Spelling Error Patterns in Spanish for Word Processing Applications

Flora Ramírez Bustamante, Enrique López Díaz

Microsoft Corporation
One Microsoft Way, Redmond WA, 98052-6399, U.S.A.
{florarb, eldiaz}@microsoft.com

Abstract

This paper reports findings from the elaboration of a typology of spelling errors for Spanish. It also discusses previous generalizations about spelling error patterns found in other studies and offers new insights on them. The typology is based on the analysis of around 76K misspellings found in real-life texts produced by humans. The main goal of the elaboration of the typology was to help in the implementation of a spell checker that detects context-independent misspellings in general unrestricted texts with the most common confusion pairs (i.e. error/correction pairs) to improve the set of ranked correction candidates for misspellings. We found that spelling errors are language dependent and are closely related to the orthographic rules of each language. The statistical data we provide on spelling error patterns in Spanish and their comparison with other data in other related works are the novel contribution of this paper. In this line, this paper shows that some of the general statements found in the literature about spelling error patterns apply mainly to English and cannot be extrapolated to other languages.

1. Introduction

This paper reports results on a research on error types in Spanish found in real-life texts. These human produced errors can be grouped in two main types: cognitive errors (i. e. misconceptions about orthographic rules or lack of language knowledge), and errors related to typing mistakes (motoric errors when using a computer keyboard).

These results have been applied to the implementation of a spell checker that detects context-independent misspellings in general unrestricted texts. This speller provides isolated-word error correction by offering a set of candidate corrections that are close to the misspelled word. The technique that underlies the spelling correction algorithm is that of *minimum edit distance* (number of additions, substitutions, omissions, and transpositions of two adjacent characters), as explained in Damerau (1964). Every operation is rated with a score. The best suggestion is the one with the lowest score. In this context, the elaboration of the error typology aimed at (a) finding a categorization based on the frequency of the above mentioned four edit operations in Spanish, and (b) discovering the most frequent error/correction pairs. These two factors allowed us to improve the ranking of suggestions offered by the speller as candidate corrections for misspellings.

This paper discusses the most common generalizations about spelling error patterns, and presents further results about frequency and types of misspellings found in real-life corpora. Our findings confirm Damerau's results in that the majority of errors tend to be single instances of insertions, deletions, substitutions and transpositions. We have also found that the majority of misspelling errors in Spanish are:

1. omissions (mainly, of accent or one character),
2. substitutions of lower case for upper case at the beginning of a proper noun,
3. cognitive errors,
4. addition,
5. substitution of one character, and
6. transposition.

Other conclusions are:

a) the percentage of errors in the first letter of a word is higher than that reported in other studies,

b) keyboard adjacency effects are less important than other factors, and

c) there is no necessarily direct correlation between frequency of a character in a corpus and its chances to be an error.

2. The Error Data Collection

The corpus used contains over 8 million words of edited and unedited texts. It comprises three different sets of texts. The first one, with about 4 million words, is a balanced set of edited and unedited texts. The second one is a highly edited corpus with about 2 million words. The third one, with about 2 million words, contains unedited data from the Original Works Creation (OWC) sites.[1] Table 1 below shows the breakdown for each set.

The corpus was analyzed morphologically using a knowledge-based syntactic parser (the parser which underlies the Microsoft Office grammar checkers developed by Microsoft). This task produced a list of unknown words.[2] The total number of unique unknown words discovered in the corpus was 76K, of which almost 27K (over 35%) were unique misspellings. The rest of them (65%) were not misspellings, but proper nouns, foreign words, derived words, etc. The misspellings were manually revised, and a correction was assigned to each of them based on the context, with the result that the same misspelled word could have more than one correction and therefore appear in more than one error pair. Every misspelling was classified as belonging to a more general class of errors (i.e. error type). The frequency of each error type and error pair occurrence was calculated.

1 The Original Works Creation (OWC) sites were created to gather unedited text. The participants from Madrid and Mexico City chose from several topics and wrote for about 30 minutes. The configuration of the software and the hardware was set in such a way that authors were only allowed to make minimal immediate corrections. The files have only been sentence-separated, and were not spell-checked or grammar-checked.

2 Our goal was to find the most common spelling error patterns in Spanish. For this reason, we started looking for unknown words in the corpus. In this sense, it was not the aim of this study to look for errors in which a correctly spelled word is substituted for another word that exists in the same language (e.g., *from-form* in English, and *más-mas* in Spanish).

	# of Words	# of Unique Unknown Words	# of Unique Mispell.	# of Mispell.
Balanced (BAL)	4,159,886	32,742	6,634	14,096
Highly edited (HE)	1,933,586	11,678	359	584
OWC – Mexico (MEX)	305,529	8,729	5,641	28,117
OWC – Madrid (MAD)	1,782,657	22,777	14,041	31,590

Table 1. Breakdown of the corpora

3. The error typology

We derived from the corpus an initial classification of 142 error types that exhaustively exploit combinations of the four editing operations. This classification incorporates also multi-error misspellings, position of the error inside of a word, character distance from the intended word, and edit operation type. Additionally, we have introduced finer-grained categorization on idiosyncratic features like diacritics, space, and case. A new distinction was also made in order to collect frequency information on cognitive errors of the phonetic type (substitutions of a phonetically correct but orthographically incorrect letter or sequence of letters for the intended word) (Kukich, 1992). Table 2 shows some examples from this initial classification.

Error type	Example	%
Omission of diacritics	*dia→día*	51.5
Omission one character	*mostar→mostrar*	6.8
Capitalization beginning word	*windows→Windows*	6.2
Cognitive errors	*biene→viene*	5.9
Addition of one character	*aereopuerto→aeropuerto*	4.7
Substitution one character	*calavara→calavera*	4.1
Addition of diacritics	*fuí→fui*	2.9
Omission space	*esque→es que*	2.0
Transposition one character no beginning word	*Interpetración→interpretación, movimineto→movimiento*	1.7
Repetition of same letter by addition	*dirrección→dirección*	1.1
Capitalization whole word	*fifa→FIFA*	1.3
Substitution of diacritic character	*informaciòn→información*	0.5
Addition space	*bue na→buena*	0.4
Substitution of the letter with diacritic	*fotografícas s →fotográficas*	0.1
Transposition one character beginning word	*haora→ahora*	0.1
Transposition space	*hayq ue→hay que*	0.04
Total single error misspellings		89.34
Multi-error: substitution (including diacritics) + addition, omission	*paguina→página, muestame →muéstrame, informacio→información, comenze →comencé*	3.6
Multi-error: other	*Nesecitaria→Necesitaría*	2.8
Multi-error: capitalization + addition, omission, substitution (including diacritics)	*jose →José*	1.1
Multi-error: addition + substitution (including diacritics)	*aficcion→afición, desfraccmentar→defragmentar*	0.6
Multi-error: repetition of same letters by addition	*tratataba→trataba*	0.4
Multi-error: substitution vowel + addition or omission accent	*sombolo→símbolo, sabámos→sabemos*	0.06
Multi-error: omission space	*alomejor→a lo mejor*	0.03
Multi-error: double omission of diacritics	*linguistica→lingüística*	0.02
Total multi-error misspellings		8.61
Rest	*serguiovamos →seguro vamos, verdadramnet →verdaderamente*	2.05

Table 2: Error type distribution in the corpus

This typology shows that a full 51% of the misspellings found in our corpus are omissions of a diacritic on a vowel.[3] An analysis of the error rate variation across the different sets of texts ratifies that the high frequency of this error type is not due to a lack of balance among the various texts (Table 3).

Error type	BAL	HE	MEX	MAD
Omission of diacritics	57%	33%	52.6%	52%

Table 3. Breakdown of diacritic omissions

Conversely, capitalization at the beginning of the word is remarkably more frequent in one of the sets than in the others (as shown in Table 4).

Error type	BAL	HE	MEX	MAD
Cap. word	1.10%	1.8%	14.22%	2.15%

Table 4. Breakdown of word initial capitalization

Regarding cognitive errors, in Spanish this error type mostly consists of substitutions of a phonetically correct but orthographically incorrect sequence of letters (i.e. homophones, mistakes on similar phonetic pairs such as *b-v*, *s-x*, *c-s*, *ll-y* for instance, as in **archibo-archivo*, **estención-extensión*, **llendo-yendo*). This type of error is common in single and multi-error patterns. A breakdown of the total number of cognitive errors found in the corpus shows that unedited sets exhibit more frequently this error type than edited sets (Table 5). Most of cognitive errors are instances of substitutions (**biene-viene*), omission (**acer-hacer*), and addition (**hacerca-acerca*).

Error type	BAL	HE	MEX	MAD
Cogn. Errors	2.17%	6.67%	4.67%	4.62%

Table 5. Breakdown of cognitive errors

Although it could be argued that the cause of the substitution of a *b* for *v* could be a keyboard adjacency effect rather than a cognitive error, the overwhelming high frequency of this error pair (614; see Table 6 below), in comparison to the lower frequency rate of other adjacent keys (**i-u* (51) and **u-i* (38), for instance) would remain unexplained.

On the other hand, the source of errors involving diacritics and case could be considered a misconception on the part of the writer, and, hence, cognitive errors. This fact would imply that a full 63% of the errors found in the corpus would be instances of cognitive errors.[4] This figure suggests that prioritizing some instances of this error type may be very helpful for a spelling correction application for Spanish.

In addition, Figures in Table 2 show that in Spanish omission is the most common spelling error,[5] followed by substitution, addition, and transposition in that order.

4. Findings on Error Patterns

The error patterns in Spanish seem to corroborate only partially some of the pattern findings reported by Kukick (1992). Kukick's conclusions are the following:

1. most errors (i.e. roughly 80%) tend to be single instances of insertions, deletions, substitutions, or transpositions,
2. most errors tend to be within one letter in length of the intended word,
3. few misspellings occur in the first letter of a word,
4. strong keyboard adjacency effects,
5. strong letter frequency effects, and
6. phonetic errors are harder to correct because they result in greater distortion of the misspelled string.

In the following sections we revisit these statements and compare them with our data from Spanish.

4.1. Most errors are single letter errors

The vast majority of errors found in the corpus, as shown in Table 2, are single error misspellings (over 89%).[6] Multi-error misspellings are less than 9%. There is an insignificant remaining percentage of noise related to spaces in multiple locations, extreme multi-error words and indecipherable strings of characters. Demarau (1964), Pollock and Zamora (1984) and Ren and Perrault (1992) report similar rates.

4.2. Most errors tend to be within one letter in length of the intended word

Our data also confirm this statement: 77% of misspellings are in length distance 0; 12.2% are 1 character shorter than the intended word, and 9% are 1 character longer than the correction. See Figure 1. These findings are consistent with the frequency of error types in Spanish: omission of diacritics, substitution of upper case and of one character, followed by omission and addition of one character.

3 Ren & Perrault (1992) also found that accents are a special class of errors in French. They subdivide these errors into four types: accent insertion, accent deletion, substitution of one accent for another and repositioning of the accent.

4 Veronis (1988) reports the importance of phonetic errors for French. See Kukich (1992) and references therein for similar reports for Dutch and English.

5 There are other studies reporting omissions as the most frequent error in other languages as well. See (Pollock and Zamora, 1984) and (Ren and Perrault, 1992) for similar statistics.

6 In the following sections, figures reported are based only on the total number of single error misspellings.

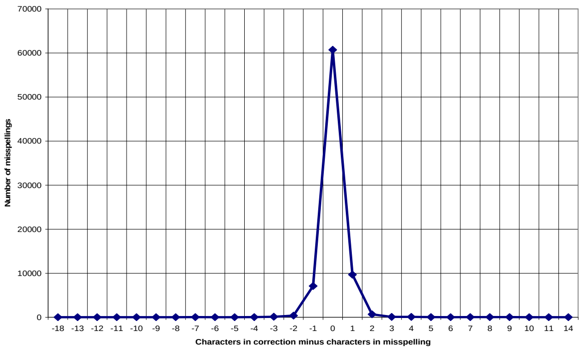

Figure 1: Comparison of the length of the error and the length of the intended word

4.3. Few misspellings occur in the first letter of a word

Based on the results of studies carried out by Pollock and Zamora (1984) and Yannakoudakis and Fawthrop (1983), Kukich (1992) considers that few misspellings occur in the first letter of a word (see also Ren & Perrault, 1992). In Spanish, given the high rate of substitution of capital letters in proper nouns, first-position errors total more than 6% of the misspellings, the third most frequent type of error. Error typologies from other languages provide interesting figures for comparison.[7] Cf. Figures 2-4.

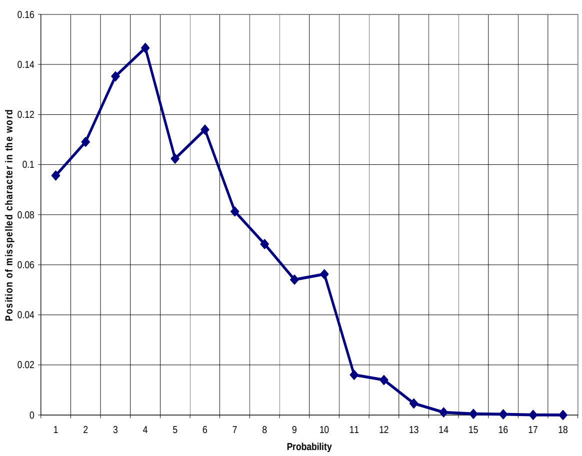

Figure 2: Probability of a misspelling happening in a given position in a word in Spanish

French surpasses German in first-position misspellings (because of substitution of capital letter in proper nouns, omission of diacritic and omission/substitution of first character, e.g. **marcel-Marcel*, **eglise-église*, **aper-taper*, **fint-vint*). The most probable position for misspellings in Spanish is around the 3rd, 4th or 5th character in the word. In German, the 4th character is the critical position for a misspelling. Zamora and Pollock (1984) report that 23% errors occurred in 3rd position in English.

7 With a similar process than the one done for Spanish, a typology of errors was elaborated as well for French and German. For French we derived 26,680 misspellings occurrences; for German, 12,765.

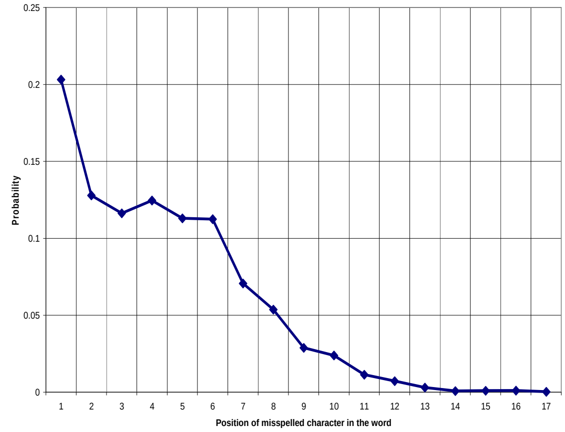

Figure 3: Probability of a misspelling happening in a given position in a word in French

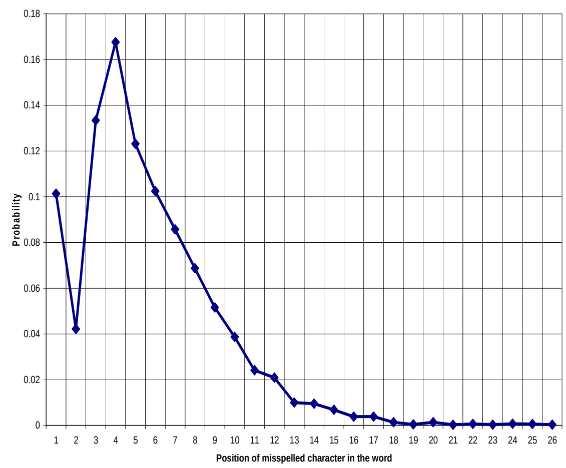

Figure 4: Probability of a misspelling happening in a given position in a word in German

4.4. Strong keyboard adjacency effects

Keyboard adjacency typing errors are caused by hitting an adjacent key on the keyboard instead of the intended one, or by hitting two keys at once instead of one. Typing mistakes have also been found in the corpus, although their frequency is not comparable to that of cognitive errors. Table 6 shows the number of times a given letter is substituted for another letter, space (SP) or nothing (NULL).[8] Higher frequencies appear associated to cognitively based misspellings.

Let's observe errors in which the character *c* is substituted, for instance. The character *c* is frequently misspelled as the topologically close in the keyboard, but not adjacent, *s* (399 times) and *z* (141 times). These are differ-

8 We have constrained the table to 35 characters from the 102 ones we worked with because of lack of space. The complete list is: <SPACE>, &, ‘, ,, -, ., ;, ?, A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z, `, a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, <NULL>, ¡, ¢, ¤, º, Á, Â, Ä, Ç, É, Ë, Ì, Í, Ï, Ñ, Ó, Ö, Ú, Û, Ü, à, á, ä, ç, è, é, ê, ë, ì, í, ï, ñ, ò, ó, ô, ö, ù, ú, ü, ý, ƒ.

ent characters that have the same pronunciation in some dialects. Substitutions of strictly adjacent keys *d* (20), *v* (8), *x* (2) for *c*, however, are much more scarce.

An interesting example are the adjacent keys *m* and *n*. The substitution of *n* for *m* happens 238 times and that of *m* for *n*, 122 times. There is a specific orthographic rule in Spanish that says that before a *b* or a *p* the only nasal letter that can be written is *m*, hence the expected higher amount of *n* for *m* error substitutions.

This frequency is similar to other substitutions totally unrelated to keyboard effects, as the ones with vowels: *e* (292 times)) that was wrongly substituted for vowel *a*, and *a* for *e* (306 times). These figures are indicative enough to suggest that some additional explanation is needed for adjacent keys that get a significant higher score than the other adjacent keys (cf. Armenta et al., 2003).

The keys adjacent to *a*, such as *s* , *q* and *z* were typed 32, 7 and 2 times, respectively, instead of *a*. *á* was typed 310 times when the intended character was *a*. Note that the reverse case (i.e. substitution of *a* (unaccented vowel) for *á* (accented vowel)) appears 7,584 times in the corpus.

Errors concerning accented vowels in Spanish could be attributed to mechanical reasons due to the configuration of the Spanish keyboard: two key strokes, one for the accent and another one for the vowel, are necessary to type an accented vowel and the user could miss one of the strokes. French refutes this hypothesis: we found that in French, the first, second and fourth most frequent single letter errors are related to *e* and its combination with diacritics *é* and *è*. Errors involving *é* represent an 18% of the errors; errors involving *e* represent 9%, and errors involving *è* amount 3.5%. The French keyboard, however, does have its own specialized keys for *é* and *è*, so misspellings on these accented vowels cannot be attributed to the omission of a keystroke for the accent.

Another reason to discard keyboard motivations for the high numbers of misspellings in Spanish that are related to accented vowels is that when an accented vowel is misspelled, it is substituted by its unaccented counterpart in almost all the cases (see Table 2 above). Only a very small number of errors in accented vowels are due to the omission of the vowel with the result of a diacritic by itself (´ instead of *á*), or to the insertion of the wrong diacritic that is in a key close to the key for the ´ accent (*ë* instead of *é*, *ì* instead of *í*). If errors in accented vowels were related to keyboard mistyping, one would expect that the omission of the diacritic were less frequent and the omission of the vowel or the insertion of a wrong diacritic were more frequent. These facts imply that when diacritics in a language represent sounds close to those of the characters without diacritics, missing diacritics cannot be reduced to an explanation based on merely keyboard reasons.

In summary, although keyboard adjacency effects are indeed relevant for a taxonomy on the nature of human misspellings, the frequencies in Table 6 suggest that other factors can be much more relevant.

4.5. Strong letter frequency effects

This statement could not be verified either. An absolute correlation between frequency of a character in a corpus and its chances to be an error in Spanish was not found. If this were the case, we should have expected that most frequent errors should have happened on the most frequent characters in the corpus. The most frequent characters in a Spanish edited corpus are unaccented vowels, and the consonants *l*, *r*, *t*, *d*, *n* and *s*, as it is shown in Figure 5. However, most errors affect the least frequent characters: accented vowels and capital letters.

4.6. General agreement that phonetic errors are harder to correct

Kukich (1992) also states that phonetic errors are harder to correct because they result in greater distortion of the misspelled string. However, this fact depends on how close or how far the orthographic system of a given language is from its phonetic system. In Spanish, correction of phonetic errors usually implies one of the edit operations, mainly substitution of one/two letters, as we have already explained in previous sections. Veronis (1988) reports the high complexity of this task for French when using a correction technique based on an *edit distance* model (cf. the Spanish error/correction pair **ipotenusa-hipotenusa* vs. the French pair **ippeautaineuz-hypotenuse*).

5. Conclusions

This paper shows that some of the general statements found in the literature about spelling error patterns apply mainly to English and cannot be extrapolated to other languages. It also reveals the beneficial and productive effects of working with real-life data in the targeted language for the elaboration of an error typology. This type of data provides clues about which error patterns should be promoted to enhance the generation of suggestions lists for corrections and, if dealing with context-dependent errors, to promote those flags related to the most common error patterns. It is obvious that errors involving homophony in Spanish are the most likely contextual errors to be flagged.

6. References

Armenta, A., Escalada, J. G., Garrido, J. M. & Rodríguez, M. A. (2003). Desarrollo de un corrector ortográfico para aplicaciones de conversión texto-voz. In *Procesamiento del Lenguaje Natural*, 31, pp. 65-72.

Damerau, F. (1964). A technique for computer correction of spelling errors. In *Communications of the ACM*, 7 (3), pp. 171-176.

Kukick, K. (1992). Techniques for automatically correcting words in text. In *ACM Computing Surveys*, 24 (4), pp. 377-439.

Pollock, J. J. & Zamora, A. (1984). Automatic spelling correction in scientific and scholarly text. In *Communications of the ACM*, 27 (4), pp. 358-368.

Ren, X. & Perrault, F. (1992). The typology of unknown words: an experimental study of two corpora. In *Proceedings of Coling*, pp. 408-414.

Veronis, J. (1988). Morphosyntactic correction in natural language interfaces. In *Proceedings of the 12th conference on Computational linguistics*, pp. 708-713.

Yannakoudakis, E. & Fawthrop, J. D. (1983). The rules of spelling errors. In *Information Processing & Management*, 19 (2), pp. 87-99.

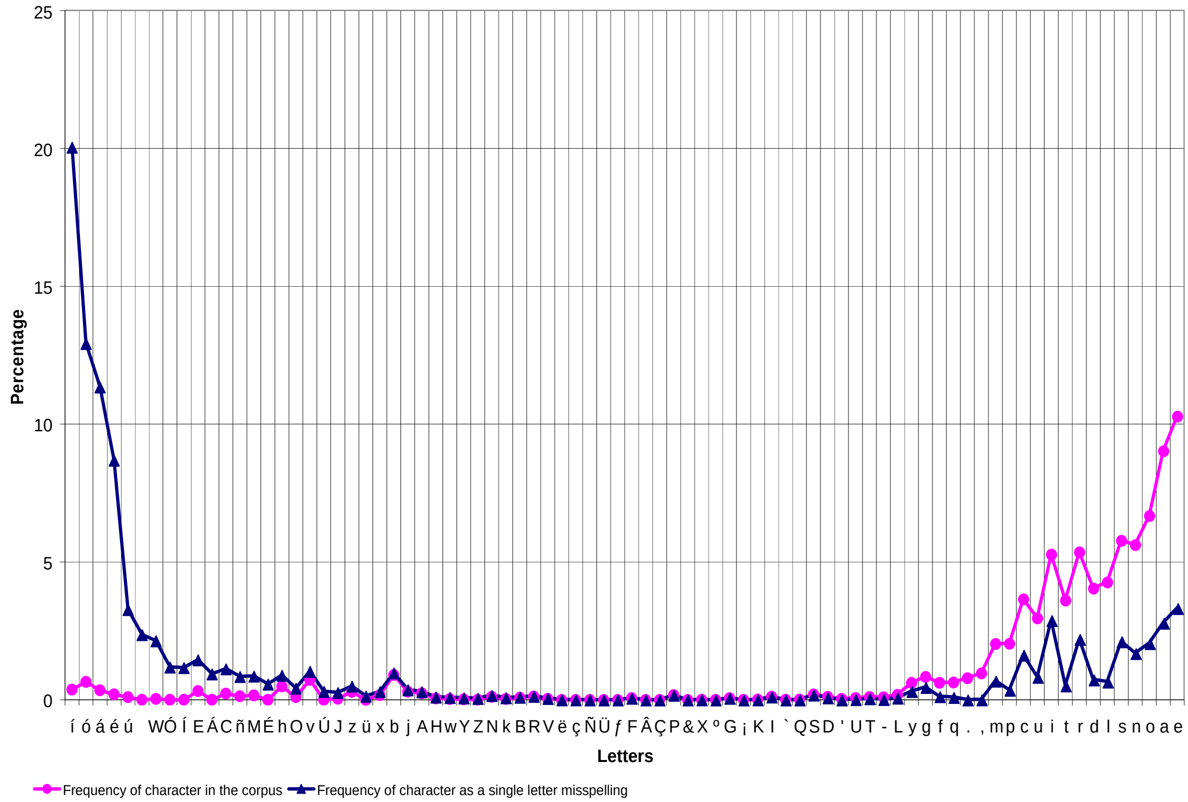

Figure 5: Frequency of characters in the corpus and frequency of misspelled characters

X \ Y	NULL	SP	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	á	é	í	ó	ú	ü	ñ
NULL	0	1606	655	15	294	229	726	10	14	314	526	4	0	155	75	588	370	99	6	889	586	138	209	18	1	4	3	11	49	82	12	27	5	1	6
SP	357	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a	518	0	0	0	0	2	306	0	0	0	13	0	0	0	0	2	85	0	2	0	40	0	6	0	0	0	0	1	7584	4	0	1	0	0	0
b	108	0	0	0	1	2	0	0	6	0	0	1	0	0	2	10	0	9	0	1	0	1	0	614	1	0	0	0	0	0	0	0	0	0	0
c	179	0	0	4	0	16	2	3	13	0	0	2	1	1	0	3	0	1	1	6	140	3	0	3	0	2	0	5	0	0	0	0	0	0	0
d	163	0	1	45	20	0	3	3	1	0	0	0	0	2	0	7	1	0	0	2	28	24	0	3	0	0	0	5	0	0	0	0	0	0	0
e	538	0	292	0	6	0	0	0	0	0	77	0	0	1	0	0	47	0	0	17	6	0	6	0	0	0	0	0	8	5776	1	0	2	0	0
f	57	0	0	1	1	5	0	0	6	0	0	1	0	0	0	0	0	0	0	4	1	2	0	2	0	0	0	0	0	0	0	0	0	0	0
g	32	0	0	4	9	1	0	6	0	2	0	162	0	0	0	0	0	0	2	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
h	287	0	0	0	0	0	0	0	2	0	0	6	0	1	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
i	371	0	5	0	0	0	75	0	0	0	0	0	0	3	0	1	77	0	0	1	1	0	51	0	0	0	39	0	1	12	13590	5	0	0	0
j	36	0	0	0	0	0	0	2	229	3	3	0	0	3	0	0	0	0	0	0	3	0	0	0	0	0	6	0	0	0	0	0	0	0	0
k	55	0	0	0	18	0	0	0	0	1	0	2	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
l	226	0	0	1	0	9	0	0	1	2	3	1	0	0	1	7	1	0	0	23	1	7	0	0	0	0	0	0	0	0	0	0	0	0	3
m	142	0	0	3	3	0	0	0	1	0	0	0	0	1	0	122	0	2	0	0	0	0	1	0	0	0	0	0	0	0	9	0	0	0	1
n	385	0	0	9	8	6	0	0	1	0	0	0	0	2	238	0	0	0	0	11	21	0	0	9	0	1	0	1	0	0	0	0	0	0	298
o	284	0	63	0	0	0	26	0	0	0	57	0	0	1	0	1	0	12	0	0	0	0	40	0	0	0	0	0	1	1	1	8641	2	0	0
p	122	0	0	10	1	1	0	0	1	0	0	1	0	0	0	0	20	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0
q	59	0	7	0	12	0	1	0	5	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	3
r	492	0	1	0	2	7	17	1	5	2	0	0	0	31	4	13	0	0	0	0	45	25	0	0	0	0	0	0	0	1	1	0	0	0	0
s	430	0	32	2	399	27	2	0	2	0	0	0	0	3	0	8	2	0	0	33	0	2	0	0	0	138	0	292	0	0	0	50	0	0	0
t	145	0	0	0	5	30	2	2	0	0	2	0	0	1	0	3	0	0	0	24	4	0	0	1	0	0	4	1	0	0	0	0	0	0	0
u	141	0	4	0	0	0	9	0	0	0	38	0	0	0	0	1	63	0	0	2	0	1	0	0	0	0	0	0	1	0	3	0	2194	95	0
v	54	0	1	470	8	1	0	2	0	0	0	1	0	0	0	1	0	1	0	0	1	1	0	0	2	0	0	0	0	0	0	0	0	0	0
w	58	0	2	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	2	0	0	0	0	0	0	0	0	0	0	0
x	27	0	0	0	2	0	0	0	1	0	0	12	0	0	0	0	0	0	0	0	38	0	0	0	0	0	0	10	0	0	0	0	0	0	0
y	57	0	0	0	0	0	0	0	0	0	21	23	0	0	1	0	0	0	0	0	0	9	7	0	0	0	0	0	0	0	5	0	0	0	0
z	13	0	2	0	141	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	91	0	0	0	0	2	0	0	0	0	0	0	7	0	0
á	0	0	310	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	1	1	0	0	0
é	0	0	0	0	0	0	550	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
í	4	0	0	0	0	0	1	1	0	0	853	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	3	2	0	0
ó	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	424	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0
ú	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	75	0	0	0	0	0	0	0	0	0	0	0	0
ü	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
ñ	54	0	0	0	0	0	0	0	0	0	0	0	0	4	1	7	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

Table 6. Substitution[X(error), Y(correct)]