Delaney and Joseph Collentine. A multidimensional analysis of a written L2 Spanish corpus. *Applied linguistics*, 32(3):299–322, 2011.

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

Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *LREC2012*, 2012.

Kristopher Kyle. Automatically assessing multiple features of lexical sophistication with taales. In *2019 conference of the American Association for Applied Linguistics (AAAL)*. AAAL, 2019.

John Frens, Ruby Davis, Jihyun Lee, Diana Zhang, and Cecilia Aragon. Reviews matter: How distributed mentoring predicts lexical diversity on fanfiction. net. *arXiv preprint arXiv:1809.10268*, 2018.

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

Scott Crossley. Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 2020.

Kenji Sagae, Alon Lavie, and Brian MacWhinney. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 197–204. Association for Computational Linguistics, 2005.

Xiaofei Lu. Automatic measurement of syntactic complexity in child language acquisition. In *International Journal of Corpus Linguistics, 14(1):3–28*, 2009.

Shannon Lubetich and Kenji Sagae. Data-driven measurement of child language development with simple syntactic templates. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2151–2160, 2014.

Scott A Crossley and Danielle S McNamara. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2):115–135, 2012.

Leonardo Zilio, Rodrigo Wilkens, and Cédrick Fairon. An sla corpus annotated with pedagogically relevant grammatical structures. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Julia Hancke and Detmar Meurers. Exploring CEFR classification for German based on rich linguistic modeling. In *Proceedings of the Learner Corpus Research (LCR) conference*, 2013.

Sowmya Vajjala and Kaidi Lėo. Automatic CEFR Level Prediction for Estonian Learner Text. In *NEALT Proceedings Series Vol. 22, pages 113–127.*, 2014.

Ildikó Pilán. Investigating the importance of linguistic complexity features across different datasets related to language learning. 2018.

Stefan Richter, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. Tracking the Evolution of Written Language Competence: an NLP–based Approach. *CLiC it*, page 236, 2015.

Xiaofei Lu. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL quarterly*, 45(1):36–62, 2011.

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

K. W. Hunt. Grammatical structures written at three grade levels. *(NCTE Research Report Number 3). Urbana, IL: National Council of Teachers of English*, 1965.

John Norris and Lourdes Ortega. Measurement for understanding: An organic approach to investigating complexity, accuracy, and fluency in SLA. *Applied Linguistics*, 30(4):555–578, 2009.

Bram Bulté and Alex Housen. Defining and operationalising L2 complexity. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, pages 23–46, 2012.

Bram Bulté and Alex Housen. Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of second language writing*, 26:42–65, 2014.

Robert J Blake and Eve C Zyzik. *El espanol y la linguistica aplicada*. Georgetown University Press, 2016.

M Rafael Salaberry. The development of past tense verbal morphology in classroom L2 Spanish. *Applied Linguistics*, 20(2):151–178, 1999.

Rod Ellis and Natsuko Shintani. *Exploring language pedagogy through second language acquisition research*. Routledge, 2013.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7147–7153, Marseille, France, May 2020. European Language Resources Association. URL https://www.aclweb.org/anthology/2020.lrec-1.883.

M. Straka, J. Hajic, and J. Strakova. UD-Pipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, 2016.

F. Dell'Orletta, S. Montemagni, and G. Venturi. READ-IT: assessing readability of Italian texts with a view to text simplification. In *Proceedings of Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, UK, July 2011a.

A. Cimino, M. Wieling, F. Dell'Orletta, S. Montemagni, and G. Venturi. Identifying Predictive Features for Textual Genre Classification: the Key Role of Syntax. In *Proceedings of 4th Italian Conference on Computational Linguistics (CLiC-it)*, pages 1–6, Rome, December 2017.

A. Cimino, F. Dell'Orletta, D. Brunato, and G. Venturi. Sentences and Documents in Native Language Identification. In *Proceedings of 5th Italian Conference on Computational Linguistics (CLiC-it)*, pages 1–6, Turin, December 2018.

Daniel Gildea. Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, 2001. URL https://www.aclweb.org/

anthology/W01-0521.

Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. Ulisse: an unsupervised algorithm for detecting reliable dependency parses. In *CoNLL*, 2011b.

Felice Dell'Orletta, Martijn Wieling, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. Assessing the readability of sentences: which corpora and features? In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173, 2014.

Sam Davidson, Aaron Yamada, Agustina Carando, Kenji Sagae, and Claudia Sánchez Gutiérrez. Word use and lexical diversity in second language learners and heritage speakers of Spanish: A corpus study. In *American Association for Applied Linguistics*, 2019.

Diego Pascual y Cabo and Inmaculada Soler. Preposition Stranding in Spanish as a Heritage Language. *Heritage Language Journal*, 12:186–209, 08 2015.

Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

Huichao Xue and Rebecca Hwa. Improved correction detection in revised ESL sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–604, 2014.

Fred J Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, 1964.

Mariano Felice, Christopher Bryant, and Ted Briscoe. Automatic Extraction of Learner Errors in ESL Sentences Using Linguistically Enhanced Alignments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 825–835, 2016.

Matthew Honnibal and Ines Montani. Spacy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1), 2017.

Madeleine Hulsen, Kees de Bot, and Bert Weltens. Between two worlds. social networks, language shift, and language processing in three generations of dutch migrants in new zealand. 2002.

Rebecca Foote. Transfer in l3 acquisition: The role of typology. In *Third language acquisition and universal grammar*, pages 89–114. Multilingual Matters, 2009.

Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

Howard Levene. Robust tests for equality of variances. *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, pages 279–292, 1961.

Bernard L Welch. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.

Bernard A Mohan and Winnie Au-Yeung Lo. Academic writing and chinese students: Transfer and developmental factors. *TESOL quarterly*, 19(3):515–534, 1985.

Tien Ping Hsiang and Steve Graham. Teaching writing in grades 4–6 in urban schools in the greater china region. *Reading and Writing*, 29(5):869–902, 2016.

Jennifer Gilbert and Steve Graham. Teaching writing to elementary students in grades 4–6: A national survey. *The elementary school journal*, 110(4):494–518, 2010.

Steve Graham. Changing how writing is taught. *Review of Research in Education*, 43(1):277–303, 2019.

Charles N Li and Sandra A Thompson. *Mandarin Chinese: A functional reference grammar*, volume 3. Univ of California Press, 1989.

Lydia White. Different? Yes. Fundamentally? No. Definiteness effects in the L2 English of Mandarin speakers. In *Proceedings of the 9th generative approaches to second language acquisition conference (gasla 2007)*, pages 251–261. Citeseer, 2008.

Terry Kit-fong Au, Leah M Knightly, Sun-Ah Jun, and Janet S Oh. Overhearing a language during childhood. *Psychological science*, 13(3):238–243, 2002.

Leah M Knightly, Sun-Ah Jun, Janet S Oh, and Terry Kit-fong Au. Production benefits of childhood overhearing. *The Journal of the Acoustical Society of America*, 114(1):465–474, 2003.

Kim Potowski, Jill Jegerski, and Kara Morgan-Short. The effects of instruction on linguistic development in Spanish heritage language speakers. *Language learning*, 59(3):537–579, 2009.

Silvina Montrul. How similar are adult second language learners and Spanish heritage speakers? Spanish clitics and word order. *Applied Psycholinguistics*, 31(1):167–207, 2010b.

Hui-Chuan Lu. An annotated Taiwanese Learners' Corpus of Spanish, CATE. 2010.

Courtney Napoles and Chris Callison-Burch. Systematically adapting machine translation for grammatical error correction. In *Proceedings of the 12th Workshop on Innovative use of NLP for Building Educational Applications*, pages 345–356, 2017.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. Automated Grammatical Error Detection for Language Learners. *Synthesis Lectures on Human Language Technologies*, 3 (1):1–134, 2010.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. Improving Grammatical Error Correction via Pre-Training a Copy-Augmented Architecture with Unlabeled Data. *arXiv preprint arXiv:1903.00138*, 2019.

Shamil Chollampatt and Hwee Tou Ng. Neural Quality Estimation of Grammatical Error Correction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2528–2539, 2018.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*, 2020.

Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

Abhisaar Yadav. Spanish Grammatical Error Correction (GEC). https://github.com/abhisaary/spanish_gec, 2022.

Jingheng Ye, Yinghui Li, Yangning Li, and Hai-Tao Zheng. Mixedit: Revisiting data augmentation and beyond for grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10161–10175, 2023.

Jared Lichtarge, Chris Alberti, and Shankar Kumar. Data weighted training strategies for grammatical error correction. *Transactions of the Association for Computational Linguistics*, 8:634–646, 2020.

Felix Stahlberg and Shankar Kumar. Seq2edits: Sequence transduction using span-level edit operations. *arXiv preprint arXiv:2009.11136*, 2020.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-scale acquisition of parallel corpora. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.417. URL https://aclanthology.org/2020.acl-main.417.

Cristóbal Lozano. CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research*, 38(4):965–983, 2022.

Kristopher Kyle and Scott A Crossley. Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102(2):333–349, 2018.

W John Wilbur and Karl Sirotkin. The automatic identification of stop words. *Journal of information science*, 18(1):45–55, 1992.

Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, 2006.

Bengt Sigurd, Mats Eeg-Olofsson, and Joost Van Weijer. Word length, sentence length and frequency–zipf revisited. *Studia linguistica*, 58(1):37–52, 2004.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020b.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Muhammad Reza Qorib, Seung-Hoon Na, and Hwee Tou Ng. Frustratingly easy system combination for grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1964–1974, 2022.

Frank Palma Gomez, Alla Rozovskaya, and Dan Roth. A low-resource approach to the grammatical error correction of ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 114–120, 2023.

Krzysztof Pajkak and Dominik Pajkak. Multilingual fine-tuning for grammatical error correction. *Expert Systems with Applications*, 200:116948, 2022.

Katerina Korre and John Pavlopoulos. Enriching grammatical error correction resources for modern greek. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4984–4991, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint* *arXiv:2106.09685*, 2021.

Dana R Ferris. Second language writing research and written corrective feedback in sla: Intersections and practical applications. *Studies in Second Language Acquisition*, 32(2):181–201, 2010.

Shannon Sauro. Computer-mediated corrective feedback and the development of L2 grammar. *UMBC Education Department Collection*, 2021.

Jianwu Gao and Shuang Ma. The effect of two forms of computer-automated metalinguistic corrective feedback. 2019.

Johanathan Woodworth and Khaled Barkaoui. Perspectives on using automated writing evaluation systems to provide written corrective feedback in the esl classroom. *TESL Canada journal*, 37(2): 234–247, 2020.

Icy Lee. Teacher written corrective feedback: Less is more. *Language Teaching*, 52(4):524–536, 2019.

Chaitanya Ramineni and Paul Deane. The criterion® online writing evaluation service. *Adaptive educational technologies for literacy instruction*, pages 163–183, 2016.

Yoonjoo Lee, Kihoon Son, Tae Soo Kim, Jisu Kim, John Joon Young Chung, Eytan Adar, and Juho Kim. One vs. many: Comprehending accurate information from multiple erroneous and inconsistent ai generations. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2518–2531, 2024.

American Academy of Arts & Sciences. The state of languages in the us: A statistical portrait, 2016.

Víctor García de la Concha, Rafael Rodríguez-Ponga Salamanca, José María Martínez Alonso, Richard Bueno Hudson, Beatriz Hernanz Angulo, and David Fernández Vítores. El español: Una lengua viva. 2013.