Experiments on Automatic Error Detection and Correction for Uruguayan Learners of English

Romina Brown **Santiago Paez** **Gonzalo Herrera** **Luis Chiruzzo** **Aiala Rosá**
Instituto de Computación, Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay
{romina.brown,santiago.paez,gonzalo.herrera,luischir,aialar}@fing.edu.uy

Abstract

This paper presents an initial experiment on Grammatical Error Correction and Automatic Grading for short texts written by Uruguayan students that are learning English. We present a set of error detection and correction heuristics, and some experiments on using these heuristics for predicting the grade. Although our experiments are limited due to the nature of the dataset, they are a good proof of concept with promising results that might be extended in the future.

1 Introduction

The kinds of errors committed by students of English as a second language could be very different depending on their background, in particular depending on their L1, but also on the different geographical varieties of their language. For example, the cognates between L1 and L2 (De Groot and Keijzer, 2000), and the homophones between languages and varieties (Kochmar and Briscoe, 2014), influence the way students learn. This could have impact on Grammatical Error Correction (GEC) and Automatic Grading systems, which are often trained in standard corpora that are not adapted to model these geographical diversities.

In Uruguay, the universalization of English teaching throughout all primary schools is one of the objectives of the National Public Education Administration (ANEP). Together with the strategic goals of ANEP, the adoption of One Laptop per Child (OLPC) program, developed as the Ceibal project in Uruguay, improved the accessibility to English classes and resources throughout the country. Uruguay is a Spanish speaking country, its Spanish variety is called Rioplatense Spanish and is shared with some regions of Argentina. This variety presents some particularities that might influence the way students learn English.

In this work, part of a research line on developing tools for Uruguayan learners of English as a second language (Chiruzzo et al., 2022), we present the results of some preliminary experiments on creating automatic GEC and grading systems adapted to the particularities of Uruguayan learners. We use a dataset of short English texts produced by students as answers to an exercise. We analyze the types of errors committed, and design heuristics for detecting and correcting them automatically. Then we carry on experiments on automatic grading using this information.

This work has an important limitation, which is that the only information available in the dataset is the answer to one specific exercise. This implies that the results obtained for this exercise might not generalize to other contexts. In order to alleviate this problem, we try to focus on creating exercise independent features for grading, but we consider this should be taken as only a proof of concept and an initial exploration on the topic, and better datasets will be needed in the future. This is, as far as we know, the first work on GEC and Automatic Grading experiments that considers text produced by Uruguayan students.

2 Related Work

Grammatical Error Correction (GEC) is an active area of research in NLP, with shared tasks and competitions organized regularly. A series of GEC related shared tasks have been proposed together with CoNLL between 2011 and 2014, for example the CoNLL-2014 shared task (Ng et al., 2014) proposed detecting and correcting errors in English essays written by students. They use the NUCLE corpus (Dahlmeier et al., 2013), that contains 1,400 essays in English written by students of the

Romina Brown, Santiago Paez, Gonzalo Herrera, Luis Chiruzzo and Aiala Rosá. Experiments on Automatic Error Detection and Correction for Uruguayan Learners of English. *Proceedings of the 12th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2023)*. Linköping Electronic Conference Proceedings 197: 45–52.

National University of Singapore.

BEA 2018 Duolingo (Settles et al., 2018) shared task proposed to build systems that predict (not correct) the mistakes a learner will make in the future, given a transcript of exercises written by the same learner annotated with word level mistakes. It is interesting in that it includes the country the learner is from, which could be used to capture the L1 variabilty and geographic diversity.

The BEA-2019 Shared Task on Grammatical Error Correction (Bryant et al., 2019) included two tracks with two datasets: one with 3,600 manually annotated submissions from Cambridge Write & Improve platform, and another LOCNESS dataset with texts produced by native English speakers. Other important datasets include: the Cambridge Learner Corpus (Nicholls, 2003), that contains answers to English exams from Cambridge by students from all over the world, and its FCE subset (Yannakoudakis et al., 2011) with 1,244 annotated answers to the First Certificate in English exam; and the Lang-8 corpus (Mizumoto et al., 2012), with around a million English sentences annotated in a crowd-sourced way from the Lang-8 website[1]. These resources are generally written in a register that is much more complex than the texts we are dealing with in this work, which are texts written by schoolchildren, and most of them are just beginning to learn English.

The main approaches to performing GEC (Ailani et al., 2019) include using rule-based heuristics, classification methods, and machine translation based methods, with the last two approaches requiring a relatively larger set of annotated examples. The related task of Automatic Grading of essays is usually approached with machine learning methods, using a variety of features such as length of the text, POS or n-grams features (Yannakoudakis et al., 2011), different types of errors such as misuse of tenses or spelling (Ballier et al., 2019), or even the use of larger structures such as multi-word expressions (Wilkens et al., 2022).

3 Dataset and Error Analysis

The dataset we worked with is a corpus of answers written by Uruguayan schoolchildren to a writing exercise. In the exercise, students had to describe a person in a picture, together with her likes and dislikes shown as icons below the picture (see Fig. 1).

[1]https://lang-8.com/

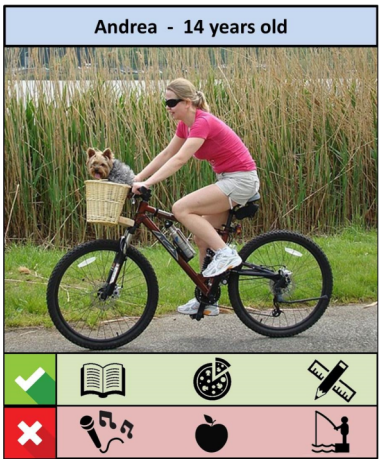

Figure 1: Picture associated to the exercise. The students had to describe the person in the picture, and her likes and dislikes.

This was part of an exam that was taken in 2017 by many schoolchildren from ages 9 to 11 that were learning English throughout the country. All short texts were graded by teachers following a rubric, with grades between 0 and 6, which roughly correspond to categories between A0 and B1 in CEFR.

There are 65,528 texts in total, but after filtering

Grade	Count	Example
0	13746	le gusta leer comer pipza y escribir lo que no le gusta es cantar comer fruta y pescar
1	11428	i like reading,pizza and rite. i don't like apple,to sing and fish his she andrea 14 years old
2	17699	she wears a pink shirt and jeans shorts he likes to ride a bicycle
3	10281	She has got a dog. She has got a glass in her face. She has got a bike. She drive in a bike. She like read and draw. She like eat pizza. She hate sing. She doesn't like eat apples.
4	1350	Andrea is 14 years old, she is a blondy and athletic girl. She is wearing a pink t-shirt, a white short and sunglasses. She is reading a bike whit her pet, a little dog. She likes eat pizza but doesn't like apples. She has a lot of books because she likes to read. Andrea studies from monday to friday. She doesn't like to fish because it's boring, she doesn't know how to sing
5	135	She is Andrea. She is 14 years old. She tall and thin. She has blonde, long hair. She is wearing white trainers, beige shorts, a pink blouse and sunglasses. She is riding a bike. She likes reading books, eating pizza and geometry. She doesn't like singing, eating apples and fishing. She has a pet. It's a dog. She loves it. She hasn't got a car. She can ride a bike but she can't fly. She gets up early, has breakfast and ride a bike. After that she has a bath and watch tv. Then she has lunch and goes to high school. After high school she goes to hockey classes. After she has a bath again, does her homework and goes to bed. She lives in a big house with his mother, father and sister. She loves her family and she is very happy.
6	13	She is Andrea, she is fourteen years old. She's wearing a pink t-shirt, and a short of jean She is riding her bike with her dog, she likes reading books, she likes eating pizza, and she likes maths. She doesn't like singing, eating apples and fishing She's got a dog but she doesn't have a cat. She doesn't look like a professional bike riding, and she isn't fat but she isn't thin. Her bike is brown and black and her dog is gray and brown, her dog is super cute, I want to be the owner of that dog, but her dog isn't like mine (...) mine is cuter than hers. She's got yellow hair and a black glasses, she is riding her bike in a quiet place, like in a countryside, behind her is a big lake.

Table 1: Example and number of texts for each grade in the corpus, after filtering empty texts.

empty and a few ungraded texts, we were left with around 54k texts. Table 1 shows a sample of each grade, and the total number of texts per grade in the corpus. The corpus is highly unbalanced, with an overwhelming majority of texts for the lower grades (almost half of them are graded with a score of 0 or 1) and only a few texts with the highest grades (less than 150 examples with grades 5 or 6). As can be seen in the table, lower graded texts tend to be shorter and have much more interference of Spanish words, while higher graded texts are significantly longer and contain more varied English vocabulary and structures.

3.1 Particularities of the sample

One interesting thing about this learners corpus is that it contains particularities of Uruguayan Spanish speakers trying to learn English. It has errors that Spanish speakers would make, but also errors that only speakers of Rioplatense Spanish would commit. Here is one example of an error in the dataset that any Spanish speaker could make:

*those ***hare** the things she does not like to do*

Because the letter "h" is silent in Spanish, misspelling *are* as **hare* could be expected, as they would sound homophonous from a Spanish perspective. However, consider the following example from the dataset:

****llor** green*

In this case, the writer intended to write about *green shorts*. Here we can see two errors: writing the adjective after the noun (as is the norm in Spanish grammar), and another mistake that is very particular to Rioplatense Spanish: The misspelling of *shorts* as **llor* responds to the fact that the "ll" digraph is pronounced /ʃ/, which is equivalent to the English "sh" sound.

Also note that these are two different types of spelling errors: in the latter case *llor* is a word that does not exist in English, so it could be captured by a dictionary search, but in the former case *hare* is a perfectly valid word in English which is invalid in that context.

3.2 Types of errors

We took two small subsets of the dataset containing samples of texts for the different categories, called the *development sample* and the *evaluation sample*. The development sample contains 53 texts, and was used to manually inspect the texts and mark all the different types of English spelling and grammar errors that could be found. Two researchers participated in this annotation: They split the development sample set and each researcher evaluated one subset, then they cross-checked their corrections, and finally they discussed the cases were there was disagreement to reach a final conclusion.

After this initial manual labeling of the texts, we compiled a list of common errors and their descriptions. This list was used by two other researchers to mark down the evaluation sample, comprised of 42 texts. Table 2 shows the different types of errors considered, and how many instances of them were found in the development sample and in the evaluation sample. We focused on the most prevalent errors found in the samples

Error	Example	Dev	Eval
Spelling	✗ **reding** ✓ reading	84	69
Subject-Verb agreement	✗ She **have** a dog ✓ She has a dog	42	15
Beginning of sentence caps	✗ **she** is Andrea ✓ She is Andrea	39	68
Use of pronoun	✗ She likes riding in your bike with **your** little dog ✓ She likes riding in her bike with her little dog	26	4
Verb form	✗ She likes **sing** ✓ She likes singing	24	41
Missing subject	✗ She has blond hair, is wearing a pink sweater... ✓ She has blond hair, she is wearing a pink sweater...	15	19
Proper noun caps	✗ She is **andrea** ✓ She is Andrea	15	5
Noun number	✗ She likes **apple** ✓ She likes apples	11	6
Use of determiner	✗ and **a** white trousers ✓ and white trousers	7	14
"I" caps	✗ **i** think she is... ✓ I think she is...	6	0
Adjective order	✗ She has a t-shirt **pink** ✓ She has a pink t-shirt	4	0
Contraction	✗ **doesnt** ✓ doesn't	3	0
Missing verb	✗ She 14 years old ✓ She is 14 years old	2	3
Wrong verb	✗ She **has** 14 years old ✓ She is 14 years old	2	10
Other errors	✗ Finally she goes to bed at 0:00 a.m. **clock** ✓ Finally she goes to bed at 0:00 a.m.	24	23

Table 2: Types of errors found in the development sample and the evaluation sample.

and tried to build heuristics for detecting and correcting them, as we will see in the following section.

4 Detection and Correction Heuristics

The proposed solution for error detection and correction comprises a series of modules that try to capture each type of error, but also need to interact with each other in order to improve the effectiveness of the process. For example, some of the NLP tools we use might not work too well with noisy text such as the one found in this dataset, so it is necessary to perform spelling correction first, before running the other modules. Each heuristic focuses on detecting one type of error, and also providing an appropriate suggestion for correction.

4.1 Spelling

We experimented with three widely used spellcheckers: Hunspell[2], the spellchecker used in open source systems like LibreOffice and the Mozilla suite which combines morphological analysis and pronunciation; Norvig's Spellchecker[3], based on Levenshtein distance search with dictionary filtering; and SymSpell[4], an improvement on Norvig's focused on speed and accuracy.

To capture particular errors like the ones mentioned in section 3.1, we made an adapted dictionary including common mistakes found in the texts. We tried using the different spellcheckers and combinations of them with a voting mechanism. Furthermore, we experimented with the use of BERT (Devlin et al., 2018) for predicting the correct word: We calculated the probability of each word suggested by the spellcheckers in the context of the text, using the `bert-base-uncased` model from Hugging Face.

Method	Acc
All spellcheckers with voting resolution	0.84
All spellcheckers with adapted dictionary	0.71
All spellcheckers with BERT resolution	0.74
Only SymSpell for detection and resolution	0.89

Table 3: Performance of the different methods used for spelling errors detection and resolution over the development sample set.

As shown in Table 3, out of the different combinations of models and tools we tested, the most accurate was using only SymSpell. It was also the fastest method, so we decided to use this tool for the rest of the experiments.

4.2 Capitalization

Note from Table 2 that there are three common errors related to capitalization, which involve not using an upper case in three cases: the beginning of a sentence, the pronoun "I", and proper nouns. The first two cases can be easily detected after sentence segmentation or finding the lowercase token "i", which is never used to refer to something different than the pronoun. However, the third case is more difficult, as the students could become creative and invent names and situations for this exercise. For example, one of the texts included the name "Paco" for the dog in the picture.

We used the Named Entity Recognition module by spaCy[5] to detect proper names. It does a good job when detecting common names used in English, like Andrea, but it failed to capture names or nicknames that are common in Spanish speaking countries, like Paco. In order to overcome this problem, we complemented the use of the NER module with a search in a list of names compiled from the Spanish National Institute of Statistics[6].

4.3 Subject-Verb agreement

In English, as well as in Spanish, the subject of a sentence and its verb must agree in number, and agreement errors are a very prevalent mistake in English learners. These errors could be easily spotted once we identify what the subject and the main verb are, which could be done using a syntactic parser, for example a dependency parser. However, consider the following text from the corpus, where the expected analysis would be the root verb *like* with the subject *she*:

She ***like** pizza*

Parsers work best when the analyzed text is clean and well written, and this is of course not the case with these texts. The spaCy dependency parser for this example considers *like* as a SCONJ, so it fails to detect it as the root of the sentence. Similar errors occur frequently with noisy texts, so a solution based on a pre-trained parser seems not feasible, although other attempts at solutions

[2]http://hunspell.github.io/
[3]https://norvig.com/spell-correct.html
[4]https://github.com/wolfgarbe/SymSpell
[5]https://spacy.io/
[6]https://www.ine.es/

based on parsing exist, like capturing wrong parses using *mal-rules* as in (Da Costa et al., 2016).

In our case, given the simplicity of the texts, we opted for a different strategy. We use rules for detecting the likely subject and main verb of the sentence: pronouns and proper nouns at the beginning of the sentence are likely subject candidates, followed by verbs that belong to a list of 1000 common verbs for English learners[7] (Turnbull et al., 2010).

We split verb forms in categories according to their inflection, then we experimented with two strategies for agreement error detection: in the first one, inspired by (Gehman et al., 2020), we use BERT to calculate the probability of the verb form used and the alternative ones; the second one, inspired by (Wang and Zhao, 2015), uses a lexicon, POS-tagging and morphology for checking agreement considering pronouns, nouns, verbs, and auxiliary constructions like negations.

Table 4 shows a comparison of both approaches on the development sample. The rules and lexicon approach, although simpler, beats the BERT method on the three considered metrics.

Method	Prec	Rec	F1
BERT	0.77	0.73	0.75
Rules and lexicon	0.82	0.76	0.79

Table 4: Performance of the different methods used for subject-verb agreement errors detection over the development sample set.

4.4 Verb form

Errors in the use of verbal forms are very common when learning English, when students must learn how to use different tenses, particularities of irregular verbs, agreement and the use of infinitives and gerunds in other constructions. The two most frequent errors found in the development sample were subject-verb agreement issues (seen in the previous section) and confusion between infinitive and gerund forms.

We considered our set of 1000 common verbs and their corresponding forms, and wrote a series of manual rules based on (Swan and Walter, 2011) that cover different situations such as: the use of verbs after adjectives, prepositions, accusative pronouns, and verbs that require a specific form. Special care had to be taken when dealing with the issue of parallelism of a construction when used in conjunctions. For example, consider the following sentence:

She likes ****eat*** *pizza, walk at night and* ****singing***.

In this case, our heuristic indicates that the verb form after "likes" should be "to eat", then the use of the verb "walk" is correct, but the verb "singing" should also be changed to "sing".

4.5 Use of determiners

There are two types of errors involving the use of determiners: they are either omitted, or included unnecessarily (wrong use). The heuristic in this case involves using the POS-tagger and morphological analyzer from spaCy to check cases of nouns with or without determiners, and using a series of rules for deciding if the use of determiner is correct. For example, plural nouns should have a plural determiner, or none in some constructions, while singular nouns could use a singular determiner depending if they are countable or not. When a missing determiner is found, the heuristic always suggests including the indefinite article ("a" or "an"), so a pronunciation dictionary[8] is used to tell apart nouns which start with vowel sounds (e.g. "an umbrella" vs. "a unicorn").

4.6 Results in sample sets

Table 5 shows the results of our heuristics over the development and evaluation samples. Note that during the development of the detection and correction heuristics, we used the information obtained by manually annotating the development sample, but the evaluation sample was not seen until later. Nonetheless, the results obtained for the evaluation sample are very similar, which gives us some confidence on how good the heuristics are for capturing the errors in the whole dataset.

	Development			Evaluation		
Error	Prec	Rec	F1	Pre	Rec	F1
Spelling	0.89	0.88	0.88	0.81	0.85	0.83
Caps - "I"	1.0	1.0	1.0	-	-	-
Caps - BoS	0.99	1.0	0.99	0.92	0.79	0.85
Caps - Proper noun	0.73	1.0	0.84	0.75	1.0	0.86
Subject-Verb agreement	0.82	0.76	0.79	0.83	0.77	0.80
Verb form	0.73	0.91	0.81	0.66	0.81	0.72
Determiner - Missing	0.71	0.87	0.78	0.50	0.81	0.62
Determiner - Wrong	0.67	0.67	0.67	0.38	0.75	0.5

Table 5: Results of the error detection heuristics over the development and the evaluation sample sets.

[7] Oxford University Press. Oxford 5000 wordlist, aug 2020. https://www.oxfordlearnersdictionaries.com/us/wordlists/

[8] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

5 Automatic Grading Experiments

After creating the set of heuristics to capture many of the errors committed by the students, we wanted to assess how useful this information would be for predicting grades given by teachers. These grades were assigned following a rubric that takes into account many aspects, including the use of English or Spanish, the production of single words or phrases, the types of errors committed, the general readabilty and soundness of the text, etc. It was interesting to see if our simpler heuristics would provide sufficient information to at least roughly predict the grade. We first split the whole dataset into 70% for training, 15% for development and 15% for test (note that these are different splits than the samples described in section 3.2).

Due to the high imbalance in the dataset, we decided to cluster some grades into ranges. Grades 0 and 1 correspond to the *low* range, 2 and 3 to the *medium* range, and 4 through 6 to the *high* range. Although this does not completely fix the balance problem, by manually inspecting the texts we found these ranges left more homogeneous texts in each category. We will present results both for grade ranges and separate grades.

We ran a baseline experiment where we used bag of words and bag of bigram features. A model trained with these features would of course be highly tailored for grading this particular exercise, and would probably not generalize well to other prompts. For example, some of the most relevant BoW features found in this experiment included "Andrea", "pizza", and "14". However, we have two main motivations for these experiments: we wanted to know how likely it is to create a classifier that would emulate the grades given by teachers, and at the same time we wanted to find out if it is possible to create a classifier that works similarly but is not overfit to the specific words of this exercise.

5.1 Features and models

We trained different classifiers using different combinations of features. As mentioned before, we used BoW features, which in our case were the 750 most frequent unigrams and bigrams.

We also included one feature for each of the heuristics described in section 4, called the "correction features". The feature value is the number of errors the heuristic found for a particular text. So we have eight features counting the number of:

- spelling errors
- beginning of sentence capitalization errors
- pronoun "I" capitalization errors
- proper noun capitalization errors
- verb form errors
- subject-verb agreement errors
- missing determiner errors
- wrong determiner errors

The rationale behind the use of these features is that, if we could capture all the errors in a text, this information could help a classifier predict a grade, even when not knowing the actual words of the text. This would decouple the classifier from the prompt of the exercise and be more generalizable.

We also used a feature indicating length of the texts in tokens. This is because, as mentioned in section 3, the length of the text seems to be correlated with the grading. This could pose a problem for an automatic grading system, because it could learn that just producing a longer text would yield a better grade. However, we must also consider that when students produce longer texts they might also be introducing more errors, which could be captured by the heuristics. Of course further experiments would be needed to validate this, and it is out of the scope of this work.

All the classifiers we trained are from the `scikit-learn` suite of machine learning tools (Pedregosa et al., 2011). We experimented with Naïve Bayes (NB), Random Forest (RF), Maximum Entropy (ME), Support Vector Machine (SVM), and Multi-Layered Perceptron (MLP) classifiers.

5.2 Results

The three rounds of experiments include: using the BoW features, using only the correction features plus the length feature, and using all the combined features. Table 6 shows the results of these experiments over the test partition. The best performing classifiers are the RF model and the ME model when using all the combined features. This is expected, as using all the features provides a lot of information. However, note that the MLP and ME models with only correction and length features, although not perfect, have a performance

	BoW		Correction features + length					Combined features				
	RF	ME	NB	RF	ME	SVM	MLP	NB	RF	ME	SVM	MLP
All grades Acc.	0.67	0.62	0.48	0.56	0.59	0.59	*0.60*	0.44	**0.68**	0.63	0.33	0.32
All grades M-F1	0.48	0.40	0.32	0.37	0.35	0.36	*0.37*	0.29	**0.49**	0.41	0.12	0.08
Ranges Acc.	0.83	0.83	0.73	0.79	0.82	0.82	*0.82*	0.70	**0.86**	0.84	0.51	0.51
Ranges M-F1	0.74	0.70	0.61	0.64	0.64	0.63	*0.68*	0.61	**0.76**	0.71	0.22	0.23

Table 6: Results of the classifiers over the test set.

that is at least comparable to the top ones. This is important, because these classifiers do not use any information on the specific words of the exercise, which gives us hope that this strategy could be used to grade similar writing exercises but with other prompts. Of course, more experiments are needed to validate this with other datasets.

6 Conclusions

We presented an initial experiment on building heuristics for detecting and correcting grammatical errors in texts by Uruguayan learners of English, and then training a classifier to predict a grade to assign to those texts. The heuristics have good performance in capturing common grammar errors like spelling, capitalization, and subject-verb agreement. Our best classifier has 82% accuracy and 76% macro-F1 for separating the texts in three ranges according to grade. We found that using only features that are independent from the exercise text the performance of the classifier gets to 82% accuracy and 68% macro-F1. This is a significant drop, but we must consider that this classifier could be adaptable to other exercises as well.

This is only a proof of concept, as we are aware that it is very difficult to build a generalizable system with examples of only one exercise. There are many ideas for future work about how to improve these heuristics and make them useful in a broader context. We would like to try using a language model to produce a representation of the text that could be comparable to a set of reference texts, and measure the distance between them. Also, we could try to use positive and negative lists of words that the text should have, and create features that would be adaptable to other exercises (in this case the list would include “Andrea”, “girl”, “read”, “bike”, etc.). Another interesting research direction is trying to assess the number of texts it would take to manually grade in a corpus, so we can finetune a system that has at least a good estimate of the grades for the rest of the corpus.

We are now in the process of building a better dataset for working on these and related problems. We want to create a more varied corpus with several exercise prompts and several example answers written by Uruguayan students of English, manually corrected and graded by teachers. This dataset would help us test and compare our current heuristics and other correction methods more thoroughly.

Acknowledgements

The dataset we used in this work was created by Ceibal en Inglés, part of the Ceibal project[9]. We want to thank them for letting us use it for research purposes.

References

Sagar Ailani, Ashwini Dalvi, and Irfan Siddavatam. 2019. Grammatical error correction (gec): research approaches till now. *International Journal of Computer Application*, 178(40):1–3.

Nicolas Ballier, Thomas Gaillat, Andrew Simpkin, Bernardo Stearns, Manon Bouyé, and Manel Zarrouk. 2019. A supervised learning model for the automatic assessment of language levels based on learner errors. In *Transforming Learning with Meaningful Technologies: 14th European Conference on Technology Enhanced Learning, EC-TEL 2019, Delft, The Netherlands, September 16–19, 2019, Proceedings 14*, pages 308–320. Springer.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. 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, Florence, Italy. Association for Computational Linguistics.

Luis Chiruzzo, Laura Musto, Santiago Góngora, Brian Carpenter, Juan Filevich, and Aiala Rosá. 2022. Using nlp to support english teaching in rural schools. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 113–121.

[9]https://ceibal.edu.uy/

Luis Morgado Da Costa, Francis Bond, and Xiaoling He. 2016. Syntactic well-formedness diagnosis and error-based coaching in computer assisted language learning using machine translation. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 107–116.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 22–31.

Annette MB De Groot and Rineke Keijzer. 2000. What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language learning*, 50(1):1–56.

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

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Ekaterina Kochmar and Ted Briscoe. 2014. Detecting learner errors in the choice of content words using compositional distributional semantics. Association for Computational Linguistics.

Tomoya Mizumoto, Yuta Hayashibe, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2012. The effect of learner corpus size in grammatical error correction of esl writings. In *Proceedings of COLING 2012: Posters*, pages 863–872.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. 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, Baltimore, Maryland. Association for Computational Linguistics.

Diane Nicholls. 2003. 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. Cambridge University Press Cambridge.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. 2018. Second language acquisition modeling. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 56–65.

Michael Swan and Catherine Walter. 2011. *Oxford English grammar course*. Oxford University Press Oxford.

Joanna Turnbull, D Lea, D Parkinson, P Phillips, B Francis, S Webb, V Bull, and M Ashby. 2010. Oxford advanced learner's dictionary. *International Student's Edition*.

Yuzhu Wang and Hai Zhao. 2015. A light rule-based approach to english subject-verb agreement errors on the third person singular forms. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters*, pages 345–353.

Rodrigo Wilkens, Daiane Seibert, Xiaoou Wang, and Thomas François. 2022. Mwe for essay scoring english as a foreign language. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 62–69.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189.