Using bigrams to detect written errors made by learners of Spanish as a foreign language

Miguel Blázquez Carretero (mblazquez@up.edu.ph)
University of the Philippines Diliman, Philippines

Abstract

Based on previous satisfactory experiences generated by Grammar Checker (UNED, 2017), a spell and grammar check package that figured among the finalists of the ELTons awards 2016, the *Universidad Nacional de Educación a Distancia* (UNED) is developing a Prototype of Grammar Checker (PGC) specifically designed to correct grammatical errors committed by learners of Spanish as a Foreign Language (SFL). The PGC relies on a corpus of reference of 100 million words and uses Sinclair's well-known logarithm to detect errors. Such a logarithm, analyses words-pairs (bigrams). Since errors made by native speakers often differ from those made by Second Language (L2) learners, this study´s core objective was to find out where to establish the thresholds that this software will use to locate incorrect bigrams and to highlight them differently depending on their probability of being an error. To do so, each bigram found in a sample of 21 compositions written by L2 learners of SFL was first analysed. Three thresholds were provisionally recommended for the PGC depending on the bigram's frequency and its probability to occur randomly. Then, the capacity of these thresholds to detect grammatical errors was later tested using another sample of 21 compositions. Results show that bigrams are a powerful tool to detect L2 learners' grammatical errors. In word-pairs that do not usually occur together ($R \leq 0.1$), the threshold has an accuracy of 90%. These results draw attention to the importance of using real data to better adapt learning tools to L2 learners needs.

Keywords: Grammar checker; Bigram filter; Detection thresholds; Grammatical and spelling mistakes; Self-correction; Spanish as a Foreign Language.

Introduction

Writing accurately is important. Errors influence the reader's perception of the quality of the ideas written (Kiuhara, Graham & Hawken, 2009), may render a text difficult to read, thus distracting the audience from the intended message (Graham & Santangelo, 2014) and can potentially impede meaning (Wilcox, Yagelski & Yu, 2013). Moreover, it is considered essential for L2 learners to identify errors not only because it is the predominant method of evaluating one's writing skills but also because written errors might impair exam performance (Graham, Harris & Herbert, 2011). The Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) came up with different levels to trace learner's accuracy in written production. The following are the three out of six levels: B1 – "accurate enough to be followed most of

the time"; B2 – "show signs of mother tongue influence"; C1 – accurate, apart from occasional slips of the pen".

According to Heffeman and Otoshi (2015), one of the best ways to improve spelling and grammar accuracy is to write as often as possible, so that common errors and mistakes will be corrected. However, a lack of time and resources dedicated to correcting student compositions limits the opportunity for feedback (Lawley, 2015). Experience suggests that around 15 minutes is required to correct and provide feedback in a free-form composition written by an L2 learner at the B1 level CEFR (Council of Europe, 2001). At the UNED, a distance-learning university in Spain, only four hours of teaching time is allocated per week to students of language courses; students are thus encouraged to self- and peer-correct. This process is guided by teachers who provide detailed instructions on how to self-correct effectively. Even though several studies suggest that self-correction can be efficacious (Lázaro Ibarrola, 2013; Quinn, 2014; Vickers & Ene, 2006), there are still difficulties encountered by the students when teachers are not available (Lawley, 2016). As Lee (1997) points out, recognition of errors is crucial to self-correction, which can be tough for L2 learners if they are not emphasized. Unlike native speakers who have the ability to effortlessly find and correct their errors, it is less intuitive for L2 students since these learners lack knowledge in the accepted rules of a language.

In line with that, a software able to detect spelling and grammar errors would be convenient for L2 learners. Furthermore, it might aid the students in self-correcting properly. There are mainstream spell and grammar check packages, such as Microsoft Word. However, they have several limitations, including auto-correction, misdiagnoses, and missing errors, especially in compositions of low-proficiency students (Blazquez & Fan, 2019; Heift & Rimrott, 2008).

Background to the study

Proofreading software

In order to detect and correct grammar and spelling errors in a written text, editing packages use spell and grammar checkers.

Spell checkers usually use a corpus as a reference to assess whether they are written correctly. If words do not appear in the corpus, they are highlighted followed by either a list of alternatives or a correction done by the software (Mitton, 2010). The detection rate of this method is higher than 85%; however, it only detects single-word errors (unigrams) (Blazquez & Fan, 2019). Mainstream spell checkers are mostly built for native speakers. Since their main objective is to facilitate and speed up a user's writing, spelling errors are automatically corrected. This is beneficial for proficient writers whose errors are mostly due to failure on their part to use a known rule system correctly (performance errors or mistakes [Brown, 1994]). In the case of L2 students whose majority of errors are due to actual gaps in their knowledge of the target language (competence errors [*ibid*.]), auto-correct does not prove as practical. They would not be able to learn from their errors since the software instantly modifies their mistakes, hence impeding them from noticing, and subsequently learning, from these errors (Schmidt, 1990). Moreover, when these mainstream spell checkers detect an error, they suggest a list of replacement words. While the native speaker, for whom the spelling error is a performance error, might have no

problems in selecting the right word from the list, it is less intuitive for L2 learners. There is sufficient evidence that L2 students are often misled by these suggestions because they usually presume that the correct word is amongst the suggested ones (Heift & Rimrott, 2008), although this is not often the case (Mitton & Okada, 2007).

A different approach is needed to detect grammatical errors. Individually looking at words is not effective since it cannot identify misplaced words (e.g. **were are you?*). To detect and correct grammatical errors, grammar checkers may opt to use three different strategies: compare erroneous sequences, evaluate grammatical sequences, and analyse the words in context (Wu, 2014). The first error detection method analyses the syntax of written sentences by comparing them with a database containing frequent sentences and errors (Thurmair, 1990). The second one, known as "tagging and parsing", assigns a grammatical function to each word and constructs a statistical language model based on the linguistic probability of certain grammatical sequences. For instance, if a word has been labelled as an article, there is more probability that it would be followed by a noun or adjective than by a verb (Chen, 2009). However, as Lawley (2015) points out this method misdiagnoses errors such as **I'm happy living hear* where the L2 learner has confused the homophones *here* and *hear*. This is because the tagger labels *living* as a subject and *hear* as a verb and the parser suggests that there is a problem of agreement between them. Finally, grammar checkers can analyse how frequency words tend to appear together (collocations), which also allows detecting punctuation errors. This is achieved by separating the text into sequences of words (n-grams) and statistically comparing their frequency with those sequences in correctly-written texts (San Mateo, 2016). For example, if the text contains the sentence **She love me*, the sentence would be divided into two segments: **she love* and *love me*. Each of these segments of two words (bigrams) will be searched in the corpus. While *love me* is likely to occur many times in the corpus, the sequence **she love* would hardly appear and, subsequently, it will be highlighted for its probability of being an error. Although the designers of this type of software use several of these strategies together, it seems that the latter (n-grams) is the most reliable (Briscoe, Medlock & Andersen, 2010; Harvey-Scholes, 2018) and most used recently by main designers of this type of software, such as Microsoft, Google, or Grammarly (Wu, 2014).

Building a grammar checker for SFL learners

To determine if the same approach might be used in the case of Spanish, a PGC based on bigrams was built at the UNED. It has a reference corpus of 100 million words of contemporary Spanish. As explained in San Mateo (2016), this corpus was made using only written texts since speech has features like false starts, backtracking, self-corrections, and interruptions which are inappropriate models for most modes of writing. Only the current usage —post '80s— was used. Poetry and drama, which sometimes imitate features of spoken Spanish, were also excluded. General texts, both fictional and non-fictional, were favoured over the technical ones. All kinds of articles were taken from contemporary newspapers and magazines, predominantly from 2012. Texts written in dialects or non-standard Spanish were avoided. Standard Spanish —peninsular and Latin-American— were preferred, since this type of Spanish is most used by the students (San Mateo, 2016).

The algorithm that the bigram filter uses to detect grammatical errors (Figure 1) is the same one used by Sinclair (1991), which compares the frequency in the reference corpus of each bigram (Fab) and the probability that these two words (a & b) occurs together at random considering the frequency of each word separately (Pab). If the frequency of the bigram written by the writer is lower (in the reference corpus included in the software) than the probability that these words occur randomly, it is probably an error (San Mateo, 2016). For example, in the sentence **yo eres guapo* (**I are handsome*) the bigram **yo eres* (**I are*) occurs only once (Fab = 1) in the corpus of 100 million words (Cn = 100,000,000). However, the words *yo* (I) and *eres* (are) are very frequent, since *yo* appears 226,946 (Fa) times and *eres* 27,427 (Fb) times in the reference corpus. This means that one out of every 441 words is the personal pronoun *yo* (100,000,000 / 226,946 = 441) and one out of every 3,646 is the verb inflection *eres* (100,000,000 / 27,427 = 3,646). Therefore, the chances that they might appear consecutively at random in the corpus of 100 million words equate to 62.2 (100,000,000 / (440.6 * 3,646) = 62.2). Due to their individual high frequency, by pure probability, both words should appear together many times, specifically 62. However, between 100 million words only appear together on one occasion. If both figures were compared (1 / 62.2 = 0.02) it will give a ratio (R) of 0.02 which indicates that it is most likely an error. As San Mateo (2016) indicated, the ratio is an indicator of how likely two words attract each other. A ratio greater than one (R > 1) suggests that these words show a tendency to be used together in that precise order. The more they tend to occur together, the greater the number. A ratio below one (R < 1) suggests the opposite: that those words do not usually appear together in correct Spanish; and the greater the degree of rejection, the smaller the number. A ratio of precisely one (R = 1) means that words usually appear together in this order as chance would predict. So, in the case of the bigram *yo eres*, a ratio of 0.02 indicates that this segment occurs much less than would be expected at random and, consequently the PGC will highlight it. On the contrary, in the case of the bigram *eres guapo* (*are handsome*), the bigram occurs 7 times in the corpus of 100 million words; the word *eres* appears 27,427 times in the corpus and the word *guapo* (*handsome*) appears 1,437 times. If the words were randomly distributed in the corpus of 100 million words, considering the individual frequency of both words, the bigram *eres guapo* would occur 2.54 times, but, in fact, it occurs 7 times. As explained above, the ratio is an indication of attraction. In this example, *eres guapo* has a threshold value of T = 2.76; this is above one, which is what chance could predict. The fact that *eres guapo* appears 2.76 times more than what chance predicts is a strong indication not only that it is correct, but is a common bigram in Spanish and, therefore, will not be highlighted.

$$R = \frac{Fab}{Pab} \qquad \text{being} \qquad Pab = \frac{Cn}{\left(\frac{Cn}{Fa}\right) x \left(\frac{Cn}{Fb}\right)}$$

Figure 1. Algorithm used by the bigram filter of the PGC (Sinclair, 1991 in San Mateo, 2016)

However, not all low ratio bigrams are necessarily wrong. For instance, in the sentence *mi hermano pequeño tiene pecas* (my younger brother has freckles), which

numbers are shown in Table 1, the pair *pequeño tiene*, despite having a low pair frequency (9) and a low ratio (0.58), is not actually an error.

Table 1.
Analysis of the Fab, Pab and Ratio of *mi hermano pequeño tiene pecas*

Pair of words	(Fab) Pair frequency	1st word frequency	2nd word frequency	(Pab) Pair probability	(R) Ratio
Mi hermano	3,932	271,588	20,959	56.92	69.08
hermano pequeño	110	20,959	22,482	4.71	23.35
pequeño tiene	9	22,482	69,592	15.65	0.58
tiene pecas	7	69,592	306	0.21	33.33

One solution to this problem would be, as Lawley (2015) suggests, is to establish different thresholds depending on the frequency and the ratio. According to him, these bigrams need to be highlighted differently so that students can prioritise those pair of words that are more likely to contain grammatical errors.Therefore, the PGC will use three colours to highlight the bigrams that are incorrectly written (yellow, orange and red) depending on how frequent the bigram is. This study aims to establish the thresholds for each frequency range or colour. Since the PGC's main objective is to help L2 learners in Spanish writing, the thresholds must be in line with their needs. Because of that, it is compulsory to examine real sample of SFL interlanguage to investigate the aim of this study: which thresholds proved most accurate before coding the software to guarantee its efficacy?

Method and analysis

Research design

A preliminary study was carried out to establish the potential thresholds for the bigram filter based on the correlation between the ratio of the bigrams and the percentage of being an error. Another study was then conducted to measure the effectiveness of the established thresholds. This investigation made use of 42 compositions, written by 20 different English L2 learners of Spanish at level B1 of the CEFR, selected randomly from the corpus CORANE (Cestero Mancera y Penadés Martinez, 2009). This corpus is a collection of written texts written in Spanish by learners of SFL with levels ranging from A2 to C1 of the CEFR (Council of Europe, 2001).

Procedure and data analysis

Preliminary phase: establishing the thresholds of the bigram filter of the PGC.

For the preliminary study, a sample of 21 compositions written by 10 English learners of SFL with levels ranging from A2 to B2 of the CEFR was analysed to obtain data for each of the bigrams in the compositions that had an expected ratio of 2.0 or less. The hypothesis

was that such bigrams, which are not used or are used relatively sparsely by educated native speakers of Spanish, are more likely to contain errors than bigrams which educated native speakers of Spanish often use.

The ratio of each bigram, given by the total occurrences of each bigram in the PGC correct Spanish corpus (Fab), and the expected occurrences of the bigram if the words were distributed randomly (Pab) were recorded (as shown in Table 1). When this data was collected, it was noted whether the bigrams were an error in the context in which they were used. The data was analysed by first grouping the low ratio bigrams (ratio of less than 2.0) into categories according to the size of the ratio, and then plotting a histogram with these ratio categories along the x-axis and the proportion of these bigrams that were errors plotted on the y-axis (Figure 3).

To see the number of occurrences that produced significantly higher predictive power and, at the same time, overlooked a small number of errors (see Figure 4), several histograms were then plotted. To do so, a trial-error procedure at intervals of 25 (Fab = 25) was followed. Considering that if a bigram appears in a corpus of correct Spanish numerous times, then it is probably a correct sequence, all the bigrams with a total occurrence above 150 (Fab > 150) were ignored.

On the other hand, a different approach was required to deal with bigrams that had no occurrences in the PGC corpus of reference, as they all had a ratio of 0, but varied wildly in terms of the expected occurrences. After all, it is not unexpected to find that a bigram which has an expected frequency of less than one does not occur, and this is not necessarily indicative of an error. For example, the bigram *demasiado extrovertido* (too extroverted), which is correct, has a Pab of 0.01; this is because the adjective *extrovertido* is quite infrequent in Spanish (20 appearances in the reference corpus of the PGC). Even if the adverb *demasiado* is frequent (44,295), due to the scarce number of appearances of *extrovertido* when the Pab was calculated (see Figure 1) the denominator becomes relatively large [if Pab = 100,000,000 / (100,000,000 / 44,295) + 100,000,000 / 20) → 100,000,000 / (2,258 * 5,000,000) = 100,000,000 / 11,290,000,000], reducing the results significantly (Pab = 0.01). Therefore, a histogram was plotted displaying the expected occurrence of a bigram along the x-axis (to the nearest integer), and the proportion of bigrams that were an error on the y-axis (Figure 5).

Second phase: measuring the efficacy of the thresholds of the bigram filter of the PGC.

As a result of the preliminary study, a set of thresholds were established and programmed in the PGC. To measure the effectiveness of these thresholds, a second study was carried out. Subsequently, further 21 compositions from the CORANE corpus written in Spanish by another 10 English students of SFL at level B1 of the CEFR, were introduced in the PGC (See Figure 2).

Figure 2. PGC analysis of a fragment of a text written by an L2 learner of Spanish

The bigrams highlighted in different colours by the PGC were analysed in their linguistic context to corroborate if, in fact, they contain a grammatical error or not. For instance, as shown in Figure 2, the bigram highlighted in yellow (23%-39% of chances of error) *de democracia* (of democracy) was analysed in its linguistic context *la difusión de democracia en este siglo* (the dissemination of democracy in this century) to determine if an error is actually present. There is, in fact, one since a definite article is needed before *democracia*; hence the correct sentence should be *la difusion de **la** democracia en este siglo*. Similarly, it was seen that both bigrams highlighted in red (87% of chances of error according to the preliminary study) were actually errors of gender agreement. Once done with this, the results were then compared with the ones obtained on the preliminary study, in order to test if the previously established thresholds were accurate in detecting grammatical errors or they needed to be modified.

Results

Establishing the thresholds of the bigram filter of the PGC

The 21 compositions contain 2,958 words. There were 887 bigrams in these 21 compositions which had a ratio of 2 or less. Plotting a histogram with these ratio categories along the x-axis and the proportion of these bigrams that were errors plotted on the y-axis, immediately provided evidence to support the idea that the ratio is a good predictor for the probability of a bigram being an error. As can be seen in Figure 3, a ratio between 0 and 0.2 ($0 \leq R \leq 0.2$) means there is a 58% chance of finding an error in a bigram. The percentage increases to 87% when $R \leq 0.1$.

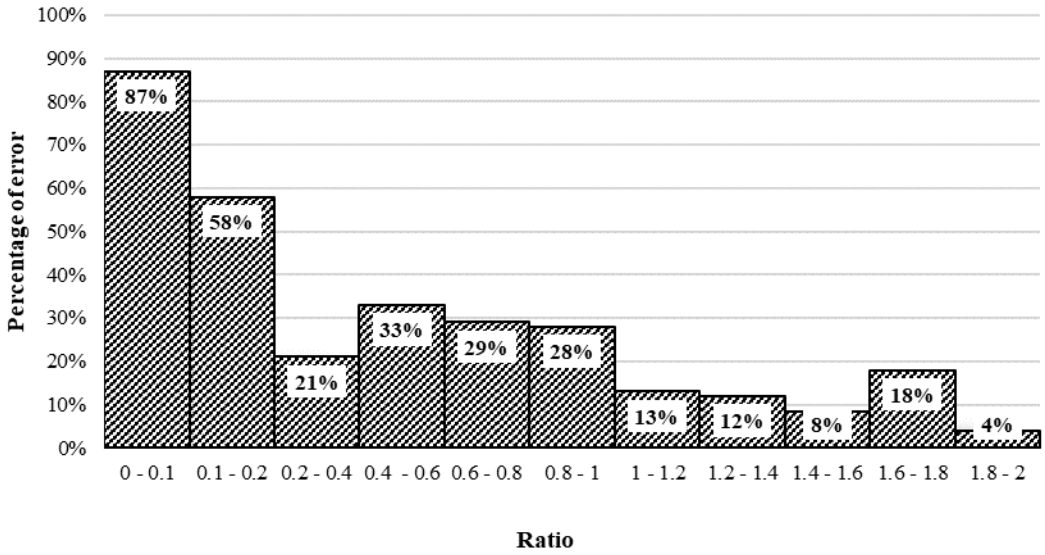

Figure 3. Probability that a bigram contains an error based on its Ratio.

It was found that if the bigrams that occur often in the PGC's reference corpus of correct Spanish are ignored, the ratio becomes a more powerful predictor of grammatical errors, as shown in Figure 4, compared with the previous histogram. Results show that ignoring all bigrams above 75 total occurrences yielded the best results, as it produced significantly higher predictive power whilst overlooking only a small number of errors (36 from over 200).

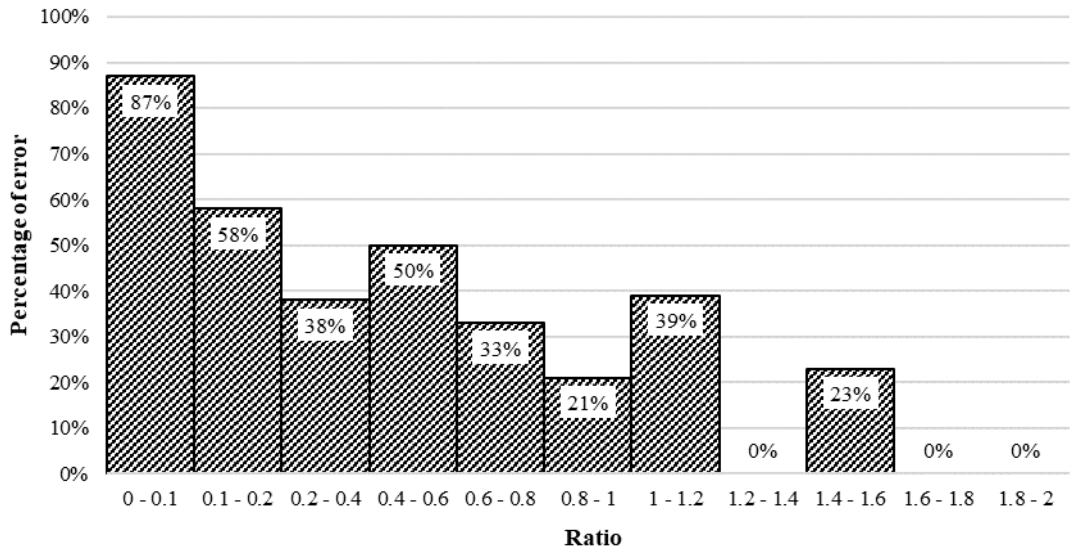

Figure 4. Probability that a bigram, which appears ≤ 75 in the PGC's corpus contains an error based on its Ratio.

As specified in the method section, to analyse the bigrams that had no occurrences (Fab = 0 → R = 0) in the PGC corpus of reference, another approach was needed, since a Pab < 1 is not necessarily indicative of an error. Subsequently, to analyse the probability that a bigram contains an error based on its Pab, a histogram was plotted displaying in the x-axis the expected occurrence of a bigram (to the nearest integer), and the percentage of bigrams that were an error on the y-axis (Figure 5). From the graph, a higher expected frequency translates to a higher probability of being an error. The graph below demonstrates how an expected frequency of 4.0 suggests a drop in the proportion of bigrams that were classified as errors, from 88% to a range of 23% to 46%.

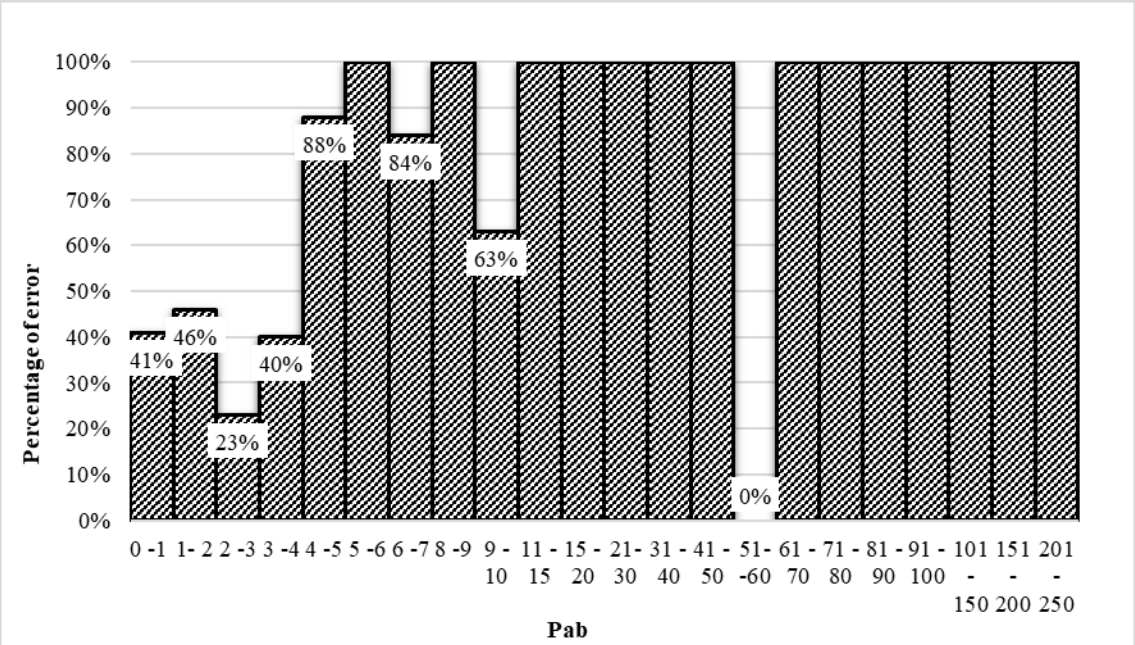

Figure 5. Probability that a bigram contains an error based on its Pab.

As a result of this preliminary study, the following specifications were provisionally recommended for the PGC:

- If there are no occurrences in the corpus for a bigram (R = 0), and when the number of expected occurrences is less than 4.0, the bigram should be highlighted in orange, meaning that the probability of error is below 46%. On the other hand, when the number of expected occurrences exceeds 4.0, the bigram should be highlighted in red, which would raise the probability of finding an error to 88%.
- On the contrary, if the bigram appears in the corpus (R > 0), the PGC should ignore any entries with more than 75 occurrences. For those entries with fewer than or equal to 75 total occurrences, the PGC should highlight the bigram in red when ratio is less than 0.1, which means an 87% chance of finding an error, in orange when ratio is between 0.2 and 0.6, which means a probability of an error between the 38% and the 58%, and in yellow when ratio is between 0.7 and 1.6, meaning that the probability of finding an error is between 23% and 39%.

It was apparent that the PGC highlights the infrequent bigrams in different colours depending on their probability of being an error to indicate to L2 learners which errors

they should prioritise. Providing indirect feedback where errors are simply singled out or highlighted somehow has been proved an effective L2 learning strategy (Van Beuningen, De Jong, & Kuiken, 2012) as well as colour-coding (Dziemianko, 2015; Kang & Lee, 2013).

Measuring the efficacy of the thresholds of the bigram filter of the PGC

To measure the effectiveness of the thresholds, an analysis of 21 compositions written in Spanish by 10 different English learners of SFL at level B1 of the CEFR was conducted. These compositions were selected randomly from the CORANE corpus, considering the proficiency level of the students and their First Language (L1). The sample of 21 compositions consisted of 3,680 words, in which 328 were found to be incorrect sequences in correct Spanish. A total number of 532 bigrams were highlighted by the PGC. Table 2 shows that 155 were highlighted in red, of which only 14 were not actually incorrect, suggesting that there is an effectiveness rate of 91% for the first threshold. Out of the 307 that were highlighted in orange, 152 were incorrect sequences (49.5%). Finally, 70 bigrams were highlighted in yellow, which means that their ratio was between 0.7 and 1.6. Only 21 of those were an error, which means a probability of error of 30%.

Table 2.
Percentage of bigrams highlighted by colour by the PGC that were incorrect sequences

	RED				ORANGE				YELLOW			
	Bi-grams	No error	Error	% accuracy	*Bi-grams*	No error	Error	% accuracy	*Bi-grams*	No error	Error	% accuracy
B1 062	5	0	5	100%	10	5	5	50%	1	1	0	0%
B1 063 (text 1)	0	0	0	----	13	9	4	31%	4	3	1	25%
B1 063 (text 2)	1	1	0	0%	11	10	1	9%	4	3	1	25%
B1 064	13	1	12	92%	16	4	12	75%	1	1	0	0%
B1 065	11	2	9	82%	31	7	24	77%	3	1	2	67%
B1 066	17	1	16	94%	19	7	12	63%	6	5	1	17%
B1 067 (text 1)	2	0	2	100%	11	8	3	27%	4	4	0	0%
B1 067 (text 2)	7	1	6	86%	17	15	2	12%	4	2	2	50%
B1 068 (text 1)	7	0	7	100%	9	1	8	89%	0	0	0	----
B1 068 (text 2)	10	0	10	100%	17	4	13	76%	5	1	4	80%
B1 068 (text 3)	10	0	10	100%	11	6	5	45%	3	1	2	67%
B1 068 (text 4)	5	0	5	100%	13	1	12	92%	3	0	3	100%
B1 069 (text 1)	7	2	5	71%	17	8	9	53%	8	7	1	13%
B1 069 (text 2)	2	1	1	50%	9	8	1	11%	4	3	1	25%
B1 070 (text 1)	12	0	12	100%	17	11	6	35%	2	2	0	0%
B1 070 (text 2)	4	1	3	75%	15	9	6	40%	3	3	0	0%
B1 071 (text 1)	16	1	15	94%	14	7	7	50%	2	2	0	0%
B1 071 (text 2)	3	1	2	67%	18	11	7	39%	6	4	2	33%
B1 071 (text 3)	9	1	8	89%	8	5	3	38%	4	3	1	25%
B1 071 (text 4)	8	1	7	88%	19	11	8	42%	2	2	0	0%
B1 072 (text 3)	6	0	6	100%	12	8	4	33%	1	1	0	0%
Total	155	14	141	91%	307	155	152	50%	70	49	21	30%

The analysis of the second sample confirmed the findings of the first sample, which validates the thresholds selected in the preliminary study (Table 3).

Table 3.
Accuracy rates of the three thresholds according to both samples

	First Sample		Second Sample
	R > 0	R = 0	
Red	87%	88%	91%
Orange	38% - 58%	46%	49.5%
Yellow	23% - 39%		30%

Discussion

According to Chacon-Beltran (2017), bigram errors are approximately 46% of the written errors made by A2-B1 CEFR level L2 learners. Based on the hypothesis that bigrams which are infrequently used by educated native speakers of Spanish are more likely to contain errors than those which are commonly used by them, a bigram filter was built to help L2 learners of Spanish to correct their errors. Using real data, different thresholds were settled to create an effective tool in detecting such errors.

Thresholds of the bigram filter of the PGC

The preliminary study analysed the bigrams of 21 compositions of L2 learners of Spanish. The results of this analysis suggest that there is more than 87% probability that a bigram contains a grammatical error if: a) it does not occur in the corpus of reference (Fab = 0) and the probability that these two words occur together at random considering the frequency of each word separately is 4 or above (Pab ≥ 4), or b) they occur in the corpus and the Ratio is below 0.1. Meanwhile, there is a 58% probability that a bigram contains an error if they occur in the corpus of reference fewer than 75 times and their Ratio lie between 0.1 and 0.6. These results corroborate those from Lawley (2015), proving that the thresholds that he established in his grammar checker specifically designed to correct errors made by L2 learners of English are also valid in Spanish. However, the study suggests that, if the number of occurrences is fewer than 75 and the ratio is higher than 0.5, approximately a third of the bigrams might contain an error (23%-39%). This result differs from Lawley (2015), who found that just 20% of the bigrams with a Ratio of 20% were errors. This corroborates that there is a need to adapt learning tools to the target language since results from English and Spanish seem to differ.

Testing the thresholds of the bigram filter of the PGC

The results of this study clearly demonstrated that bigrams could be used as an error detection method in future pedagogic spell-checkers for L2 of Spanish in much the same way that they are already used in grammar checkers for L2 of English. The thresholds

established that resulted in highlighting combinations of words which are especially likely to contain errors were proven to be valid. All the bigrams errors were highlighted. The first threshold was able to detect 141 out of 328 incorrect bigrams (43%) with an accuracy of 90%. The other two thresholds identified all the errors at the risk of perpetrating false positives: 50% and 70% percent respectively. Acknowledging that identifying errors is key to helping L2 learners correct their writing (Lee, 1997), the three thresholds were kept and the probability of finding an error was shown, depending on the highlighting colour.

Whilst bigram analyses represent a powerful tool, it does not spot every single error written in a composition such as errors committed in longer strings (e.g. **in a few occasions*). A simple solution, as Harvey-Scholes (2018) suggests would be to divide the 4-gram **in a few occasions* in a unigram (*in*) and a trigram (*a few occasions*) and use the same method considering *a few occasions* as a hole (unigram). However, he pointed out that detecting these errors would "require a far larger corpus" (p. 147), since the number of occurrences of longer strings is lower. On the contrary, a broader corpus of reference will present an obvious intrinsic problem. A broader corpus will produce more infrequent words and bigrams as well as ortho-typographic errors, resulting into a spell checker and the bigram filter with an increased number of false positives. Additional research needs to be conducted in order to determine the ideal size that offers a higher predictive power whilst overlooking only a small number of errors. One suggested solution is to integrate two corpora of reference, a spellcheck and bigram filter and a larger one to detect longer word-strings into one.

In dealing with the errors that cannot be identified as n-grams such as tense errors (e.g. **every summer we work at home* when the intended sentence is *every summer we worked at home*) or false friends (**I was sleeping in my dormitory* when *I was sleeping in my bedroom* would be the appropriate sentence given the context), it may require a different approach. For example, both Harvey-Scholes (2018) and Chacón-Beltrán (2017) suggested that simply highlighting words that are commonly associated with errors such as verbs or false friends might be useful since L2 learners' attention are drawn on them, which Hernández García (2017) proved as an effective strategy in self-correction. However, which words and expressions must be highlighted or require a more thorough examination, since an excessive use of this strategy might be overwhelming and demotivating. Thus, further research is still needed to investigate and broaden the coverage of grammar-checking software by analysing real data, within the framework of the *error analysis* (Odlin, 1989), to better adapt the grammar checker to the L2 learner's interlanguage.

Conclusion and implications

Due to the limited time available for teachers to correct and provide feedback on their L2 learners' written compositions, self-correction is usually encouraged. However, students face difficulties detecting their own errors (Lee, 1997). Text-editing software equipped with grammar and spell checkers might help L2 learners in doing this task. Despite the limitations of spell and grammar check packages (Blazquez & Fan, 2019; Heift & Rimrott, 2008), spell checkers and bigram filters can detect up to 79% of the written errors (Harvey-Scholes, 2018). This widespread presence of errors in unigrams and bigrams has

great implications in the development of grammar check packages and L2 teaching alike since focusing on their treatment L2 learners can reduce their written errors dramatically. Because of that, the UNED decided to develop a grammar checker based on a bigram filter to help L2 learners of Spanish correcting their errors. This study has proven bigram filters, along with the thresholds that it used to identify errors, as an effective tool to detect L2 Spanish learners' grammatical errors. The development of an effective pedagogic spell checker, which can identify written errors and provide helpful feedback, will have positive implications for teaching schedules as well as learner's development. A student can correct many of his/her errors, and in the process also provide learning opportunities, even before the written output is submitted (Blazquez & Woore, in press). Meanwhile, teachers will significantly save time that can potentially be allocated to other, more important aspects of L2 writing acquisition where spell checkers proved to be less helpful such as content, argument, organizational structure, or register.

References

Blazquez, M. & Fan, C. (2019). The efficacy of spell check packages specifically designed for second language learners of Spanish. *Pertanika Journal of Social Science and Humanities – JSSH*, *27*(2), 847-863. Retrieved from: http://www.pertanika.upm.edu.my/Pertanika%20PAPERS/JSSH%20Vol.%2027%20(2)%20Jun.%202019/07%20JSSH-3837-2018.pdf.

Blazquez, M & Woore, R (in press). The importance of immediate feedback: does a pedagogical spell checker improve L2 Spanish learners' spelling accuracy? *Language Learning & Technology*. Accepted for publication.

Briscoe, T., Medlock, B., & Andersen, O. (2010). Automated assessment of ESOL free text examinations. *University of Cambridge Computer Laboratory Technical Report 790.* Cambridge, UK: University of Cambridge Computer Laboratory. Retrieved from http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-790.pdf

Brown, H.D. (1994). *Principles of Language Learning and Teaching.* Englewood Cliffs, NJ*:* Prentice Hall Regents.

Chacon-Beltran (2017). Free-form writing: computerized feedback for self-correction. *ELT journal, 71*(2), 141-149. Retrieved from https://doi.org/10.1093/elt/ccw064

Chen, H. J. H. (2009). Evaluating two web-based grammar checkers: Microsoft ESL Assistant and NTNU Statistical Grammar Checker. *Computational Linguistics and Chinese Language Processing, 14*(2), 161-180. Retrieved from http://www.aclweb.org/anthology/O/O09/O09-4002.pdf

Cestero Mancera, A. M., & Penedés Martínez, I. (2009). *Corpus de textos escritos para el análisis de errores de aprendices de E/ELE, CORANE.* Universidad de Alcalá de Henares. Alcalá de Henares.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge, UK: Cambridge University Press.

Dziemianko, A. (2015). Colours in online dictionaries: A case of functional labels. *International Journal of Lexicography*, *28*(1), 27-61. Retrieved from https://doi.org/10.1093/ijl/ecu028

Graham, S., & Santangelo, T. (2014). Does spelling instruction make students better spellers, readers, and writers? A meta-analytic review. *Reading and*

Writing, *27*(9), 1703-1743.Retreived from https://doi.org/10.1007/s11145-014-9517-0

Graham, S., Harris, K.R., & Hebert, M. (2011). *Informing writing: The benefits of formative assessment.* Alliance for Excellence in Education. Washington, D.C. (Commissioned by the Carnegie Corp. Of New York). Retrieved from https://www.carnegie.org/media/filer_public/37/b8/37b87202-7138-4ff9-90c0-cd6c6f2335bf/ccny_report_2011_informing.pdf

Harvey-Scholes, C. (2018). Computer-assisted detection of 90% of EFL student errors. *Computer Assisted Language Learning*, *31*(1-2), 144-156. Retrieved from https://doi.org/10.1080/09588221.2017.1392322

Heffernan, N., & Otoshi, J. (2015). Comparing the pedagogical benefits of both criterion and teacher feedback on Japanese EFL students' writing. *JALT CALL Journal*, *11*(1), 63-76. Retrieved from https://files.eric.ed.gov/fulltext/EJ1107988.pdf

Heift, T., & Rimrott, A. (2008). Learner responses to corrective feedback for spelling errors in CALL. *System*, *36*(2), 196-213. Retrieved from https://doi.org/10.1016/j.system.2007.09.007

Hernández García, F. (2017). La detección y corrección de errores en la deixis temporal del verbo en redacciones escritas en inglés como lengua extranjera. *Estudios de lingüística inglesa aplicada, 17*, 183-207. Retrieved from https://doi.org/10.12795/elia.2017.17.08

Kang, SoJin & Lee, Jin-Hwa (2013). Effects of color and serial position of words on L2 vocabulary recall and retention. *Language Research, 49*(2), 147-167. Retrieved from http://s-space.snu.ac.kr.

Kiuhara, S., Graham, S., & Hawken, L. (2009). Teaching writing to high school students: A national survey. *Journal of Educational Psychology*, *101*, 136-160. Retrieved from https://doi.org/10.1037/a0013097

Lawley, J. (2015). New software to help EFL students self-correct their writing. *Language Learning & Technology*, *19*(1), 23-33. Retrieved from https://www.learntechlib.org/p/159024/

Lawley, J. (2016). Spelling: computerised feedback for self-correction. *Computer Assisted Language Learning*, *29*(5), 868-880. Retrieved from https://doi.org/10.1080/09588221.2015.1069746

Lázaro Ibarrola, A. (2013). Reformulation and self-correction: insights into corrections strategies for EFL writing in a school context. *Vigo International Journal of Applied Linguistics, 10*, 29-49. Retrieved from http://vialjournal.webs.uvigo.es/pdf/Vial-2013-Article2.pdf

Lee, I. (1997). ESL learners' performance in error correction in writing: Some implications for teaching. *System*, *25*(4), 465-477. Retrieved from https://doi.org/10.1016/S0346-251X(97)00045-6

Mitton, R. (2010). Fifty years of spellchecking. *Writing Systems Research*, *2*(1), 1-7. Retrieved from https://doi.org/10.1093/wsr/wsq004

Mitton, R., & Okada, T. (2007). The adaptation of an English spell-checker for Japanese writers. *CALICO Journal*, *16* (4), 584-595. Retrieved from http://eprints.bbk.ac.uk/id/eprint/592

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning*. Cambridge University Press.

Quinn, C. (2014). Training L2 writers to reference corpora as a self-correction tool. *ELT Journal 69*(21), 65-77. Retrieved from https://doi.org/10.1093/elt/ccu062

San Mateo, A. (2016). Un corpus de bigramas utilizado como corrector ortográfico y gramatical destinado a hablantes nativos de español. *Revista Signos*, *49* (90), 94-118. Retrieved from http://doi.org/10.4067/S0718-09342016000100005

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*, 129-158. Retrieved from https://doi.org/10.1093/applin/11.2.129

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

Thurmair, G. (1990, August). Parsing for grammar and style checking. In *Proceedings of the 13th conference on Computational Linguistics-Volume 2* (pp. 365-370). Association for Computational Linguistics. Retrieved from https://dl.acm.org/citation.cfm?id=998002

Universidad Nacional de Educación a Distancia. (2017). *Grammar checker*. Retrieved from http://www.euned.es/correctme/cmenglish/identification.php?msg=Usuario+no+identificado,+introduzca+sus+credenciales.

Van Beuningen, C.G., De Jong, N. & Kuiken, F. (2012). The effect of direct and indirect corrective feedback on L2 learners' written accuracy. *ITL International Journal of Applied Linguistics 156*, 279-296. Retrieved from https://doi.org/10.2143/ITL156.0.2034439

Vickers, C. H. & Ene, E. (2006). Grammatical accuracy and learner autonomy in advanced writing. *ELT Journal, 60*(2), 109-16. Retrieved from https://doi.org/10.1093/elt/cci097

Wilcox, K. C., Yagelski, R., & Yu, F. (2014). The nature of error in adolescent student writing. *Reading and Writing*, *27*(6), 1073-1094. Retrieved from https://doi.org/10.1007/s11145-013-9492-x

Wu, Y. L. (2014). The Impact of Technology on Language Learning. In Park J., Pan Y., Kim CS., Yang Y. (eds) *Future Information Technology. Lecture Notes in Electrical Engineering*, vol 309 (pp. 727-731). Springer, Berlin, Heidelberg. Retrieved from https://doi.org/10.1007/978-3-642-55038-6_112