"Categorizing spelling errors to assess L2 writing"

Bestgen, Yves ; Granger, Sylviane

Abstract
Based on a corpus of 223 argumentative essays written by English as a foreign language learners, this study shows that spelling errors, whether detected manually or automatically, are a reliable predictor of the quality of L2 texts and that reliability is further improved by sub-categorising errors. However, the benefit derived from sub-categorisation is much lower in the case of errors automatically detected by means of the Microsoft Word 2007 spell checker, a situation which results from Word's limited success in detecting and correcting some specific categories of L2 learner errors.

Document type : Article de périodique (Journal article)

Référence bibliographique


DOI : 10.1504/IJCEELL.2011.040201
Categorizing spelling errors to assess L2 writing
Yves Bestgen & Sylviane Granger


Abstract

Based on a corpus of 223 argumentative essays written by English as a foreign language learners, this study shows that spelling errors, whether detected manually or automatically, are a reliable predictor of the quality of L2 texts and that reliability is further improved by subcategorizing errors. However the benefit derived from subcategorization is much lower in the case of errors automatically detected by means of the Microsoft Word 2007 spell checker, a situation which results from Word’s limited success in detecting and correcting some specific categories of L2 learner errors.

Keywords
English as a foreign language, learner corpus, automatic scoring, automatic error detection, misspellings, International Corpus of Learner English, Common European Framework of Reference for Languages, spell checker, spelling error subcategorization, L1 transfer, letter doubling, word segmentation errors

1. Introduction

Research into spelling correction has long focused on the needs of native speakers. As a result, “spelling correction tools designed for language learners are rare” (Leacock et al., 2010, p.79) and foreign language learners (henceforth L2 learners) have no other choice but to use existing spell checkers which have been trained on native (L1) data and are therefore much less successful in dealing with learner errors. Rimrott and Heift (2005) evaluated the Microsoft Word 2003 spelling corrector on learner German texts and found that only 52% of the errors were detected, a substantially lower figure than that for texts written by German native speakers. This study and a few other recent ones (Al-Jarf, 2009a; Botley and Dillah, 2007; Hovermale, 2008; Hovermale and Martin, 2008; Mitton and Okada, 2007; Okada, 2004; Rimrott and Heift, 2008) bear witness to a surge of interest in L2 spelling errors. The main objective of these studies is to assess the reliability of spell checkers for L2 users and suggest avenues for improvement (Heift and Rimrott, 2008). The research focuses mainly on the specificity of L2 vs. L1 errors and the impact of the foreign language learner’s L1 on the types of errors produced. Results highlight the following characteristics of L2 misspellings: (1) there is a wide diversity of error types; (2) there is a larger number of errors than in L1 users; (3) a large proportion of errors result from a lack of knowledge of the target language.

In the field of automatic text assessment, the presence of a large number of spelling errors is often viewed as a factor that hinders effective assessment. This is typically the case for systems that score open-ended questions by comparing students’ answers to some gold standard (Landauer, Laham, and Foltz, 2003; Pérez, Alfonseca and Rodriguez, 2004). However, errors – whether orthographic or grammatical – have also been proved to make a positive contribution to the grading of texts (Chodorow and Burstein, 2004; Lonsdale and Strong-Krause, 2003), especially within the framework of L2 assessment. Judging from the literature, this research trend has been relatively marginal and results are mixed. Some automatic scoring systems, primarily aimed at native users, take spelling errors into account
and most of the systems provide feedback on these errors (see reviews in Chung and O’Neil, 1997; Dikli, 2006; Warschauer and Ware, 2006). However, this information does not play a key role in text assessment. In addition, spelling errors are treated as one undifferentiated category rather than being broken down into subtypes (Burstein, Chodorow, and Leacock, 2004). The impact of spelling errors is also usually neglected in recent L2 research. Many studies rely on lexical variation to predict text quality (Bestgen, Lories, and Thewissen, 2010; Crossley et al., 2008; Yu, 2010) and disregard the possible influence of misspellings, which Granger and Wynne (1999) have shown to be very strong.

This study has two main objectives: first, to assess whether it is possible to predict learners’ proficiency automatically on the basis of spelling errors and second, to examine the possibility that subcategorizing these errors might improve the prediction of both manual and automatic detection. The data used are authentic texts produced by L2 learners from three different mother tongue backgrounds. The analysis, which involves a comparison between a golden standard manual (i.e. a set of manually detected errors) and automatic flags by Microsoft Word 2007, provides a valuable platform from which to assess Word’s efficiency in detecting and correcting errors in authentic L2 texts. This information could be used to improve not only automatic spell checkers, but also free-text assessment tools by allowing them to remove spelling errors more efficiently before the assessment step is undertaken.

The article is structured as follows. In section 2, we describe the corpus data, the rating procedure and the methodology used to annotate the spelling errors both manually and automatically. Section 3 investigates the efficiency of manually tagged errors in predicting the quality of an essay while section 4 assesses the success rate of MS Word 2007 in terms of error detection and correction as well as prediction of essay quality. In section 5 we tackle the possible impact of L1 differences on automatic error detection and prediction efficiency. The last section discusses some limitations of the study and suggests avenues for future research.

2. Data and methodology

2.1. Learner corpus data

Our study is based on data extracted from a large computer learner corpus, i.e. an electronic collection of texts written by foreign language learners. The corpus, the International Corpus of Learner English (ICLE), is a 3.7 million word corpus of essays written by intermediate to advanced learners of English as a foreign language from 16 mother tongue backgrounds (Granger et al, 2009). Two hundred and twenty three argumentative essays were extracted from three ICLE subcorpora, i.e. 74 essays were taken from the French (FR) component, 71 from the German (GE) component and 78 from the Spanish (SP) component of the learner corpus. These texts vary between 500 and 900 words in length. The detailed breakdown of the learner corpus sample used is presented in Table 1.

<table>
<thead>
<tr>
<th>L1 background</th>
<th>Number of learner essays</th>
<th>Total tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>74</td>
<td>50,195</td>
</tr>
<tr>
<td>GE</td>
<td>71</td>
<td>49,856</td>
</tr>
<tr>
<td>SP</td>
<td>78</td>
<td>51,397</td>
</tr>
<tr>
<td>Total</td>
<td>223</td>
<td>151,448</td>
</tr>
</tbody>
</table>

2.2. Rating procedure
Although the essays in the ICLE can be broadly described as intermediate to advanced, this still leaves room for considerable differences in proficiency between the texts. With a view to assigning a more precise proficiency rating to each text, the 223 essays were assessed by two professional raters, who assigned a score based on the descriptors of the Common European Framework of Reference for Languages (CEF) (Council of Europe, 2001) (for a detailed description of the rating procedure, see Bestgen, Lories, and Thewissen, 2010).

The CEF includes six proficiency levels which can be broken down into three groups of two, viz. A1 and A2 (=basic users; elementary proficiency learners), B1 and B2 (=independent users; intermediate proficiency learners), C1 and C2 (=proficient users; advanced proficiency learners). Raters were further able to use + or – signs to further specify quality within each proficiency level (e.g. B2- or C1+). We started the rating procedure at level B1 because the CEF specifies that learners need a B1 proficiency level to attempt essay writing. In order to attribute one final CEF score to each text, we computed the mean of the holistic scores given by the raters (each holistic score had been given a numerical value, as shown in Table 2).

<table>
<thead>
<tr>
<th>Holistic CEF score</th>
<th>B1-</th>
<th>B1</th>
<th>B1+</th>
<th>B2-</th>
<th>B2</th>
<th>B2+</th>
<th>C1-</th>
<th>C1</th>
<th>C1+</th>
<th>C2-</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerical value</td>
<td>0.67</td>
<td>1</td>
<td>1.33</td>
<td>1.67</td>
<td>2</td>
<td>2.33</td>
<td>2.67</td>
<td>3</td>
<td>3.33</td>
<td>3.67</td>
<td>4</td>
</tr>
</tbody>
</table>

The mean CEF score is 2.35 with a standard deviation of 0.99. However, as appears from Table 3, there are marked differences between the three subcorpora. A one-way between-groups analysis of variance (ANOVA) indicated that there are highly significant differences in mean CEF scores between the groups (F(2, 220) = 96.14 ; p < 0.0001). The Student-Newman-Keuls post-hoc test revealed that all three group means actually differ significantly from each other. As can be noted, however, the difference between the SP group and the FR/GE groups is much more marked than that between the FR and GE groups themselves.

<table>
<thead>
<tr>
<th>L1 Group</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP</td>
<td>78</td>
<td>1.44</td>
<td>0.64</td>
</tr>
<tr>
<td>FR</td>
<td>74</td>
<td>2.64</td>
<td>0.67</td>
</tr>
<tr>
<td>GE</td>
<td>71</td>
<td>3.03</td>
<td>0.87</td>
</tr>
</tbody>
</table>

2.3. Error detection and correction

Manual detection and correction of the errors was carried out by a native speaker of English with considerable expertise in English language and linguistics. The same analyst subsequently tagged each error in accordance with the Louvain error tagging system (Dagneaux et al. 1998)¹ and inserted the tags into the text files with the help of the Université catholique de Louvain Error Editor (UCLEE), a menu-driven editor by means of which error tags and corrections are inserted into the text files with the appropriate markup. The Louvain

---

¹ The error tagging procedure was carried out within the framework of a PhD that is under completion at the Centre for English Corpus Linguistics (Thewissen, J. Accuracy across L1 backgrounds and proficiency levels: Insights from an error-tagged EFL learner corpus. Université catholique de Louvain: Centre for English Corpus Linguistics).
error tagging system includes eight major error domains (Formal, Grammatical, Lexical, Lexico-grammatical, Punctuation, Word redundant/missing/order, Style and Infelicity), which are further broken down into 56 error categories. For the present study, only the errors pertaining to the form of words have been selected. The majority of these errors are tagged F and further broken down into pure spelling errors (FS) and morphological errors (FM). While the majority of FS errors result in non-existing word forms, the category also includes words with similar forms which are easily confused, i.e. pairs such as their-there or to-too. We also included erroneous forms such as the plural forms of uncountable nouns (advices) or adjectives (responsibles) which are categorized as grammatical or lexico-grammatical errors according to the Louvain error tagging system. Examples 1 to 4 illustrate the different categories with the appropriate markup, viz. the error tag written between brackets in front of the erroneous word and the correction presented between dollar signs after the error.

(1) The fast spread of television can transform it into a double-edged (FS) weapon.
(2) Today (FS) its $it's$ the Germans who have turned to killing foreigners living in Germany.
(3) There are just as many weak, emotional, (FM) impractical $impractical$ men as there are strong and practical women.
(4) The experiments of breeding hundreds of (XNUC) cattles $cattle$ out of one cell have raised a storm of protest.

Alongside the manual error detection process, a process of automatic error detection was carried out on the texts, using Microsoft Word 2007 with only the spell checker turned on. We chose to ignore words in upper case and words with numbers to avoid Word flagging abbreviations and words like B170 or 54m2. The spellchecking was carried out twice: first with the UK English dictionary and then with the US English dictionary. Only the words that were flagged by the two analyses were considered as errors for Word. The first correction suggested by Word was consistently accepted. When no suggestion was made, a # was inserted instead.

2.4. Spelling error subcategorization

As stated above, one of the objectives of this study was to investigate the potential of breaking the errors down into subcategories on the basis of the differences between the forms produced by the learners and the correct forms proposed by the analyst or the spell checker. The variables underlying the categorization are: the element that carries the error (letter, word boundary, apostrophe) and the error type (single letter addition, omission, substitution, or transposition and multiple letter errors). Capitalization errors have been disregarded. Table 4 lists the 9 categories together with their codes and illustrates them with examples extracted from the corpus.

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>0X</td>
<td>Omission of a letter</td>
<td>completely completely concious conscious</td>
</tr>
<tr>
<td></td>
<td></td>
<td>distinct distinct eople people</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mechanisms mechanisms throughout throughout</td>
</tr>
<tr>
<td>X0</td>
<td>Addition of a letter</td>
<td>develope develop</td>
</tr>
<tr>
<td>Error Type</td>
<td>Description</td>
<td>Examples</td>
</tr>
<tr>
<td>-------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Doub12</td>
<td>Single letter instead of double letter</td>
<td>especially especially robbed robbed addicts addicts carrots carrots occurred occurred occurring occurring</td>
</tr>
<tr>
<td>Doub21</td>
<td>Double letter instead of single letter</td>
<td>apartments apartments allmighty allmighty detailed detailed losing losing professors' professors'</td>
</tr>
<tr>
<td>XY</td>
<td>Substitution of one letter</td>
<td>lifes lives dependend dependent consequently consequently comfortable comfortable engeneering engineering uncredible incredible</td>
</tr>
<tr>
<td>Swap</td>
<td>Interchange of two adjacent letters</td>
<td>conceived conceived birht birth life life peoples peoples entreprises enterprises</td>
</tr>
<tr>
<td>Apost</td>
<td>Error involving an apostrophe</td>
<td>Its it's womans woman's children's children's</td>
</tr>
<tr>
<td>SplitW</td>
<td>Erroneous splitting or joining of words (word segmentation error)</td>
<td>business_man businessman every_one everyone free-time free time everyday every_day airpollution air_pollution eventhough even_though</td>
</tr>
<tr>
<td>Many</td>
<td>Two or more errors of the same type or of different types</td>
<td>unbalance imbalance politic political payed paid weter whether dustbinman dustman theirselves theirselves begginning beginning configurating configuring divorcion divorce hitted hit</td>
</tr>
</tbody>
</table>

Note: The underscore represents a space character.

Besides traditional categories like omission, addition, substitution, swap and many (Pollock and Zamora, 1984), our categorization also contains categories that are less often used in the literature such as the erroneous doubling (or lack of doubling) of letters or erroneous splitting.

2 Note that Doub12 is a special case of 0X and Doub21 of X0. As shown below, the letters involved in both categories are predominantly consonants.
(or joining) of words. These categories have been added because they seemed to be particularly recurrent features of learner spelling (cf. Al-Jarf, 2009b; Bebout, 1985; Cook, 1997).

2.5. Statistical technique

The first question we were addressing in this study was the extent to which the total number of spelling errors and the different categories of misspellings can be used to predict the CEF scores attributed to each learner text. As is customary in the literature on automated writing evaluation (see for review, Warschauer and Ware, 2006; Yang et al. 2002), we used the statistical technique of linear regression to assess the predictive power of misspellings. In these analyses, the variable to be predicted is the CEF score of a text and the predictors are the relative error frequencies in that same text. We used relative frequencies to take into account the variable length of the texts. The frequencies in each different category were divided by the total number of words in the text and multiplied by 1000. For the prediction based on categories of misspellings, the best predictors were selected by means of the Stepwise procedure (with a significance level to enter and to stay of 0.05). The quality of the prediction was measured by the $R^2$ coefficient which corresponds to the part of the variance that can be explained by the predictor(s). Its value ranges from 0 to 1, where 1 represents perfect prediction. As the present study compares models based on different numbers of predictors, it was preferable to use the Adjusted $R^2$ which takes this factor into account.

Admittedly, spelling errors are but one variable among many that can be used to predict the quality of a text and considered in isolation, their predictive power is necessarily limited. However, our analyses provide an initial indication of the role played by spelling errors in general and make it possible to assess whether subcategorization improves the prediction. The comparison between manual and automatic error detection and correction also gives us an idea of the challenges faced in attempting to fully automate the process of text evaluation.

3. CEF score prediction: manual approach

3.1. Total number of errors

A total of 1614 errors were initially identified manually (10.7 errors per 1000 words). The analysis of the distribution of the number of errors per text showed that one of the Spanish L2 texts contained 65 spelling errors, meaning that this text alone contained 4% of the total number of errors in the whole corpus. As the error frequency was markedly higher in this text than in all the other error-dense texts (which contained in the region of 30 errors each), the decision was made to discard this ‘rogue’ text from subsequent analyses, which reduced the number of errors to 1549 for 222 texts.

Results of the regression analysis demonstrate that the relative frequency of spelling errors in a text can significantly predict ($p<0.0001$) the CEF score of that text with an Adjusted $R^2$ of 0.34. According to Cohen’s (1988) classical guidelines for effect size, this value can be considered as large.

3.2. Error categories
The second stage in the analysis involved comparing the results obtained for the total number of spelling errors with those obtained when subcategorization is introduced. The frequency of the different error categories in the learner corpus is shown in Figure 1 (N= 1549).

![Figure 1: Frequency of spelling error categories](image)

The figure shows that Many represents more than 23% of the errors and SplitW more than 18%. These two types of errors tend to be neglected by classic automatic detection systems which mainly focus on single letter errors which are predominant in L1 writing (Pollock and Zamora, 1983). The frequency of multi-letter errors has also been underlined by Rimrott and Heift (2008) for learners of German. Figure 1 also brings out a sizeable proportion of doubling errors (Doub12 and Doub21). Although this is a classic spelling difficulty in English which affects both native and L2 writers (cf. Cook, 1997), a recent study based on the LOCNESS corpus, a corpus of essays written by American undergraduates (cf. Granger and Wynne, 1999), gives frequencies below 0.1 per 1000 words for Doub12 and Doub21 errors, while the frequencies in our L2 corpus are above 0.7 and 0.5 for Doub12 and Doub21 respectively.

Multiple regression analysis based on the different categories of errors yields a better prediction than that achieved for the total number of undifferentiated errors. The Adjusted $R^2$ based on the 6 predictors selected by the Stepwise procedure was equal to 0.43, compared to 0.34 for the previous analysis. The 6 selected predictors are given in Table 5. They are ordered according to their importance in predicting the CEF score, the importance being measured by the Squared Semi-Partial Correlation between the predictor and the CEF score. It corresponds to the decrement in $R^2$ that would result from the elimination of this predictor from the model (Howell, 2007, p.526).

In order to facilitate comparison between the different analyses, we present in Table 5 an overview of all the results. The discussion of some of these results is deferred to the next section, however.
Table 5: CEF score prediction: manual vs. automatic detection

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of errors</td>
<td>0.34</td>
<td></td>
<td>0.27</td>
<td>0.32</td>
<td></td>
<td>0.070</td>
</tr>
<tr>
<td>Error categories</td>
<td>0.43</td>
<td>X0</td>
<td>0.054</td>
<td>0.32</td>
<td>Doub12</td>
<td>0.070</td>
</tr>
<tr>
<td></td>
<td></td>
<td>XY</td>
<td>0.036</td>
<td></td>
<td>0X</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Doub12</td>
<td>0.036</td>
<td></td>
<td>X0</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Many</td>
<td>0.030</td>
<td></td>
<td>XY</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0X</td>
<td>0.014</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SplitW</td>
<td>0.014</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: SSPC stands for Squared Semi-Partial Correlation. Where there is only one predictor, SSPC is equal to the model $R^2$.

4. CEF score prediction: automatic approach

For the automatic detection stage of our analysis we used Word 2007, a commercial, widely-distributed software tool frequently used by L2 writers to correct their spelling. Investigating Word’s success at predicting CEF scores also gave us the opportunity to assess its accuracy in L2 error detection and correction. Rimrott and Heif (2008), the most in-depth study that has been carried out of a very limited number of studies of L2 spelling errors, indicates that Word 2003 is not highly successful. When interpreting their results, it is important to bear in mind that their study only assesses whether Word detects errors that are present in the texts. It does not investigate whether Word flags correct words as errors. In a study like ours which aims to predict text quality, incorrect flags need to be included.

4.1. Total number of errors

4.1.1. Precision and recall rates of Word 2007

Word’s spell checker flagged 1513 words in the corpus, which seems at first sight to be close to the number of manually detected errors (1549). However, subsequent analysis revealed that these two figures did not cover the same sets of errors. Of the 1513 words flagged by Word, 267 were overflags (OFs), i.e. words that were not detected as errors in the manual analysis. Only 1246 words could therefore be considered as correct flags (CFs). In addition, Word missed 303 manually detected spelling errors, henceforth referred to as underflags (UFs).

To assess Word’s success at error detection we used the measures of precision and recall recommended by Leacock et al (2010, p.38) to evaluate error detection systems in the language learner's context. Salton (1989, p.248) defines these measures as follows: “Two main parameters of retrieval effectiveness have been used over the years, defined as the proportion of relevant materials retrieved, or recall (R), and the proportion of retrieved materials that are relevant, or precision (P)”. Applied to our study, this produces the following results:

\[
\text{Recall rate} = \frac{\text{Number of relevant spelling errors flagged by Word07}}{\text{Total number of manually detected spelling errors}} = \frac{1246}{1549} = 80.43\%
\]

\[
\text{Precision rate} = \frac{\text{Number of relevant spelling error flagged by Word07}}{\text{Total number of automatic spelling error flagged by Word07}} = \frac{1246}{1513} = 82.35\%
\]
An analysis of the OFs shows that the vast majority (84%) is made up of foreign words (liégeois, Grüne), abbreviations in lower case or including lower case (a.s.o., CFC’s), proper names (Garbo, Testarossa) or stylistic effects (GE46 : when she suddenly lisped, I'm terribly thorry, but the liht ith full now. Can you try it again nekht themether?). While Word can hardly be blamed for flagging these words, the situation is different with the remaining 16%, which are also more serious in an L2 perspective as they may mislead L2 learners. Here we find clear detection errors such as The parents too have a role to play in the education of their children -> The parents to have... and English words that are absent from Word’s dictionary (categorial, hoovering). While the second category of OFs is particularly confusing for L2 learners, both categories are just as problematic when it comes to predicting the quality of a text.

4.1.2. Prediction of CEF scores

As shown in Table 5, the relative frequency of errors detected by Word in an L2 text makes it possible to predict the CEF score significantly with an Adjusted R² of 0.27. This value is significantly lower than that obtained on the basis of manual detection, which stands at 0.34 (Z = -2.20; p<0.05; comparison of correlated correlation coefficients: Meng, Rosenthal, and Rubin, 1992).

4.2. Error categories

4.2.1. Efficiency in error detection: breakdown per category

When Word’s efficiency in error detection is assessed in terms of error categories, it soon appears that the precision and recall rates given in the preceding section conceal marked discrepancies between categories. For example, of the 282 errors which fall into the category of SplitW according to manual annotation, only 88 (31%) are flagged as an error by Word. All the other instances are underflags. SplitW errors actually account for 64% of the spell checker’s 303 UFIs. Particularly striking is the fact that Word accepts as correct 108 of the 109 hyphenated words which should have been written unhyphenated or solid (e.g., honey-moon -> honeymoon). It also misses 80 of the 101 unhyphenated words which should either have been written hyphenated or solid (e.g., sheep_dog -> sheepdog). However, it proves more effective at detecting solid words which should have been written hyphenated or unhyphenated as two words, as only 6 of the 72 instances are missing (e.g., good-hearted -> good-hearted, everyday -> every_day). Another noteworthy category is the Many category. Only 60% of the words flagged by Word were annotated manually as errors. The other 40% are overflags (cf. section 4.1.1).

4.2.2. Error correction success rate: breakdown per category

So far we have only investigated Word’s success at detecting spelling errors. Another particularly important issue for L2 writers is whether it is capable of suggesting the correctly spelt word as first correction. In this section, we only take into account the cases where both Word and manual annotation have detected an error and provided a correction. Our aim is to determine whether automatic and manual analyses suggest the same word or not. If Word presents the same word as in the manual analysis as first suggestion, the occurrence is counted as correct flag (CF). The results are presented in Table 6. On the left-hand side of the table, the percentage of correct flags is calculated according to the error categories based on manual detection and correction, while on the right-hand side, the percentage is calculated on the
basis of the error categories based on Word 2007. The difference between the two sides of the table is highlighted clearly by the Swap category. When there is a Swap error according to manual analysis, Word provides the right correction in 97% of cases, but when Word says it is a Swap error, the percentage of correct flags is much lower (72%). This difference probably reflects a tendency for Word’s algorithms to prioritize Swap errors over other error types. This prioritization might well turn out to be more appropriate for native texts than L2 texts.

<table>
<thead>
<tr>
<th>Category</th>
<th>Total</th>
<th>CF (%)</th>
<th>Total</th>
<th>CF (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0X</td>
<td>187</td>
<td>86.10</td>
<td>215</td>
<td>77.67</td>
</tr>
<tr>
<td>X0</td>
<td>140</td>
<td>80.71</td>
<td>152</td>
<td>74.34</td>
</tr>
<tr>
<td>Doub12</td>
<td>107</td>
<td>97.20</td>
<td>116</td>
<td>91.38</td>
</tr>
<tr>
<td>Doub21</td>
<td>76</td>
<td>98.68</td>
<td>79</td>
<td>97.47</td>
</tr>
<tr>
<td>XY</td>
<td>301</td>
<td>77.08</td>
<td>322</td>
<td>72.98</td>
</tr>
<tr>
<td>Swap</td>
<td>34</td>
<td>97.06</td>
<td>46</td>
<td>71.74</td>
</tr>
<tr>
<td>Apost</td>
<td>14</td>
<td>85.71</td>
<td>49</td>
<td>24.49</td>
</tr>
<tr>
<td>SplitW</td>
<td>88</td>
<td>88.64</td>
<td>102</td>
<td>76.47</td>
</tr>
<tr>
<td>Many</td>
<td>299</td>
<td>26.09</td>
<td>165</td>
<td>39.39</td>
</tr>
</tbody>
</table>

Note: CF = correct flag (Word 2007 provides the correct spelling as first suggestion)

The table shows marked differences between categories, some (e.g., Doub21) achieving very high CF rates while others (e.g., XY or X0) fare much less well and would merit closer investigation of the underlying algorithms. Two categories deserve particular attention in view of their very low CF rates. The most problematic category is Many, a result which confirms Rimrott and Heift’s (2008) study of L2 German errors. This result is hardly surprising as Word is likely to experience more difficulty in finding the correct word when the original word contains many errors. For example, for *sacrifying* Word uses Swap and suggests *scarifying* when in fact the target word was *sacrificing*, which belongs to the Many category (XY + 0X). Another problematic category is Apost (right-hand side of the table). Word makes more use of this category than it should, mainly because it frequently introduces an apostrophe when the incorrect word contains ‘s’ as its last letter. This is the case, for example, for the word *lifes*, which is corrected as *life’s* instead of *lives* (16 occurrences).

### 4.2.3. Prediction of CEF scores

The analysis revealed that multiple regression achieves a better prediction rate when it is based on the different error categories than when it is based on global scores. The Adjusted R² based on the 4 predictors selected by the Stepwise procedure is equal to 0.32 (vs. 0.27 for the undifferentiated error totals). However, the gain is much lower than that based on categories of manually detected errors (0.43). The selected predictors and their Squared Semi-Partial Correlations are presented in Table 5. It appears clearly from the table that Many and SplitW are not selected as predictors by Word. This was to be expected as these are the categories of errors that Word was found to be least good at detecting (SplitW, see section 4.2.1) and correcting (Many, see section 4.2.2).
5. The impact of learners’ L1

Up to now, this article has presented analyses of the learner data as a whole, with no distinctions made relating to the three different learner populations – FR, GE and SP – it represents. However, studies such as Bebout (1985), James et al. (1993), Al-Jaef (2009a) and Mitton and Okada (2007) conclude that the learners’ L1 affects the type of spelling errors they produce. It therefore seems likely that an automatic detection system which is better at detecting and correcting some error categories than others – and our study shows that this is the case for Word – runs the risk of favouring some learner populations and penalizing others and this risk exists whether the analysis is based on undifferentiated or subcategorized spelling errors. In this section we investigate the potential impact of the learners’ mother tongue by providing the L1 breakdown of the error categories and some examples of L1-related categories of errors.

5.1. Error frequency: L1 breakdown of manually detected errors

The potential impact of the learners’ L1 on the frequency of error types was assessed by means of ANOVAs, taking the relative frequency of each error category as dependent variable. In cases where the difference is significant, the means of the three L1 groups were compared by means of the Student-Newman-Keuls procedure.

<table>
<thead>
<tr>
<th>Category</th>
<th>FR Mean</th>
<th>FR SD</th>
<th>GE Mean</th>
<th>GE SD</th>
<th>SP Mean</th>
<th>SP SD</th>
<th>R²</th>
<th>p</th>
<th>Diff.Sign</th>
</tr>
</thead>
<tbody>
<tr>
<td>0X</td>
<td>0.6</td>
<td>1.1</td>
<td>0.9</td>
<td>1.6</td>
<td>2.5</td>
<td>3.1</td>
<td>.14</td>
<td>***</td>
<td>SP&gt;(FR=GE)</td>
</tr>
<tr>
<td>Doub12</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
<td>0.6</td>
<td>1.8</td>
<td>2.1</td>
<td>.25</td>
<td>***</td>
<td>SP&gt;(FR=GE)</td>
</tr>
<tr>
<td>X0</td>
<td>0.6</td>
<td>1.1</td>
<td>0.7</td>
<td>1.4</td>
<td>1.6</td>
<td>2.0</td>
<td>.09</td>
<td>***</td>
<td>SP&gt;(FR=GE)</td>
</tr>
<tr>
<td>Doub21</td>
<td>0.4</td>
<td>0.8</td>
<td>0.5</td>
<td>1.1</td>
<td>0.6</td>
<td>1.1</td>
<td>.01</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>XY</td>
<td>1.1</td>
<td>1.8</td>
<td>1.5</td>
<td>3.0</td>
<td>3.7</td>
<td>3.4</td>
<td>.14</td>
<td>***</td>
<td>SP&gt;(FR=GE)</td>
</tr>
<tr>
<td>Swap</td>
<td>0.2</td>
<td>0.6</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>1.0</td>
<td>.05</td>
<td>**</td>
<td>SP&gt;GE,FR=GE,FR=SP</td>
</tr>
<tr>
<td>Apost</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.6</td>
<td>0.0</td>
<td>0.2</td>
<td>.08</td>
<td>***</td>
<td>GE&gt;(FR=SP)</td>
</tr>
<tr>
<td>SplitW</td>
<td>1.3</td>
<td>1.9</td>
<td>3.0</td>
<td>3.0</td>
<td>1.4</td>
<td>1.7</td>
<td>.11</td>
<td>***</td>
<td>GE&gt;(FR=SP)</td>
</tr>
<tr>
<td>Many</td>
<td>1.8</td>
<td>2.1</td>
<td>1.4</td>
<td>1.9</td>
<td>3.9</td>
<td>3.0</td>
<td>.17</td>
<td>***</td>
<td>SP&gt;(FR=GE)</td>
</tr>
</tbody>
</table>

Note: SD = Standard Deviation; NS = Not Significant; ** = p<0.01; *** = p<0.0001. The Diff Sign column presents the statistically significant differences between the means according to the Student-Newman-Keuls procedure. For example, SP>(FR=GE) indicates that the mean for the FR group does not differ significantly from that of the GE group, while both of these means are significantly lower than the mean for SP.

As appears from Table 7, there is a statistically significant difference between the SP group and the other two groups for nearly all the error categories. The higher error rate in the SP texts ties in with the difference in the quality of the essays described above (section 2.2), which showed that the difference between the SP group and the FR/GE groups is much more marked than that between the FR and GE groups.

5.2. Illustrations of L1-related error categories
The SplitW category deserves special attention as it is the only one which is more prominent in the GE group than in the other two. The doubling categories are also worthy of closer investigation. They display opposite tendencies: Doub12 is significant and the effect size is large ($R^2=0.25$), while Doub21 is non-significant and the effect size is close to zero. A more fine-grained qualitative analysis will enable us to assess whether, as suggested in the literature, these differences originate in transfer from the learners’ L1.

5.2.1. Word segmentation

SplitW errors can be further subcategorized into three categories:
- unhyphenated words that should be hyphenated or solid (e.g. *business man* -> *businessman*)
- hyphenated words that should be unhyphenated or solid (e.g. *free-time* -> *free time*)
- solid words that should be hyphenated or unhyphenated (e.g. *airpollution* -> *air pollution*)

The L1 breakdown of the three categories is given in Table 8.

<table>
<thead>
<tr>
<th>Category</th>
<th>FR Mean</th>
<th>SD</th>
<th>GE Mean</th>
<th>SD</th>
<th>SP Mean</th>
<th>SD</th>
<th>$R^2$</th>
<th>p</th>
<th>Diff. Sign</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unhyphenated</td>
<td>0.5</td>
<td>0.9</td>
<td>0.9</td>
<td>1.7</td>
<td>0.7</td>
<td>1.1</td>
<td>0.02</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>Hyphenated</td>
<td>0.5</td>
<td>1.4</td>
<td>1.4</td>
<td>1.9</td>
<td>0.3</td>
<td>0.9</td>
<td>0.09</td>
<td>***</td>
<td>GE&gt;(FR=SP)</td>
</tr>
<tr>
<td>Solid</td>
<td>0.3</td>
<td>0.9</td>
<td>0.8</td>
<td>1.8</td>
<td>0.4</td>
<td>0.7</td>
<td>0.03</td>
<td>*</td>
<td>GE&gt;(FR=SP)</td>
</tr>
</tbody>
</table>

Note: SD = Standard Deviation; NS = Not Significant; * = p<0.05 ; *** = p<0.0001.

While there is no difference between groups for the unhyphenated category, the other two categories – hyphenated and solid – are significantly more frequent in GE than in FR or SP. Examples include *honey-moon* > *honeymoon*, *lorry-drivers* > *lorry_drivers*, *alarm-clock* > *alarm_clock*; *selfesteem* > *self-esteem*, *familylife* > *family_life*, *eachother* > *each_other*. This difference is most probably due to transfer from German which is characterized by a high number of compounds, which are usually written as one word, i.e. solid (*Schulkind*, *Strassenverkehr*, *Haustür*) but may in some cases be hyphenated (*CD-Laufwerk*, *Lufthansa-Pressesprecher*). Unhyphenated compounds do not exist in German (e.g. *rush hour* = *Verkehrsspitze*; *lorry driver* = *Lastwagenfahrer*).

5.2.2. Letter doubling

The results for letter omission (Doub12) and letter addition (Doub21) are strikingly different. As appears from Table 9, Doub12 errors are significantly more frequent in SP than in FR and GE. Over 80% of the Doub12 errors are found in the SP subcorpus.

<table>
<thead>
<tr>
<th></th>
<th>FR Mean</th>
<th>SD</th>
<th>GE Mean</th>
<th>SD</th>
<th>SP Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.2</td>
<td>0.6</td>
<td>0.2</td>
<td>0.6</td>
<td>1.8</td>
<td>2.1</td>
</tr>
</tbody>
</table>

Note: SD = Standard Deviation.

Here too the influence of the learners’ L1 appears as a likely factor, as shown by the list in Table 10 that compares the erroneous forms used by the SP learners to the corresponding

---

3 The Apost category has a similar profile but the limited number of errors (15 in the manual annotation) does not allow for a more fine-grained analysis.

4 We are grateful to Jennifer Thewissen for providing these examples.
words in Spanish. As the table shows, the letters involved in Doub12 errors are predominantly consonants.\(^5\) As pointed out by Bebout (1985, p.583), although consonant doubling is a difficulty for any learner or writer of English, it is particular treacherous for native speakers of Spanish who “are less used to paying attention to the presence or absence of doubled consonants or to making decision about doubling when writing”.

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>communication ; comunicación</td>
<td>impossibe; imposible</td>
</tr>
<tr>
<td>comunity ; comunidad</td>
<td>metallic; metálico</td>
</tr>
<tr>
<td>diference ; diferencia</td>
<td>officially; oficialmente</td>
</tr>
<tr>
<td>dificul; dificil</td>
<td>oportunity; oportunidad</td>
</tr>
<tr>
<td>eficient; eficiente</td>
<td>oposición</td>
</tr>
<tr>
<td>exagerate; exagerar</td>
<td>opresión</td>
</tr>
<tr>
<td>excelent; excelente</td>
<td></td>
</tr>
</tbody>
</table>

Conversely, the infrequency of this type of error in FR and GE can be explained by the lack of difference between German/French and English for this category. To take but one example, the English word *communication* corresponds to *communication* in French and *Kommunikation* in German.

As shown in Table 11, no such difference between the three learner groups can be observed as regards Doub21 errors.

<table>
<thead>
<tr>
<th>FR</th>
<th>GE</th>
<th>SP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>0.4</td>
<td>0.8</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Note: SD = Standard Deviation.

The influence of the L1 is mainly felt in the FR group where most of the errors can be related to the existence of an equivalent word with two letters in French (e.g., *educationnal* - *éducationnel*; *mentionned* - *mentionné*).\(^6\) The GE subcorpus contains a mixture of potentially L1-related errors (*detailed* – *detailliert*) and errors where L1 influence is absent (*beeing*, *neccessary*). The SP subcorpus contains three times fewer occurrences of Doub12 errors than Doub21 and in this case, the errors are clearly not L1-related (*carefull*; *compells*). This confirms Bebout’s (1985, p.579) study of Spanish learners which revealed significantly more errors involving the failure to double a consonant than the unnecessary doubling of one.

Cook (1997) gives further illustrations of L1 influence. He shows, for example, that Japanese learners are the only ones to confuse <l> and <r> as in *familiality* or *grobal*. However, his general conclusion is that “transfer from the L1 is less important than had been believed”. Our own results are more in keeping with an earlier study by James et al. (1993) which assigned a large proportion of errors (38.5%) to the learners’ L1. However, the issue is clearly far from settled and further research is needed to clarify the role of L1.

\(^5\) 98% of the Doub12 errors in the whole learner corpus involve consonants. In Doub21 errors consonants also predominate but to a lesser extent (78%).

\(^6\) Had the proficiency level of the learners been lower, one could have expected many more instances of this type of error, as is the case for the Doub21 errors in the SP corpus.
6. Conclusion

Our study has generated a number of interesting findings relating to L2 spelling errors and the role they can play in automatic text evaluation.

First, our study shows that spelling errors, whether detected manually or automatically, are good predictors of the quality of L2 texts and that prediction scores can be further improved by subcategorizing errors. The benefit derived from subcategorization is high in the case of manually detected but much lower in the case of automatically detected errors. The analysis shows that this is mainly due to the difficulties encountered by Word in detecting and correcting errors of the types Many and SplitW. Thirty-eight percent of the errors detected by Word which fall into the category of Many are overflags resulting from its unawareness of some proper names, abbreviations and foreign words. A better handling of some of these errors would make it possible to improve prediction accuracy. Our study also shows that Word encounters difficulties when it comes to identifying SplitW errors, only 31% of which are detected. The question that remains unanswered in this connection is whether such errors are more typical of L2 writers than native writers.

Our analyses also highlight the impact of the learners’ L1 on the success rate of spell checkers. This has implications for both the automatic grading of L2 texts and the customization of spell checkers for different categories of learners. As suggested by Haggan (1991, p.61), progress will only be made in the field if further studies are conducted “to add to our stock of cross-language error profiles”.

Overall our study strongly underlines the necessity to adapt spell checkers to L2 learners, a need that has been voiced by other researchers and led to some preliminary implementation. Mitton and Okada (2007), for example, have adapted a spell checker to some distinctively Japanese error patterns and report some promising results. Our study also underlines the need already identified by Granger and Wynne (1999) to take spelling errors into account when computing other indices of text quality such as lexical variation. On this point it is interesting to note that spelling errors seem to be a considerably higher predictor of text quality than lexical diversity (Yu, 2010). Admittedly, this comparison can only be tentative in view of differences in corpus and methodology. A more reliable comparison can however be made with Bestgen, Lories, and Thewissen (2010), which uses the same corpus and the same evaluation method. In that study, the indices for lexical variation account at best for 24% of the variance of the CEF score, which is markedly less than in the analyses presented here.

One major limitation of our study is that we only used one spell checker, viz. Word 2007. Further research is necessary to ascertain whether other spell checkers encounter the same difficulties as those highlighted in our study. Other error classifications such as those based on competence and performance put forward by Rimrott and Heift (2008) or those based on the sources of error (Cook, 1997; Haggan, 1991) also need to be assessed. There is no doubt that those classifications are linguistically more meaningful than those used in our study but it remains to be seen to what extent they can be automated.

References


