Word Processing in Spanish Using an English Keyboard: A Study of Spelling Errors

Nestor J. Rodriguez and Maria I. Diaz

Institute for Computing and Informatics Studies
University of Puerto Rico at Mayagüez
nestor@ece.uprm.edu, mdiazfigueroa@gmail.com

Abstract. This article describes a study of spelling errors made by writers while typing in Spanish using an English keyboard. The most important contribution of this study is the identification of a profile of errors made by writers using a word processor and an English keyboard to write in Spanish. The study revealed that a large number of the errors are related with words that have a character such as á, é, í, ó, ú or ñ. Another important finding of the study was that a substantial number of errors (approximately one third) are not corrected and that backspace was used to correct approximately two thirds of all the words corrected. The study supports the conclusion that the lack of straightforward support for characters such as á, é, í, ó, ú or ñ in the Spanish language can cause a significant number of errors.

Keywords: spelling errors detection, spelling errors correction, spell checking, word processing, Spanish writing.

1 Introduction

Grammatical correctness in writing is emphasized since a person starts learning to write because it is essential for clear transmission of thoughts, ideas, concepts or facts. Since we are unable to recall the correct spelling of every word of a particular language we usually seek help in order to achieve grammatical correctness in writing. The best resource for helping writers produce grammatically correct texts has been the dictionary. This resource is often used when the writer does not know the correct spelling of a word. However, the writer must has an idea of the spelling of the word to be able to search it through the dictionary. The effectiveness of the dictionary depends on the ability of the writers to detect words with incorrect spelling. Thus, there is a need for tools that can help them become aware of misspelled words an appropriately correct them. This type of help has been made available to writers through computer technology in the form of word processors with spell checking tools. Word processors with spell checkers are able to detect any pattern of letters that does not match a word in its dictionary. They can also suggest words for correcting the misspelled ones and automatically correct some of them.

There is no doubt that word processors with spell checking software have been tremendously effective in helping writers detect and correct misspelled words. However, this technology is not always 100% effective. A word processor could

N. Aykin (Ed.): Usability and Internationalization, Part II, HCII 2007, LNCS 4560, pp. 219–227, 2007.

indicate a correctly spelled word as being incorrect and a misspelled word as being correct. This problem is treated in a study conducted by D. Galletta et al. [1]. The study identified three outcomes that can result from using spell checkers: correctly identified errors, false positives errors and false negatives errors. A “correctly identified error” occurs when the spell checker detects and actual misspelled word. A “false positive error” occurs when spell checkers indicates that a correctly spelled word is misspelled. A “false negative error” occurs when the spell checker does not detect a word that is misspelled. The study revealed that when the spell checkers correctly identified errors, it helps low verbal people (people with less experience in a given language) to write almost as high verbal people. On the contrary, high verbal people that rely on spell checkers end up making more errors than when they don’t rely on them. This is mainly due to false negative and false positive errors.

Spelling errors are generated for a variety of reasons. In a study by Huang and Powers [2] six types of common errors were identified: typographical, homophone, grammatical, frequency disparity, learners, and idiosyncratic. Typographical errors typically manifest when a writer types a letter that is adjacent in the keyboard instead of the correct letter. Homophone errors are words that sound similar but they have a different meaning (i.e. piece and peace). Grammatical errors occur when a word with similar meaning is written instead of the intended word (i.e. “among” instead of “between”). Frequency disparity errors result when a writer tries to type the abbreviation of a word but types a similar unintended word instead (“their“ instead of “they’re”). Learner errors are those made by writers that are learning to write in a language that is not their first language. Idiosyncratic errors are those made for an unknown reason [3].

Spelling errors detection and correction has made its way through the World Wide Web as evidenced by two research studies. In the first study [4] the World Wide Web was used as a database to correct grammar and spelling errors. A client/server system was implemented in which the client sends a string or a phrase to the server and the server makes a search using a search engine on the web. The system counts the occurrence of that word or phrase in the web and lets the writer know the number of hits of the incidence of that word or phrase. It assumes that words with high frequency are very likely to be correct. In the second study Bolshavok and Gelbukh [5] proposed a solution for malapropism, writing words with similar sound but different meaning. For example, in the phrase "the boy is eating a peace of pizza" the word “peace” was used instead of “piece”. Collocations and a search engine were used to correct this kind of error. Collocations are phrases composed of words that co–occur for lexical rather than for semantic reasons. If a specific combination of words does not exist in the collocation database, a search engine searches that combination. It is assumed that if a combination of words occurs several times in the web, it is correct. The problem with these two studies is that they assume that text on web pages is grammatically correct.

Spelling errors detection and correction has not been well documented for the Spanish language. The Spanish language has some characters that are not used in languages such as English. These characters are not well supported by English keyboards. This lack of support may cause writers to make spelling errors. These issues with the Spanish language motivated the study presented in this document. The goal of this work was to study the spelling errors made by writers while writing in Spanish using an English keyboard and how these errors are corrected.

2 Methodology

As a first step in this research, a study was conducted in which twenty people were asked to write for an hour using MS Word. The participants were college students and recent college graduates. They were asked to write in Spanish something related with their lives. They were asked to type as they normally do. All the participants used Microsoft ® Office Word 2003.

The participants' interaction with the computer was recorded using the TechSmith Morae software. This software records and synchronizes user and system data for usability analysis. The software consists of three components: Morae Recorder, Morae Remote Viewer, and Morae Manager. Morae Recorder is the component of the software that captures the interaction of the user while he/she is using the computer. This part of the software was installed in the users' computers. This component can be configured to capture important activity from the screen, keyboard, and the mouse to be used in the analysis of the interaction. The Morae Remote Viewer allows experimenters to watch the interaction of a user remotely through Internet. For this study, this component was not used because it was not necessary to monitor the participants while interacting since the recording of the interaction provided the necessary data for the study. Finally, Morae Manager was used to analyze the recorded interaction of every participant. Morae Manager allows the researcher to place markers on the recording, so he/she can easily move to that point of the recording while reviewing it.

The Morae Recorder software was configured to record the keystrokes (input from the keyboard), screen text, and mouse clicks (highlight mouse cursor, left and right mouse clicks). It was set to record the writer activity for a period of one hour. The recording was done once for each participant.

The collected data was analyzed to identify the type of errors made by writers while typing. The behavior of the spell checker was also studied to identify words automatically corrected. In addition the strategies used by the participants to correct misspelled words were also studied. Pilot tests helped identify the types of errors commonly made by people while typing. Eleven errors categories were identified. Some of these errors such as typographical, homophone, transposition, extra letter, wrong letter and missing letter have been previously identified in the literature for the English language [3, 6]. Three of the error categories identified (Accent, Ñ Error and Special Accent) are non-existent in the English Language but very common in the Spanish language. The description of these errors follows.

- Extra Letter - The writer types an extra letter in a word (i.e. "estuudio" instead of estudio).
- Missing Letter – The writer does not type a letter in a word (i.e. "esudio" instead of "estudio").
- Homophone - Words that sound similar but they have a different meaning (i.e. "ciervo" and "siervo").
- Typographical – The writer types an adjacent letter in the keyboard instead of the correct one (i.e. "sin" instead of "son").
- Transposition & Disorder – Disorder corresponds to the case where the writer types a word with all its letters but in an incorrect order (i.e. Aoccdrnig instead of

According), while transposition is a special case of disorder in which two adjacent letters in a word are exchanged (i.e. “etsudio” instead of “estudio”).

- Wrong Letter – The writer types a wrong letter in a word (i.e. “estgdio” instead of “estudio”).
- Grammatical – The writer types a word with similar meaning instead of the intended word (i.e. “among” instead of “between”).
- Caps Lock – The writer is typing the first letter of a word in a sentence and turns on the Caps Lock and continues typing in capital letters.
- Ñ – The word has a ñ or Ñ but the writer does not type it and usually types a n instead.
- Accent – The writer types a word that must has a vowel with an accent and writes the vowel without the accent.
- Special Accent – The writer does not place an accent on a word that should has an accent or the writer places an accent on a word that should not has an accent (i.e. cambie, cambié). In both cases, both words are correct words of the dictionary but only one of them is correct in the context of a sentence.

3 Results

In order to make a fair analysis of the results the number of errors made by each of the participants was normalized by dividing the number of errors by the total number of

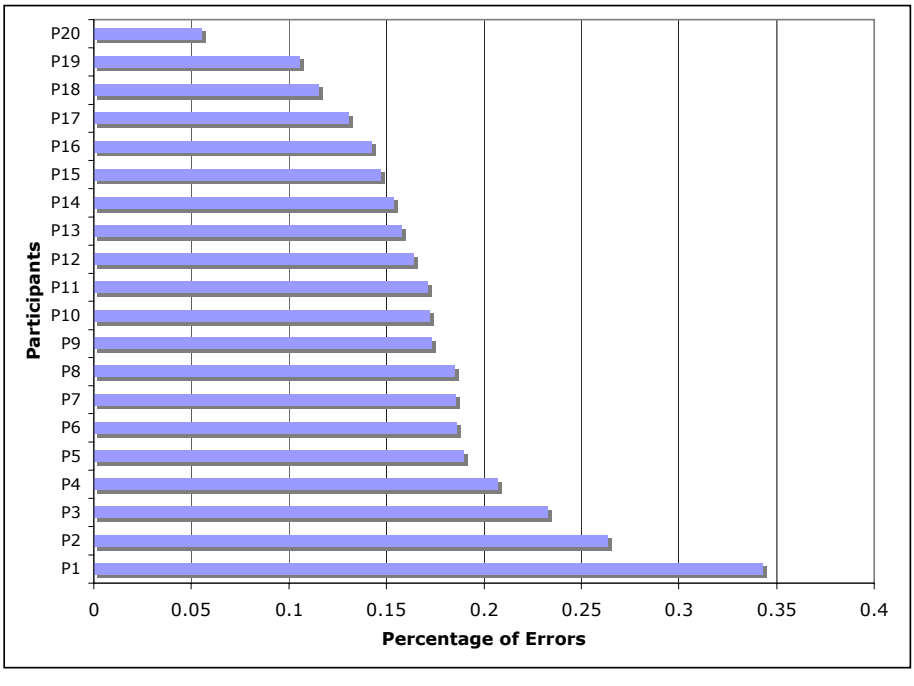

Fig. 1. Percentage of All Words that Were Errors Made by Each Participant

words written. Figure 1 shows the variability in the percentage of errors made by the participants. Results indicate that of all the words written by the participants an average of 17.31% were erroneous with a standard deviation of 6.16%. The maximum percentage of errors made by a participant was 34.29% and the minimum 5.78%.

Table 1 shows the mean, standard deviation, minimum and maximum values of the normalized number of errors made by the participants in each error category.

Table 1. Errors Made by the Participants Normalized by Total Number of Words Written

	Mean %	*Standard Dev. %*	*Minimum %*	*Maximum %*
Transposition & Disorder	0.91	0.61	0.07	2.22
Extra Letter	2.16	1.11	0.89	5.00
Wrong Letter	1.80	0.98	0.59	4.33
Missing Letter	2.29	1.15	0.91	5.47
Typographical	1.51	1.18	0.47	5.39
Homophones	0.48	0.50	0.00	1.72
Grammatical	0.02	0.05	0.00	0.19
Caps Locks	0.16	0.19	0.00	0.72
Accent	5.29	2.14	0.15	7.83
Special Accent	2.12	1.28	0.07	4.73
Ñ	0.57	0.44	0.00	1.49

Figure 2 presents the percentages of all errors that each category constitutes. These percentages were calculated using the averages of the normalized number of errors made by the participants for each category. Results revealed that the Accent error category is the one with the highest occurrence with over 30%, while the grammatical error category is the one with the lowest occurrence. The errors that are unique when writing in Spanish (Accents, Special Accents and Ñ) constitutes over 46% of all the errors made by participants.

An important action observed during the study was how the errors made by participants were corrected. Results indicate that an average of 73.0% of all the errors were corrected with a Standard Deviation of 19.0%. The participant that corrected the fewer number of errors fixed 29.6% of them while the one that corrected the most fixed 100.0% of them. The spell checker identified many of the errors but some writers did not review the document to fix them.

The participants used different ways to correct errors made while typing. Four techniques were identified: backspace, right click, spell checker and undo. The backspace technique was used to correct errors that were detected by the writers immediately while typing. The right click technique consisted in doing right click on the mouse on a word marked as incorrect by the word processor. When this is done the word processor displays a menu of words from which the writer can select the correct one if available and substitute the erroneous word. The spell checker technique is the case when the writer types in all the text and then go back to correct the erroneous words using the spell checking command in the Tools menu of Microsoft Word. The undo technique consists of hitting the undo button after the word processor automatically corrects a word that was typed correctly.

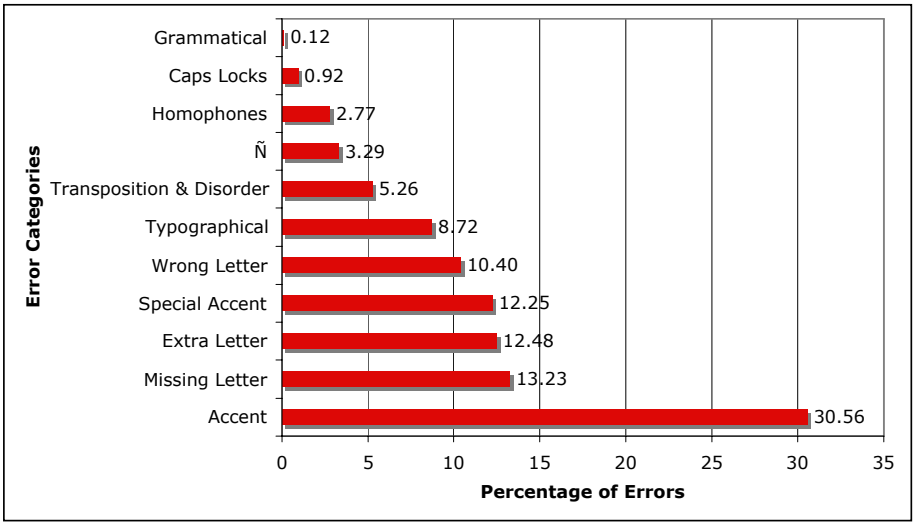

Fig. 2. Distribution of Errors Made by the Participants by Error Category

The percentage of errors made with each technique is presented on Figure 3. The results reveal that most of the errors were fixed with the backspace technique. Thus, most of the errors were fixed right at the moment they occurred.

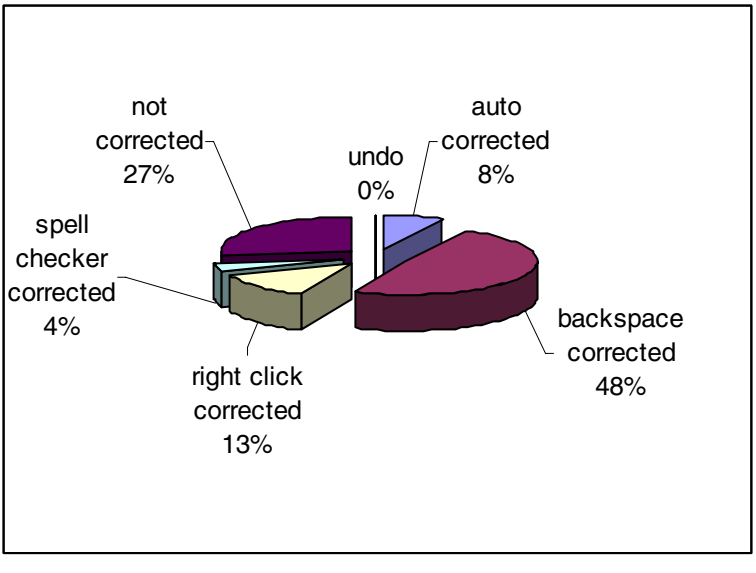

Fig. 3. Percentage of Errors Corrected with Each Error-Correction Technique

The study revealed that the percentage of errors corrected varies with the type of error. Error types that are easily identified by the writers or the word processor such

as Wrong Letter, Extra Letter or Missing Letter, Typographical and Transposition & Disorder errors exhibit high percentage of correction. On the other hand, error types that are not detected by the word processor or easily identified by the writers such as Special Accent and Ñ errors exhibit a lower percentage of correction.

Another interesting aspect observed during the study was how many words the spell checker fixed automatically. The results indicate that the word processor attempted to fix 7.57% of the errors automatically. However, approximately 13% of the words automatically corrected resulted in false positive errors [1]. The words were correct but the spell checker identified them as erroneous and substituted them with a words that ended up being incorrect.

4 Discussion

The most significant finding of the study presented in this document was that a large number of the errors made while writing in Spanish is related with words that have accented vowels (á, é, í, ó, ú) or a "ñ" character. The study revealed that the errors related with typing words with these characters (Accents, Special Accents and Ñ errors) constitute over 46% of all the errors made by the participants. For obvious reasons these types of errors do not happen when writing in English. The errors that are common in the English language are the other eight error types identified with this study. In our study, if the errors related with the special characters are removed, the combination of Transposition, Wrong Letter, Extra Letter or Missing Letter errors constitutes 77% of all the errors made by the participants. These results are very similar to the Damerau study [7], that revealed that these four error types constitute over 80% of all the error made by the writers.

The Accents, Special Accents and Ñ errors occur mostly because the methods for typing the vowels with the accent and the ñ are very cumbersome in most computer systems and the writers do not remember how to do it. In the Windows platform these characters can be typed by pressing the Alt key and a sequence of digits. Another way of entering these characters is using the English International Keyboard Setting. With this setting the writer holds the Right Alt key and presses the vowel they want to accent. It also allows writing the ñ letter just by holding the Right Alt key and the n letter. On a system with a Spanish keyboard, the writer places accents by pressing the accent key and the vowel to be accented. Also there is a ñ key on the Spanish keyboard. MS Word offers a Symbol map that includes the accent letters. In addition, Windows platforms have a Character Map under System Tools that includes special characters such as letters with accent.

Result revealed that approximately two thirds of the errors made by the participants were corrected. In the large majority of the cases the errors were corrected using the backspace key. The remaining words were corrected using the spell checking features of the MS Word.

The study revealed that the ways in which the errors are corrected varies with the types of errors. The large majority of Wrong Letter, Extra Letter or Missing Letter, Typographical, Transposition & Disorder errors are corrected by the writers using backspace. This is because these types of errors are easy to identify by the writers and they corrected most of them on the spot. All the Caps Lock Errors were corrected

either by the writers or the word processor. This is due to the fact that words in capital letters are easy to spot and writers can identified them easily. On the contrast, none of the Grammatical Errors were corrected. This is because Grammatical Errors are errors in which another word is written instead of the intended word. Since the word written is a correct word the word processor does not detect the error and it passes unnoticed by the writers.

About half of the Accent errors were corrected by the writers, most of them with the speller features of the word processor (spell checker, right click and autocorrect). The large majority of the Special Accent errors were not corrected. This is because the MS Word spell checker does not detect the large majority of this type of errors and in most of the cases the writer is unaware of the problem. Most of these errors were corrected with the speller features of the word processor. In the case of the Ñ errors about half of them were corrected by the writers and almost another half were left uncorrected. Some of these errors can be easily detected because of the presence of a tilde character (~) on the word. However, other cases are difficult to detect by the word processor and the writers because the word is written with an "n" instead of the "ñ" and that word is a correct word of the dictionary.

5 Conclusion

The most important contribution of this study is the identification of a profile of errors made by people using a word processor to write in Spanish. Eleven error categories were identified. The most significant finding was that a large number of the errors made are related with words that have a character such as á, é, í, ó, ú or ñ. These characters are very common in the Spanish language but non-existent in the English language. The errors caused when typing words with these characters were classified as Accents, Special Accents and Ñ errors. Most of these errors were made because the methods for typing the á, é, í, ó, ú and ñ characters were very cumbersome and the writer usually did not recall them. Thus, we conclude that the lack of straight-forward support for special character of languages such as Spanish can cause a significant number of errors.

Our study produced results consistent with the Damerau [7] study. If the Accents, Special Accents and Ñ errors are not considered, the percentage of the combination of transposition, wrong letter, extra letter and missing letter errors found in our study is very similar to the percentage reported in [7]. Thus, the Accents, Special Accents and Ñ errors are additional errors that are associated with the Spanish language. These findings supports the conclusion that writers can make a significant number of additional errors when writing in Spanish using Microsoft Word and a standard English keyboard than when they do it in English.

Another important finding of the study was that a substantial number of errors (approximately one third) are not corrected. Most of these errors pass undetected because the word processor does not detect them and thus does not provide a warning to the writers. From the study we identified that writers used basically three techniques to correct errors while writing in Spanish language: backspace, right click and spell checker. The backspace technique was used by the writers to correct approximately two thirds of all the words corrected. The writers used this technique

to correct most of the errors that they detected at the moment they made them. Thus, we conclude that writers correct most of the errors on the spot by recalling the correct spelling of the word.

The study revealed that the percentage of errors corrected varies with the type of error. Error types that are easily identified by the writers or the word processor such as Wrong Letter, Extra Letter or Missing Letter, Typographical and Transposition & Disorder exhibit a high percentage of correction. On the other hand, error types that are not detected by the word processor or easily identified by the writers such as Special Accent and Ñ exhibit a lower percentage of correction.

As it is documented in [8], with simple algorithms most of the Accent, Special Accent and Ñ errors can be detected. In addition the writer can be provided with alternatives to correct the error. The adoption of such algorithms by commercial word processor can improve error correction for Spanish writers.

Acknowledgments. This study was made possible in part by NSF grant EIA 99-77071.

References

1. Galletta, D., Ducikova, A., Everard, A., Jones, B.: Does Spell-checking Software Need a Warning Label? Communications of the ACM, 48(7), (July 2005)
2. Huang, J.H., Powers, D.: Large Scale Experiments on Correction of Confused Words. In: Computer Science Conference, Proceedings 24th. Australasian, pp. 77–82 (2001)
3. Powers, D.W.: Learning and Application of Differential Grammars. In: CoNLL97: Computational Natural Language Learning, ACL Association for Computational Linguistics, pp. 88–96 (1997)
4. Fallman, D.: The Penguin: Using the Web as a Database for Descriptive and Dynamic Grammar and Spell Checking. In: CHI'02 Extended Abstracts on Human Factors in Computing Systems (2002)
5. Boshakov, I.A., Gelbukh, A.: On Detection of Malapropisms by Multistage Collocation Testing. In: 8th International Conference on Applications of Natural Language to Information Systems (2003)
6. Durham, I., Lamb, D., Saxe, J.: Spelling Corrections in User Interfaces. Communications of the ACM. 26(10), (October 1983)
7. Damerau, F.J.: A Technique for Computer Detection and Correction of Spelling Errors. Communications of the ACM. 7(3), (March 1964)
8. Diaz, M.I.: A Study of Spelling Errors in Word Processing: Detection and Correction. M.S. Thesis, University of Puerto Rico-Mayaguez, Mayaguez, Puerto Rico (2006)