Spelling Error Patterns in Brazilian Portuguese

Priscila A. Gimenes
EACH / USP

Norton T. Roman*
EACH / USP

Ariadne M. B. R. Carvalho
Institute of Computing / Unicamp

Fifty years after Damerau set up his statistics for the distribution of errors in typed texts, his findings are still used in a range of different languages. Because these statistics were derived from texts in English, the question of whether they actually apply to other languages has been raised. We address this issue through the analysis of a set of typed texts in Brazilian Portuguese, deriving statistics tailored to this language. Results show that diacritical marks play a major role, as indicated by the frequency of mistakes involving them, thereby rendering Damerau's original findings mostly unfit for spelling correction systems, although still holding them useful, should one set aside such marks. Furthermore, a comparison between these results and those published for Spanish show no statistically significant differences between both languages—an indication that the distribution of spelling errors depends on the adopted character set rather than the language itself.

1. Introduction

Almost 50 years since Damerau's groundbreaking work was published (Damerau 1964), the figures he set for the proportion of spelling errors in typed text, along with their classification, still remain in use. According to Damerau, over 80% of all misspelled words present a single error which, in turn, falls into one out of four categories, to wit, **Insertion** (an extra letter is inserted), **Omission** (one letter is missing), **Substitution** (one letter is wrong), and **Transposition** (two adjacent characters are transposed). In fact, these very same figures lie at the heart of much current mainstream research on related topics, from automatic text spelling correction (e.g., Pirinen and Lindén 2014) to more advanced information retrieval techniques (e.g., Stein, Hoppe, and Gollub 2012) to optical character recognition techniques (e.g., Reynaert 2011). The reason for such popularity may rest in the simplicity of the approach, whereby numbers can be assigned

* EACH–USP, Arlindo Béttio, 1000. 03828-000. São Paulo, SP – Brazil. E-mail: norton@usp.br.

Submission received: 30 January 2014; revised submission received: 21 July 2014; accepted for publication: 18 September 2014.

doi:10.1162/COLI_a_00216

to the probability that a certain type of spelling error might take place, simply because that seems to be the frequency with which people make that kind of mistake.

Surprisingly enough, even though Damerau derived his statistics uniquely from texts in English, his findings are applied, with very little to no adaptation at all, to research in a range of different languages, such as Basque (e.g., Aduriz et al. 1997), Persian (e.g., Miangah 2014), and Arabic (e.g., Alkanhal et al. 2012). The question then becomes how appropriate these figures are when applied to languages other than English. In fact, some researchers have already noticed this potential flaw, and tried to adapt Damerau's findings to their own language, usually by modifying Damerau-Levenstein edit distance (Wagner and Fischer 1974) to match some language-specific data, but still without verifying the appropriateness of Damerau's statistics in the target language (e.g., Rytting et al. 2011; Richter, Stranák, and Rosen 2012), or by taking into account some feature of that language, such as the presence of diacritics, for example (e.g., Andrade et al. 2012).

In this article, we move a step further by analyzing a set of typed texts from two different corpora in Brazilian Portuguese—a language spoken by over 190 million people[1]—and deriving statistics tailored to this language. We then compare our statistics with results obtained for Spanish (Bustamante, Arnaiz, and Ginés 2006). As we will show, the behavior demonstrated by native speakers of these languages follow a very similar pattern, straying from Damerau's original findings while, at the same time, holding them still useful. As a limitation, in this research we only account for non-word errors, that is, errors that, once made, result in some non-existing word, and which may be detected by a regular dictionary look-up (Deorowicz and Ciura 2005).

As an indication of the impact of these results, we added a new module (Gimenes, Roman, and Carvalho 2014)[2] to OpenOffice Writer[3]—a freely available word processor—so as to reorder its suggestions list according to the statistics presented in this article. Within Writer, when typing *Pao*, instead of *Pão* ('bread' in Portuguese), the correct form comes in fifth place in the suggestion list, with *Poca*, *Pocá*, *Poça*, and *Próça* coming in the first four positions. With this module, it was observed that, for the first testing corpus, in 27.34% of the cases there was an improvement over Writer's ranking (i.e., the correct form was ranked higher in the list), whereas in 5.84% the new list was actually worse, and in 66.82% no changes were observed. In the second corpus, an improvement was observed in 19.90% of the cases, in 9.00% figures were worse, with 71.10% remaining unchanged. Some 10–21% increase in accuracy is not something to be neglected, especially at the price of changing weights in an edit distance scheme.

2. Related Work

One of the first efforts to derive statistics similar to those by Damerau in a language other than English was made by Medeiros (1995), who analyzed a set of corpora in European Portuguese. In his work, Medeiros found that 80.1% of all spelling errors would fit into one of Damerau's categories. He also noticed that around 20% of all linguistic errors would be related to the use of diacritics, meaning that they should be taken into account during string edit distance or probability calculations. By

1 According to 2010's demographic census: http://www.ibge.gov.br/home/estatistica/populacao/censo2010/caracteristicas_da_populacao/resultados_do_universo.pdf.

2 Available at http://ppgsi.each.usp.br/arquivos/RelTec/PPgSI-001_2014.pdf.

3 https://www.openoffice.org/.

linguistic errors, the author meant linguistically motivated errors, as opposed to those caused by slipping fingers in a keyboard, for example. However illustrative, Medeiros's findings suffer from a major drawback, to the extent that he relied on pre-existing lists of erroneous words in the related literature, along with handwritten errors made by high school students, with only part of the data coming from e-mail exchanges—that is, from actual typed text. The problem with this approach is that these data come without any frequency analysis, which renders them inappropriate for the application of computer techniques based on statistics, also making it hard to compare to any findings related to error frequency.

Spanish was another language found to confirm Damerau's findings (Bustamante, Arnaiz, and Ginés 2006). In that case, however, the authors made a much more detailed categorization of the errors they found, beyond the four classes originally proposed by Damerau, in a corpus of written (presumably typed) texts from Mexico and Spain. Still, a careful inspection into these categories (by grouping insertions and repetitions of the same letter, and taking errors related to the misuse of diacritics, capitalization, and cognitive replacements to be substitutions) leads one back to Damerau's four basic errors. In that case, a total of 86.7% of all errors were found to fit into one of these groups. The main difference, however, between both Damerau and Medeiros, and the work by Bustamante, Arnaiz, and Ginés (2006), is that the majority of spelling errors (54.9% of all errors, corresponding to around 49% of all single errors) were related to the misuse of diacritics, with special emphasis on their omission, which was responsible for 51.5% of all spelling errors (i.e., around 46% of all single errors) found in the corpus.

Finally, one of the latest additions to this list was made by Baba and Suzuki (2012), in which the authors analyzed spelling errors both in English and Japanese, with the last one typed on a Roman keyboard (that is, transliterated). Interestingly, the authors report almost no difference in the distribution of errors among Damerau's four classes, both for English and Japanese. By looking at the details, however, they found some specific errors in Japanese, which they attribute to the phonological and orthographic characteristics of the language. In fact, a breakdown analysis of the four classes show accentuated differences in the type of substitutions (that is, a vowel by another vowel, a vowel by a consonant, etc.), insertions (inserted vowel vs. consonant), deletions (whether or not the deleted character was a repetition of some of its neighbors), and transpositions (adjacent characters *versus* syllables) that were made.

As we will show in the following sections, this overall confirmation of Damerau's findings, but with remarked differences caused by each language's idiosyncrasies, seems to be the common behavior both in the related work and ours. This, in turn, may shed some light on the reasons why English-based spelling checking software does not seem to perform so badly in other languages, but still not as well as it could, should each language's own characteristics be taken into account.

3. Materials and Methods

Because we were interested in source texts written in Brazilian Portuguese, where authors were free to type with little to no intervention by spelling correction facilities, we decided to use the corpus presented in Roman et al. (2013), which comprises 1,808 texts, typed in by 452 different participants in a Web experiment, with a total of 62,858 words in length. As a source, this corpus has the advantages of (1) being produced by a number of different people of different ages, educational attainment, and background;

(2) being produced through a Web browser, without any help from spelling correction modules usually present in text editors; and (3) being freely available for download over the Web. We will refer to this corpus as C_1.

In order to verify the statistics gathered on C_1, in June 2011 we collected a corpus of blog posts from four different Web sites[4] that describe travel diaries along with comments by visitors. With a total of 26,418 words, spread over 192 posts, and being written in Brazilian Portuguese, this corpus presents the same advantages as C_1, except for the fact that we cannot verify participants' details, which means we cannot make any statement on the participants' distribution according to age, background, and so forth. Still, the main characteristics of availability and free text production remain. We will refer to this corpus as C_2. By comparing the statistics from both corpora we intend to reduce the effect of any bias related to writing style and characteristics of the participants.

Given that our goal was to identify non-word errors, we relied on a commercial text editor to highlight misspellings in both corpora. The highlighted words were then manually inspected by one of the researchers, and errors were grouped in categories. Besides Damerau's four original sets, we identified three extra categories: errors involving the use of diacritics, errors related to the use of the cedilla, and space-related errors. The first of these groups was inspired by the observation that the use of diacritics is a key issue in Portuguese (Andrade et al. 2012), potentially playing an important role in the making of spelling mistakes. Following Damerau, this group can be subdivided into four other subcategories: **missing diacritic**, **addition of diacritic**, **right diacritic applied to the wrong character**, and **wrong diacritic applied to the right character**.

The second group concerns the misuse of the cedilla—a special diacritical mark that in Portuguese can only be applied to the character c, resulting in ç. The reason for this character deserving a whole category of its own lies in the existence of two distinct keyboard layouts in the Brazilian market: ABNT-2 and US-accents. The main difference between both layouts is that whereas ABNT-2 presents a separate key for ç, that key does not exist on the US-Accents keyboard. Hence, while on ABNT-2 the user hits a single key to get a ç, on US-Accents it is a composite—two keys must be pressed: the single quotation mark and c. Ultimately, errors involving the cedilla may be interpreted both as a regular character mistake, or an error related to the use of diacritics, depending on the keyboard.

Finally, because space-related errors, although not so common (see Section 4), are usually difficult to deal with via spelling correctors, we decided to give them an independent set of categories. As a mistake, spaces have the annoying characteristic of not being handled by edit distance operations (Attia et al. 2012), thereby passing undisturbed by systems that use string distance-based ranking when giving alternative spellings to the user. The reason for this problem is that, by joining two words together, for example, the number of mistakes dramatically rises, if one takes that new string to be a single word. Along with these three categories, there is a "leftover" fourth one—**others**—comprising linguistic expression errors, first-letter capitalization, and wrong diminutive or augmentative.

4 http://bragatte.wordpress.com/.
http://guilhermebragatte.blogspot.com.br/.
http://forasteironairlanda.wordpress.com/tag/nomadismo/.
http://sussuemdublin.wordpress.com/2011/01/.

Table 1
Number of spelling errors, separated according to the number of mistakes per word in C_1.

	Errors per Word				
Error Category	**Single**	**Two**	**Three**	**Over Three**	**Total (%)**
Insertion	80	2	9	8	119 (10.45)
Omission	135	45	11	4	195 (17.12)
Transposition	40	2	0	0	42 (3.69)
Substitution	135	10	1	0	146 (12.82)
Missing diacritic	395	34	0	0	429 (37.66)
Addition of diacritic	17	2	0	0	19 (1.67)
Wrong diacritic in right letter	11	0	0	0	11 (0.96)
Right diacritic in wrong letter	9	0	0	0	9 (0.79)
Missing cedilla	46	23	0	0	69 (6.06)
Substitution by space	1	0	0	0	1 (0.09)
Space insertion	1	0	0	0	1 (0.09)
Space transposition	1	1	0	0	2 (0.17)
Missing space	26	1	0	0	27 (2.37)
Other	69	0	0	0	69 (6.06)
Total	966	140	21	12	1,139 (100)

4. Results and Discussion

Results both for C_1 and C_2 are shown in Tables 1 and 2.[5] Even though the total number of mistakes is almost the same in both corpora (1,139 for C_1 vs. 1,260 for C_2), the proportion of spelling errors in C_2 is more than twice as high as that of C_1. In this case, C_1 had a mean error rate of 1.81% (that is, 0.0181 error per word), while C_2 reached 4.77%. This difference may be due to the fact that C_1 was generated by students from a university, that is, people with a higher educational level. It may also have something to do with the fact that blogs are more conversation-like, which may lead people to relax the spelling rules they choose to follow, especially when it comes to errors involving diacritics. Still, despite this discrepancy, a two-sample Kolmogorov-Smirnov test showed no statistically significant differences between these corpora, on the distribution of the proportion of errors among categories, neither for the total number of errors ($ks = 0.4286, p = 0.1528$), nor for the number of errors per word ($ks = 0.4286, p = 0.1528$ for single, $ks = 0.2143, p = 0.9048$ for double, $ks = 0.2857, p = 0.6172$ for triple, and $ks = 0.1429, p = 0.9988$ for multiple errors).

As it turns out, errors related to the use of diacritics correspond to approximately half of all spelling errors in both corpora (overall, 47.15% in C_1 and 50.08% in C_2, corresponding to 49.48% of all single errors in C_1 and 58.84% in C_2), if we include in this sum cedilla-related errors. By taking the cedilla as a simple substitution, the numbers drop to 41.09% in C_1 and 45.63% in C_2 (overall), with 44.72% in C_1 and 57.08%

5 In these tables, for example, 22 insertions found in words with two errors mean that, from all errors found in such words, 22 were insertions, disregarding whether a word presents two insertions or an insertion along with some other type of error. To save space, rows filled with 0, that is categories with no examples, were removed. Space-related errors were also broken down into **Substitution by Space**, as in *pre qualified* instead of *pre-qualified*; **Space Insertion**, as in *w ord* instead of *word*; **Space Transposition**, as in *m yword* instead of *my word*; and **Missing Space**, as in *myword*.

Table 2
Number of spelling errors, separated according to the number of mistakes per word in C_2.

	Errors per Word				
Error Category	**Single**	**Two**	**Three**	**Over Three**	**Total (%)**
Insertion	54	20	23	22	119 (9.44)
Omission	186	65	37	16	304 (24.13)
Transposition	2	1	0	0	3 (0.24)
Substitution	32	17	16	8	73 (5.79)
Missing diacritic	519	41	11	2	573 (45.48)
Addition of diacritic	1	0	1	0	2 (0.16)
Missing cedilla	16	31	9	0	56 (4.44)
Other	101	15	14	0	130 (10.32)
Total	911	190	111	48	1,260 (100)

Table 3
Wrong word count, separated by the number of errors per word.

	With repetitions		Without repetitions	
Errors	**Words in C_1 (%)**	**Words in C_2 (%)**	**Words in C_1 (%)**	**Words in C_2 (%)**
1	966 (92.3)	911 (86.3)	442 (86.9)	513 (86.4)
2	70 (6.7)	95 (9)	62 (12.2)	67 (11.3)
3	7 (0.7)	37 (3.6)	4 (0.8)	11 (1.8)
Over 3	3 (0.3)	12 (1.1)	1 (0.1)	3 (0.5)
Total	1,046 (100)	1,055 (100)	509 (100)	594 (100)

in C_2 for single errors—still a substantial proportion. Interestingly, if one takes errors involving diacritics to be substitutions, then we end up with a total of 89.86% of all single errors falling into one of Damerau's categories in C_1, with 88.91% in C_2, thereby confirming Damerau's statistics. An analysis of the distribution of single errors among Damerau's four original categories (i.e., disregarding diacritic-related misspellings), and the distribution among the categories related to the use of diacritics also shows no statistically significant difference between both corpora ($ks = 0.5, p = 0.6994$ for Damerau's categories and $ks = 0.6, p = 0.3291$ for the diacritics—including cedilla—set).

Finally, regarding the number of misspelled words, we have once again confirmed Damerau's results, in that over 85% of all wrong words, be they repetition of existing words or not, have a single spelling error. Table 3 shows these results. In this table,[6] we present the figures both when taking word repetitions into account and when ruling them out (that is, when keeping the statistics only for new words). This is also in line with the results obtained for Spanish (Bustamante, Arnaiz, and Ginés 2006), in which it was found that 86.7% of all spelling errors were single errors.

6 Total = $966 \times 1 + 70 \times 2 + 7 \times 3 = 1,127$ errors, with another 12 distributed in the last category, for C_1, and $911 \times 1 + 95 \times 2 + 37 \times 3 = 1,212$ errors, with another 48 distributed in the last category, for C_2.

Table 4
Absolute wrong word frequency in (%) in Spanish and Brazilian Portuguese.

		Brazilian Portuguese	
Error Category	Spanish	C_1	C_2
Insertion	5.80	7.65	5.12
Omission	6.80	12.91	17.63
Transposition	0.00	3.82	0.19
Substitution	16.20	12.90 (17.30)	3.03 (4.55)
Missing diacritic	51.50	42.16 (37.76)	50.71 (49.19)
Addition of diacritic	2.90	2.49	0.10
Substitution of diacritic	0.50	1.05	0.00
Space related	2.40	2.77	0.00
Other	1.94	6.60	9.57
Multi-Errors	11.96	7.65	13.65

4.1 Comparison to Other Languages

For reasons noted in Section 2, we have chosen the work by Bustamante, Arnaiz, and Ginés (2006), on European and Mexican Spanish, as a benchmark against which to compare ours. Table 4 shows the comparison results. In this case, for the sake of comparison, and given that each research has a classification scheme of its own, we had both to reorganize some of the categories and present the figures in terms of number of words, instead of number of errors. Although results are shown considering cedilla-related errors to be cases of a missing diacritic, whenever appropriate the figures for taking such errors as simple substitutions are also presented within parentheses.

By comparing the data distributions in Table 4, one sees no differences[7] between our results (both for C_1 and C_2) and those for Spanish, no matter whether one takes cedilla-related errors to be a missing diacritic or a substitution. This, in turn, might indicate that such errors and their distributions are not related to the language or culture themselves, but instead to the set of characters (and corresponding diacritical marks) allowed within each specific language. Also, the data show the importance of taking diacritical marks into consideration when designing spelling correction systems for these languages. As it turned out, this type of error is responsible for over 40% of all wrong words, both in Brazilian Portuguese and Spanish.

5. Conclusion

In this article we presented some statistics collected from two different corpora of typed texts in Brazilian Portuguese. Our first contribution is a description of the error rate distribution among the four original categories defined by Damerau (1964), along with the distribution of errors related to the misuse of diacritical marks. Results show not only that Damerau's figures still hold, should we put aside diacritics, but also make a

7 With cedilla being a diacritic omission: $ks = 0.3, p = 0.7591$ both between Spanish and C_1, and between Spanish and C_2. With cedilla being a character substitution: $ks = 0.3, p = 0.7591$ both between Spanish and C_1, and between Spanish and C_2.

point about the importance such marks have on the misspelling of words in Brazilian Portuguese, as indicated by the frequency with which this type of error is made, thereby rendering spelling correction systems that rely solely on Damerau's results unfit for this language. On this account, a straightforward experiment with a commercially available text editor has shown a 10–21% improvement, depending on the testing corpus, in the ranking of suggestions for misspelled words.

As an additional contribution, we have shown that our results are very much like those obtained for Spanish (Bustamante, Arnaiz, and Ginés 2006), in that we could see no statistically significant difference on error distribution between these two languages. This, in turn, may be taken as an indication that the distribution of errors does not depend on language or culture, but instead on the character set that people are allowed to use. As such, it would not come as a surprise to find out that the same distribution can also be observed in other languages, such as Italian, French, and, to some extent, Turkish, for example, which make an intensive use of a similar set of characters. This is something that is worth investigating, and we leave it for future research.

References

Aduriz, Itziar, Iñaki Alegria, Xabier Artola, Nerea Ezeiza, Kepa Sarasola, and Miriam Urkia. 1997. A spelling corrector for Basque based on morphology. *Literary and Linguistic Computing*, 12(1):31–38.

Alkanhal, Mohamed I., Mohamed A. Al-Badrashiny, Mansour M. Alghamdi, and Abdulaziz O. Al-Qabbany. 2012. Automatic stochastic Arabic spelling correction with emphasis on space insertions and deletions. *IEEE Transactions On Audio, Speech, and Language Processing*, 20(7):2,111–2,122.

Andrade, Guilherme, F. Teixeira, C. R. Xavier, R. S. Oliveira, Leonardo C. da Rocha, and A. G. Evsukoff. 2012. Hasch: High performance automatic spell checker for Portuguese texts from the Web. In *Proceedings of ICCS-2012*, pages 403–411, Omaha, NE.

Attia, Mohammed, Pavel Pecina, Younes Samih, Khaled Shaalan, and Josef van Genabith. 2012. Improved spelling error detection and correction for Arabic. In *Proceedings of COLING-2012*, pages 103–112, Mumbai.

Baba, Yukino and Hisami Suzuki. 2012. How are spelling errors generated and corrected? A study of corrected and uncorrected spelling errors using keystroke logs. In *Proceedings of ACL-2012*, pages 373–377, Jeju Island.

Bustamante, Flora Ramírez, Alfredo Arnaiz, and Mar Ginés. 2006. A spell checker for a world language: The new Microsoft's Spanish spell checker. In *Proceedings of LREC-2006*, pages 83–86, Genoa.

Damerau, Fred J. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

Deorowicz, Sebastian and Marcin G. Ciura. 2005. Correcting spelling errors by modeling their causes. *International Journal of Applied Mathematics and Computer Science*, 15(2):275–285.

Gimenes, Priscila Azar, Norton Trevisan Roman, and Ariadne Maria Brito Rizzoni Carvalho. 2014. An OO writer module for spelling correction in Brazilian Portuguese. Technical Report PPgSI-001/2014, EACH-USP, São Paulo, SP – Brazil.

Medeiros, José Carlos Dinis. 1995. Processamento morfológico e correc cão ortográfica do português. Master's thesis, Instituto Superior Técnico – Universidade Técnica de Lisboa, February.

Miangah, Tayebeh Mosavi. 2014. Farsispell: A spell-checking system for Persian using a large monolingual corpus. *Literary and Linguistic Computing*, 29(1):56–73.

Pirinen, Tommi A. and Krister Lindén. 2014. State-of-the-art in weighted finite-state spell-checking. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 8404 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 519–532.

Reynaert, Martin W. C. 2011. Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal on Document Analysis and Recognition*, 14(2):173–187.

Richter, Michal, Pavel Stranák, and Alexandr Rosen. 2012. Korektor—a system for contextual spell-checking and diacritics completion. In *Proceedings of COLING-2012*, pages 1,019–1,028, Mumbai.

Roman, Norton Trevisan, Paul Piwek, Ariadne Maria Brito Rizzoni Carvalho, and Alexandre Rossi Alvares. 2013. Introducing a corpus of human-authored dialogue summaries in Portuguese. In *Proceedings of RANLP-2013*, pages 692–701, Hissar.

Rytting, C. Anton, David M. Zajic, Paul Rodrigues, Sarah C. Wayland, Christian Hettick, Tim Buckwalter, and Charles C. Blake. 2011. Spelling correction for dialectal Arabic dictionary lookup. *ACM Transactions on Asian Language Information Processing*, 10(1):3:1–3:15.

Stein, Benno, Dennis Hoppe, and Tim Gollub. 2012. The impact of spelling errors on patent search. In *Proceedings of EACL-2012*, pages 570–579, Avignon.

Wagner, Robert A. and Michael J. Fischer. 1974. The string-to-string correction problem. *Journal of the ACM*, 21(1): 168–173.