A Longitudinal & Cross-sectional Corpus of Annotated and Corrected L2 and Heritage Learner Spanish

Anonymous ACL submission

Abstract

The study of language learner behavior as well as the ability to train and evaluate NLP tools for language learners requires the availability of large corpora of annotated learner data. This is especially true of tasks such as error analysis, grammatical error correction, and the study of learning trajectories. In this paper, we introduce a significantly updated version of a previously released corpus of L2 and Heritage learner Spanish collected in the context of a large American university. This updated version of the corpus, collected and annotated over the course of seven years, includes an expanded collection of error-corrected parallel learner text and annotations for common errors of interest to researchers. We also add a large set of longitudinal data from students who submitted three or more essays to the corpus, allowing for analysis of individual learner second language acquisition. The corpus includes detailed metadata about each student's linguistic experience to allow further study of questions such as transfer effects in L2 learners of Spanish. Finally, we present results for a grammatical error correction model trained on the newly released data, demonstrating its utility for researchers relative to previously available datasets. The updated corpus is freely available to researchers and the general public.

1 Introduction

Understanding and analyzing the nuances of language acquisition among learners is crucial for educational development and Natural Language Processing (NLP) research. This paper introduces an updated corpus focusing on Spanish language learning, comprising 5,383 essays authored by 1,934 distinct student participants enrolled in L2 and Heritage Spanish language courses at a prominent west coast American university. The corpus features essays representing various proficiency levels, ranging from introductory to upper-division content courses. This diversity allows for examination of Spanish language acquisition across different educational stages. The corpus includes longitudinal data from students participating in the project over multiple academic terms, offering insights into individual and cohort-based language development.

Moreover, a notable aspect of this updated corpus lies in its annotation and correction methodologies. More than half of the essays, particularly those from students engaged for over three quarters, have undergone correction and annotation by experienced graduate-level Spanish instructors. This annotation process facilitates the examination of specific grammatical error types, providing a unique resource for researchers studying Spanish second language acquisition.

The paper also reports significant improvements in grammatical error correction (GEC) tasks gained when trained on the updated corpus data, demonstrating the benefits of the updated corpus relative to other Spanish learner corpora. The release of essays, metadata, corrected and annotated data, and associated code on the corpus GitHub repository[1] offers a valuable resource to the =research community.

2 Motivation

While interest in Spanish learner corpora (SLC) has increased in the past years as evidenced by recent publications on the topic (Sánchez-Gutiérrez et al., 2022), few large-scale SLC exist to date. Previous to the updated version of the corpus presented here, only two SLC included over a million tokens: APRESCRILOV (Buyse and González Melón, 2013) and CEDEL2 (Lozano and Mendikoetxea, 2013). None of these, however, contain samples from heritage speakers of Spanish in an academic context, longitudinal data, or systematic error annotations or corrections, which limits the type and breadth of studies and NLP applications that can be

[1] https://anonymous.4open.science/r/corpus-616D

Course Level	Essays	Tokens	Students
Beginner	3,018	711,630	1,658
Intermediate	562	150,279	311
Composition	883	284,746	474
Heritage	579	164,723	313
Upper Div.	222	62,404	114
Not Disclosed	119	29,476	83
Total	5,383	1,367,258	2,953

Table 1: Summary of collected essays & token count. Please note that students may have participated in more than one of the listed categories as they progressed through the Spanish course sequence. The number of unique student participants is 1,934.

Prompt description	Essays
A special person in your life	972
A famous person	893
A terrible story	833
Describe yourself	833
A perfect vacation	807
A beautiful story	694
A place you dislike	128
Chaplin clip: The Kid	104
Total	5,383

Table 2: Essay counts by prompt presented to students. Note that at any given data collection interval, only one of these prompts is used.

carried out with the SLC currently available (Rojo et al., 2022; Sánchez-Gutiérrez and Fernández-Mira, 2022).

Specifically, when it comes to NLP applications, automatic writing evaluation (AWE) and corrective feedback systems (ACF) (Shermis et al., 2013) offer unique affordances in the context of second language learning. Indeed, providing more frequent and timely feedback, without overwhelming the instructor with additional grading work, can improve students' writing (Biber et al., 2011; Graham et al., 2016). However, training and evaluating models that underlie such systems requires large parallel corrected and/or annotated corpora that are representative of specific populations of learners (Chollampatt et al., 2016; Nadejde and Tetreault, 2019).

3 Data Collection

Since 2017 we have gathered 5,383 essays written by 1,934 unique student participants enrolled in L2 and Heritage Spanish language courses at a large west coast American university. These courses range from introductory Spanish to upper-division content courses, providing samples from a variety of proficiency levels. The distribution of corpus data by proficiency level is shown in Table 1.

Over the course of an academic quarter (10 weeks of instruction), student participants are asked to write two essays in Spanish, with one essays collected during week 4 and the second during week 8. Students may participate during any quarter in which they are enrolled in a Spanish language course at the host institution. We ask students to write at least 500 words in each essay, however students in Spanish 1 are permitted a lower word count due to their relative lack of proficiency. At each data collection interval, all students are asked to respond to the same prompt, regardless of course level. The consistency of prompts across levels avoids prompt selection based on perceived difficulty (Polio and Glew, 1996), and mitigates the potential issue of prompt effects (Kroll and Reid, 1994; Way et al., 2000) when comparing texts. We have updated the prompts multiple times since beginning the project in an effort to elicit different grammatical and lexical forms from students; most prompts were used for at least one full academic year (3 quarters of instruction). To date, we have collected essays written in response to eight different prompts, as shown in Table 2.

4 Longitudinal data

One of the unique features of the updated version of the 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 (Tracy-Ventura et al., 2016) contains longitudinal written data gathered from L2 learners of Spanish, the scale of our data and the annotations available in the corpus make it a unique resource to SLC 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. 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 present corpus represents the first dataset of error-corrected and annotated longitudinal data of L2 and Heritage learners of Spanish writing available to the research community.

5 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 (2,498 essays) for several specific types of grammatical error of interest to project collaborators. Although annotations for target errors were included in the previous version of the corpus, the updated corpus significantly expands the number of annotated essays and adds additional targeted error types.

The corpus currently includes annotated errors in the following grammatical categories: gender and number agreement, personal pronouns, articles, prepositions *por* and *para*, linking verbs (*ser* and *estar*), adjective word order, and verbal morphology. 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.

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 remainder of the target data. Overall, our annotators achieved a Cohen's $\varkappa$of 0.62, indicating reasonable inter-annotator agreement.

6 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 textTo date, the compositions collected in this project have been corrected by graduate student instructors of Spanish, both at the university where the essays were collected and at a partner university in Spain.

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 3: 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 70,397 sentence count above and the numbers in this table.

The present version of the corpus greatly increases the number of parallel texts available to researchers. The previously released version contained 571 corrected essays, or roughly 13,000 parallel corrected sentence pairs. By contrast, the present version contains 2,922 corrected essays for 70,397 corrected sentence pairs. The distribution of corrected essays by student level is shown in Table 3. Note that 572 essays have been corrected by two annotators, providing two references for more flexible evaluation of GEC systems trained on the corpus data.

The parallel nature of the corrected essays allows errors to be automatically tagged using tools such as ERRANT (Bryant et al., 2017), which was 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. Our updated corpus includes errors extracted using ERRANT. To our knowledge, our corpus represents the largest parallel dataset of holistically corrected Spanish text available to researchers.

7 Corpus Application

To demonstrate the efficacy of our updated corpus data to NLP researchers, we train a grammatical error correction (GEC) model by fine-tuning mT5 (Xue et al., 2021), a multilingual text-to-text transformer model, on our parallel corrected corpus data, as described in Yadav (2022). This choice follows other recent research in using mT5 fine-tuning for GEC in lower-resourced languages (Palma Gomez et al., 2023; Pająk and Pająk, 2022;

Korre and Pavlopoulos, 2022). We also implement the BiLSTM GEC model 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, a approach which has proven fruitful in GEC applications (for example, Napoles and Callison-Burch (2017). Unlike many of the referenced works in lower-resourced GEC, we do not augment our training data with synthetically generated error data; as our goal is to demonstrate the utility of our updated corpus, we leave experiments to further improve the efficacy of Spanish GEC to future work.

To generate training data from the parallel essays provided in the corpus, we 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, we must use string matching to ensure that sentences are correctly aligned. We use The Fuzz string matching package in Python (https://github.com/seatgeek/thefuzz) and align each sentence in the original with the most similar sentence in its corrected counterpart. We release both the code used to generate the parallel dataset, as well as the extracted aligned sentences themselves, including a 70/15/15 train/test/validation split.

We implement our mT5 model in Python using Huggingface Transfomers and PyTorch Lightning, using the *mt5-base* variant of the model. During fine-tuning we use a batch size of 16 and a maximum sequence length of 128, and fine-tune for 2 epochs. Our training code and model parameters are released along with the corpus.

In order to compare to results reported in previous Spanish GEC work (Yadav, 2022; Davidson et al., 2020), we train the BiLSTM and mT5 models on our previously released corpus data, which is of comparable size to the dataset used in these works. Since both of these works use synthetic data augmentation to achieve their best performing models, we include results using the data augmentation approach described in Davidson et al. (2020).

7.1 Results

We find that the increased size of our training corpus markedly improves Spanish GEC performance, even with no synthetic data augmentation. Training mT5-base on our current 70,397 sentence pairs more than doubles the $F_{0.5}$ score achieved by 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 (both with and without data augmentation), respectively. Results are shown in Table 4. These results clearly demonstrate the utility of the parallel original-corrected essays pairs provided in our updated corpus.

Model	Recall	Precision	$F_{0.5}$
Previous corpus w/o data augmentation			
BiLSTM	0.094	0.139	0.101
mT5-base	0.30	0.102	0.216
Previous corpus w/ data augmentation			
BiLSTM	0.254	0.153	0.224
mT5-base	0.29	0.108	0.217
Updated corpus w/o data augmentation			
mT5-base	**0.619**	**0.326**	**0.525**

Table 4: GEC results demonstrating the marked improvement in model performance when trained using the updated version of the corpus.

8 Conclusion

In this paper, we present an updated corpus of 5,383 Spanish-language essays written by L2 and Heritage learners of Spanish. Unlike previous corpora of learner Spanish, the present corpus includes both cross-sectional and longitudinal data, providing researchers with a new tool for investigating Spanish language acquisition in a university classroom setting. Additionally, we have annotated a large portion of the essays for a number of error tags of interest to researchers of second language acquisition, and provide an annotation schema allowing future researchers to expand upon the existing annotation set. Finally, more than half of the essays in the corpus, including all essays from students who participated in more than 3 quarters, have been holistically corrected by graduate student instructors of Spanish. All essays, metadata, and related code are freely available in the corpus GitHub repository. We hope that this expanded corpus proves to be a valuable resource to the NLP and language research communities.

9 Ethical considerations

The key ethical consideration when collecting student data for public release is ensuring the anonymity of student participants. We have taken great care to anonymize all essays in the corpus that contain personally identifiable information (PII). All original essays were anonymized manually as part of the overall annotation process. Corrected and annotated essays were anonymized by aligning anonymized originals with non-anonymized corrected/annotated versions, and automatically replacing PII with the corresponding anonymization tag. All data that was anonymized using this automated process was subsequently reviewed by project collaborators to ensure that anonymization standards were met.

While we take great pains to remove all PII from the corpus, there is always the remote possibility that a student could be identified by the content of their writing (for example, writing a "terrible story" that is particularly unusual). Student participants are informed of the anonymization process and the remaining potential risks when they agree to participate in the project, and sign an informed consent prior to providing any writing samples or personal information to researchers.

Annotators are fairly compensated as per the standard for graduate student pay their country of residence and institution (either Spain or the United States).

This project was approved by the institutional review board of the host insitution.

10 Limitations

The data presented in this paper was all collected at a single university from students enrolled in L2 and Heritage Spanish language courses. Given the fact that all students participating in the project are learning Spanish in the same language program, the diversity of styles and the trajectory of form and lemma acquisition present in the data are likely skewed by the pedagogical goals and methodology of the program. This aspect of the corpus can be considered either a limitation, in that it reduces the diversity present in the corpus, or a feature, as it reduces variables that may affect students' patterns of acquisition of Spanish, thus facilitating analysis.

Similarly, the fact that the data was drawn from a single US university creates commonalities in the linguistic experience of participants that may affect their acquisition patterns. For example, all students must demonstrate English proficiency prior to admission to the university. Thus all students, regardless of their first language (L1) are either L1 or proficient L2 speakers of English. This may result in transfer effects from English that would not be present in non-English speaking students. Again, this may be viewed either as a limitation that reduces diversity, or a feature that provides additional avenues for research in language acquisition.

References

Douglas Biber, Tatiana Nekrasova, and Brad Horn. 2011. The effectiveness of feedback for l1-english and l2-writing development: A meta-analysis. *ETS Research Report Series*, 2011(1):i–99.

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

Kris Buyse and Eva González Melón. 2013. El corpus de aprendices aprescrilov y su utilidad para la didáctica de ele en la bélgica multilingüe. *Plurilingüismo y enseñanza de lenguas*, pages 247–52.

Shamil Chollampatt, Duc Tam Hoang, and Hwee Tou Ng. 2016. 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.

Sam Davidson, Aaron Yamada, Paloma Fernandez Mira, Agustina Carando, Claudia H Sanchez Gutierrez, and Kenji Sagae. 2020. Developing nlp tools with a new corpus of learner spanish. In *Proceedings of the 12th language resources and evaluation conference*, pages 7238–7243.

Steve Graham, Karen R Harris, and Amber B Chambers. 2016. Evidence-based practice and writing instruction. *Handbook of writing research*, 2:211–226.

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

Barbara Kroll and Joy Reid. 1994. Guidelines for designing writing prompts: Clarifications, caveats, and cautions. *Journal of Second Language Writing*, 3(3):231–255.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

Cristóbal Lozano and Amaya Mendikoetxea. 2013. Learner corpora and second language acquisition. *Automatic treatment and analysis of learner corpus data*, 59:65–100.

Maria Nadejde and Joel Tetreault. 2019. Personalizing grammatical error correction: Adaptation to proficiency level and l1. *W-NUT 2019*, page 27.

Courtney Napoles and Chris Callison-Burch. 2017. 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.

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

Frank Palma Gomez, Alla Rozovskaya, and Dan Roth. 2023. 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, Dubrovnik, Croatia. Association for Computational Linguistics.

Charlene Polio and Margo Glew. 1996. Esl writing assessment prompts: How students choose. *Journal of Second Language Writing*, 5(1):35–49.

Guillermo Rojo, Ignacio Palacios, María Sampedro Mella, and Aurélie Marsily. 2022. Los corpus de aprendices de español le/l2: panorama actual y perspectivas futuras. *Journal of Spanish Language Teaching*, 9(2):174–189.

Claudia Sánchez-Gutiérrez, Barbara De Cock, and Nicole Tracy-Ventura. 2022. Spanish corpora and their pedagogical uses: challenges and opportunities. *Journal of Spanish Language Teaching*, 9(2):105–115.

Claudia Sánchez-Gutiérrez and Paloma Fernández-Mira. 2022. Datos longitudinales en corpus de aprendientes de español. In *Lingüística de corpus en español/The Routledge Handbook of Spanish Corpus Linguistics*, pages 374–387. Routledge.

Mark D Shermis, Jill Burstein, and Sharon Apel Bursky. 2013. Introduction to automated essay evaluation. In *Handbook of automated essay evaluation*, pages 1–15. Routledge.

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

Denise Paige Way, Elizabeth G Joiner, and Michael A Seaman. 2000. Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of french. *The Modern Language Journal*, 84(2):171–184.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. 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.

Abhisaar Yadav. 2022. [link].