# On Correction of Semantic Errors in Natural Language Texts with a Dictionary of Literal Paronyms[*]

Alexander Gelbukh,[1,2] Igor A. Bolshakov[1]

[1] Center for Computing Research
National Polytechnic Institute, Mexico City, Mexico
{gelbukh,igor}@cic.ipn.mx; www.Gelbukh.com

[2] Department of Computer Science and Engineering, Chung-Ang University,
221 Huksuk-Dong, DongJak-Ku, Seoul, 156-756, Korea

**Abstract.** Due to the open nature of the Web, search engines must include means of meaningful processing of incorrect texts, including automatic error detection and correction. One of wide-spread types of errors in Internet texts are malapropisms, i.e., semantic errors replacing a word by another existing word similar in letter composition and/or sound but semantically incompatible with the context. Methods for detection and correction of malapropisms have been proposed recently. Any such method relies on a generator of correction candidates—paronyms, i.e., real words similar to the suspicious one encountered in the text and having the same grammatical properties. Literal paronyms are words at the distant of few editing operations from a given word. We argue that a dictionary of literal paronyms should be compiled beforehand and that its units should be grammeme names. For Spanish, such grammemes are (1) singulars and plurals of nouns; (2) adjectives plus participles; (3) verbs in infinitive; (4) gerunds plus adverbs; (5) personal verb forms. Basing on existing Spanish electronic dictionaries, we have compiled a dictionary of one-letter-distant literal paronyms. The size of the dictionary is few tens thousand entries, an entry averaging approximately three paronyms. We calculate the gain in number of candidate search operations achievable through the proposed dictionary and give illustrative examples of correcting one-letter malapropisms using our dictionary.

## 1 Introduction

Linguistic studies based on various Internet search engines repeatedly show that completely correct texts in any language are rather rare exceptions among lots of texts containing errors of various kinds. For any task of creation of specialized corpora or

[*] Work done under partial support of Mexican Government (CONACyT, SNI, COFAA-IPN) and Korean Government (KIPA Professorship for Visiting Faculty Positions in Korea). The first author is currently on Sabbatical leave at Chung-Ang University.

databases extracted from Internet search engines or robots, a module that tries to automatically correct the extracted texts seems absolutely necessary.

There are different types of errors introduced into texts by their authors. Modern widespread authoring tools, such as Microsoft Word or Microsoft FrontPage, easily detect *orthographical* errors—misspellings leading to a letter string that does not exist in the given language as a correct word, e.g., **wors* for *word*. However, such tools usually fail to detect so-called real-word errors: mutilations of a word that lead to another existing word, e.g., $^{?}$*work*, $^{?}$*wore*, or $^{?}$*worn* for *word*. What is more, such tools sometimes introduce such errors by automatically "correcting" misspelled words by substituting them with existing words without asking the user whether this was the intended word.

The presence of some of such real-word errors—namely, *syntactic* errors—is relatively easy to detect using syntactic analysis. For example, the phrase **every wore* (or *worn*) *in this text is spelled correctly* is ungrammatical and thus probably contains a misspelled word. However, when the grammatical characteristics of the new word coincide with those of the intended one, even a grammar checker fails to detect a problem, as, for example, in the phrase $^{?}$*every work in this text is spelled correctly* (for *every word...*), which is perfectly grammatical, while does not make sense, or at least is suspicious, semantically.

Such *semantic* errors cannot be automatically detected or corrected so far. They usually replace a real word by another one existing in the language but semantically incompatible with context, which results in violation of human knowledge and/or common sense and impedes comprehension of the text. Indeed, if the replacing word has the same syntactic function in the sentence, neither orthography nor syntax is violated, so that spelling and grammar checkers fail to detect the problem. Semantic errors replacing a word by another one similar to the intended word in letter composition and/or sound are named *malapropisms*. A Spanish example is *visitar el centro histérico* 'to visit hysterical center' for the intended word *histórico* 'historical'.

Two methods of malapropism correction have been proposed recently [4, 3]. Both of them rely on a generator of correction candidates for malapropisms, i.e., a set of real words similar to the suspicious one encountered in the text, with the same grammatical properties. Such similar words are usually called *paronyms*. Two words that are literal paronyms are at the distant of few editing operations from one another, where by operations insertion, omission, replacement of a letter, or permutation of two adjacent letters is meant. If only one such operation is applied to a string, the resulting letter string is said to be at the distance 1 from the source string. In this paper we study only such 1-distance literal paronyms.

We argue for that a dictionary of literal paronyms should be compiled beforehand for each language. The preferable unit for an entry of such a dictionary is a grammeme. For Spanish, such grammemes are (1) singulars and plurals of nouns;

- adjectives plus participles;
- verbs in infinitive;
- gerunds plus adverbs;
- personal verb forms.

Based on available lexicographic material on Spanish, we have elaborated a dictionary of one-letter-distant literal paronyms. The size of the dictionary proved to be ca. 38,000 words, with an entry (set of literal paronyms of a headword) averaging approximately three. The gain in the number of search operations necessary to find all correction candidates using the proposed dictionary is about 248 times; see calculations below.

More specifically, the objectives of this paper are:

- To discuss distances between words as letter strings;
- To choose the most suitable unit for paronym dictionaries;
- To present the compiled paronym dictionary and to theoretically evaluate the gain achievable using this dictionary as compared with the blind search of correction candidates;
- To outline shortly two possible resources as a basis of practical use of the compiled dictionary;
- To trace three demonstrative examples of correcting one-letter malapropisms.

We illustrate our considerations mostly on English examples, in all cases when it does not contradict our main motivation of applying our method to processing Spanish texts.

## 2 Distance between Words as Letter Strings

For any known method of malapropism detection, a generator of correction candidates is necessary. The candidates should be in some sense “similar” to the given word. Such generation is analogous to generation of the candidates for orthographical errors but differs in the way the similarity is measured and in the restrictions that apply.

Indeed, word forms of a natural language are rare interspersions in the space of all letter strings. For approximate evaluation of this rarefaction, consider that in such highly inflexional language as Spanish there exist ca. 800,000 different word forms, whereas in low inflexional English, say, three times less. The number of all possible strings over a given alphabet consisting of $A$ letters with the lengths equal to the mean length $L$ of a real word in a corresponding dictionary is $A^L$, i.e. $33^9 \approx 4.6 \times 10^{13}$ in Spanish and $26^8 \approx 2.1 \times 10^{11}$ in English. This means that a real word form occurs on average once among 58 million senseless strings in Spanish and 788 thousand in English. The change of the mean length of word form in a dictionary to the mean textual value decreases this contrast, still leaving it huge.

If word forms as letter strings were absolutely stochastic in structure, the probability to meet two forms at a short distance would be inconsiderable. In fact, however, words are built of few thousands of radices and even fewer prefix and suffix morphs (these are few hundreds in languages like Spanish). Some semantic and morphonological restrictions are imposed on compatibility of radixes, prefixes, and suffixes, since not all combinations are reasonable and not all reasonable ones are utterable.

Just this circumstance facilitates the candidate search for replacement of one real word by another. For *orthographical* errors, a wrong string can be arbitrary and the task to gather beforehand, for each such string, all similar real words seems impractical. However, the environments of a real word form, as our considerations show, contain few real words, which can be found beforehand. Being collected in a special dictionary, they could be used for malapropism correction, cutting down the search space of candidates.

Indeed, for correction of one-letter error in a string with the length $L$, it is necessary $A\ (2\,L + 1) + L - 1$ checks, which for a word of nine letters equals to 616. For two-letter errors, already ca. 360,000 checks are necessary. In the same time, beforehand gathered one-letter-apart candidates are only few units and for two-letter-apart ones, few tens. For words that are not in the dictionary of substitutes, the candidate search is not needed, which also cuts down the search.

Let us consider the distance between literal strings in more detail. One literal string of the length $L$ can be formed from any other one by a series of editing operations [5, 6, 9]. Consider strings over an alphabet of $A$ letters. The elementary editing operations are: replacement of a letter with another letter in any place within the source string (here there are $(A - 1)\,L$ options); omission of a letter ($L$ options); insertion of a letter (here, $A\ (L + 1)$ options); permutation of two adjacent letters ($L - 1$ options).

The string obtained with any of these $A\ (2L + 1) + L - 1$ operations is at the distance 1 from the source string, i.e., within the sphere of radius 1 in the string space. Making another elementary step off, we form a string on the sphere with radius two with regard to the source one, etc. Points obtained with minimum $R$ steps are on $R$-sphere, whereas the points of $r$-spheres with $r < R$ and the source point itself are not included in the $R$-sphere. Here are English examples:

- *word* Vs. *world, ethology* Vs. *etiology, ethology* Vs. *ethnology* are at the distance 1,
- *hysterical* Vs. *historical, dielectric* Vs. *dialectic, excess* Vs. *access, garniture* Vs. *furniture* are at the distance 2,
- *company* Vs. *campaign*, *massy* Vs. *massive, sensible* Vs. *sensitive, hypotenuse* Vs. *hypothesis* are at the distance 3 or more.

Though the mean distance between word forms is large in any language, they prove to gather together in clusters. Firstly, such clusters contain elements of morphological paradigms of various lexemes, word forms within them being usually at a distance of 0 to 3 from each other. Just such a cluster is a lexeme, and one of the composing forms is used as its dictionary name. Secondly, paradigms of various lexemes with similar morphs can be close to each other, sometimes even with intersection (such intersections give rise to morphological ambiguities).

## 3 Preferable Unit of Paronym Dictionary

For our purposes, of interest are the paradigm pairs with the same number of elements and correlative elements at the same distance. E.g., all four elements of paradigms of Eng. verbs *bake* and *cake* differ in the first letter only. Let us call such paradigms parallel. If the distance equals to 1, let us call them close parallel.

Thus, any element $\lambda(\chi)$ of the paradigm of $\lambda$, where $\chi$ is a set of intra-lexeme coordinates (i.e., morphological characteristics selecting a specific word form), can be obtained from the correlated element of the parallel paradigm using the same editing operator $R_i(\,)$, where $i$ is the cardinal number of the operator in an effective enumeration of such operators. Then the relation between dictionary names (they correspond to $\chi = \chi_0$) and specific word forms of parallel lexemes can be represented by the proportion

$$\frac{\lambda(\chi)}{\lambda(\chi_0)} = \frac{R_i(\lambda(\chi))}{R_i(\lambda(\chi_0))} \tag{1}$$

This formula means that for any suspicious form $\lambda(\chi)$ in text, it is necessary to find its dictionary form $\lambda(\chi_0)$, and, if a close parallel $R_i(\lambda(\chi_0))$ for it exists, $R_i(\lambda(\chi))$ should be tried as a correction candidate. For such tries, the syntactic correctness usually pertains, and some try can correct the error.

This parallelism permits to unite sets of word forms, storing in the dictionary only one their representative, i.e., the dictionary name of each lexeme. However, strictly parallel paradigms are not so frequent in highly inflectional languages. More usually the parallelism between subparadigms can be found. As such subparadigms, it is reasonable to take grammemes corresponding to fixed combinations of the characteristics $\chi$.

For example, noun lexemes of European languages have grammemes of singular and plural. They play the same role in a sentence but differ in the sets of collocations they can be in.[1] This division of wordforms by grammatical number serves our purposes as well.

Spanish verbs have grammemes of personal forms, infinitive, participles, and gerund. These grammemes differ in their role in a sentence, so that their separate use keeps syntactic correctness of text after restoration of the error.

A dictionary name is assigned to each grammeme, e.g., adjectives and participles are represented in Spanish by their singular forms of masculine. For such dictionary entry and specific word forms, the formula (1) pertains.

In the cases when the parallelism is not valid for some forms of two grammemes under comparison, the correction attempt will fail. If such cases are rare, it is not very problematic.

## 4 Compiled Dictionary and Gain Achieved

No general-purpose electronic dictionary proved to be useful for our goal to compile the dictionary of literal paronyms. So we used the lexicographic materials put at our disposal by our colleagues from the Polytechnic University of Barcelona, Spain.

The following grammemes were taken for the target Spanish dictionary:

- Nouns of masculine gender in singular number, together with the verbal infinitives,

[1] By collocations we mean semantically meaningful and syntactically connected (in the sense of dependency grammars; possibly through a chain of functional words) pairs of words. Some authors mean by collocations co-occurrences within a text unit [10]. We believe that even in this case collocation properties of plurals and singulars differ.

- Nouns of masculine gender in plural number,
- Nouns of feminine gender in singular number,
- Nouns of feminine gender in plural number,
- Adjectives together with verbal participles (singular masculine form representing all forms),
- Adverbs together with verbal gerunds,
- Verbs in infinitive,
- Verbs in 3rd person present indicative (representing all personal forms of non-compound tenses)

Here are some examples of the entries of the dictionary:

- ...
- **aliño** *nms*
  - alijo *nms*
  - aliso *nms*
- **ali** *nms*
  - ami *nms*
- **aliado** *nms*
  - alzado *nms*
- **aliar** *v*
  - alias *nmn*
  - altar *nms*
  - alzar *v*
  - alear *v*
  - alfar *v*
  - aviar *v*
  - adiar *v*
- **alias** *nmn*
  - aliar *v*
- **alicorear** *v*
  - alicortar *v*
- ...
- **bajas** *nfp*
  - balas *nfp*
  - batas *nfp*
  - bayas *nfp*
  - bazas *nfp*
  - babas *nfp*
  - cajas *nfp*
  - fajas *nfp*
  - lajas *nfp*
  - majas *nfp*
  - pajas *nfp*
  - rajas *nfp*
- **bajezas** *nfp*
  - majezas *nfp*
- **bajistas** *ncp*
  - bañistas *ncp*
  - cajistas *ncp*
- ...
- **apical** *adj*
  - amical *adj*
- **apilonado** *v*
  - apisonado *v*
  - apitonado *v*
  - apiñonado *adj*
- **apiojado** *v*
  - apiolado *v*
- ...
- **cuando** *adv*
  - ciando *v*
  - cuanto *adv*
  - puando *v*
  - ruando *v*
- **cuanto** *adv*
  - cuando *adv*
- **cuartando** *v*
  - coartando *v*
- **cubando** *v*
  - cebando *v*
  - cucando *v*
  - cunando *v*
  - curando *v*
  - cuñando *v*
- ...
- **alude** *v*
  - acude *v*
  - elude *v*
  - ilude *v*
- **aluja** *v*
  - aleja *v*
  - alija *v*
  - aloja *v*
  - aluna *v*
  - aluza *v*
  - amuja *v*
  - atuja *v*
  - aduja *v*
  - aguja *v*
- **alumbra** *v*
  - adumbra *v*
  - alambra *v*
- ...

The meaning of the marks next to the words is as follows: *nms* stands for noun, masculine, singular; *nmn* for noun, masculine, neutral number, *nfp* for noun, feminine, plural; *ncp* for noun, common gender, plural; *adj* for adjective, *v* for verb, *adv* for adverb. In the examples one can observe mixed groups, in which substitutes for a word of a certain morphological part of speech is a word of another morphological part of speech, though the same syntactic function: for example, (*el*) *aliar* / (*el*) *alias*.

The statistics for each type of grammemes and totaling parameters are in Table 1.

Table 1. Statistics of the dictionary on grammeme basis

| Syntactic part of speech | Source entries | Entries with paronyms | Average group |
|---|---|---|---|
| Nouns masculine singular, with infinitives | 24,805 | 8,598 | 2.73 |
| Nouns masculine plural | 9,568 | 2,034 | 2.38 |
| Nouns feminine singular | 11,656 | 2,862 | 2.89 |
| Nouns feminine plural | 7,901 | 1,864 | 2.66 |
| Adjectives with participles | 23,255 | 6,349 | 2.14 |
| Adverbs with gerunds | 13,592 | 5,775 | 3.03 |
| Verbs in infinitive | 11,831 | 5,385 | 2.99 |
| Verbs in 3rd person present indicative | 11,785 | 5,512 | 2.96 |
| **Total** | **114,393** | **38,379** | **2.74** |

One can see that the entries with paronyms cover ca. 33.6% of the source dictionary and amount ca. 38 thousand of the entries of the compiled dictionary. On average, the groups of paronyms contain 2.74 elements, this figure weakly depending on a specific grammeme.

The main gain in candidate search is reached due to looking up only the candidates given in the paronymy dictionary. Using the total number of tries for a 9-letter Spanish word, we get the gain coefficient $G_1 = 616 / 2.74 = 225$.

The source dictionary contains ca. 114,000 grammemes and the revealed paronymous grammemes are supposedly the most frequent among them. With the reasonable assumption that the rank distribution of all words in the dictionary conforms to Zipf law, we have the additional gain coefficient $G_2 = \ln 114{,}300 / \ln 38{,}400 = 1.103$ due to that all other 75,900 grammemes are ignored in the candidate search. The total gain is $G_1 \times G_2 \approx 248$.

## 5 Resources for Testing Collocations

The developed dictionary of paronyms can be used only in conjunction with a resource for testing whether or not a given pair of content words can constitute a collocation. In [3] two such resources have been proposed, with different strategy of the decision.

The resource of the first type is a machine-readable collocation base. For English, the market can only propose Oxford collocations dictionary [8] in printed form. However, the availability of the large collocation base for Russian [2] shows that such bases will be at hand for other languages in the foreseeable future.

Two words *V* and *W* are admitted to form a collocation recorded in this collocation base if both words are in its dictionary and potential syntactical link between them corresponds to the features also recorded in the base.[2]

[2] Hence a simplistic syntactic analysis of the text under revision is necessary.

The resource of the second type can be a search engine for Internet, e.g., Google [3]. For using it, a simpler criterion of statistical nature is applied. The words *V* and *W* are considered combinable into a collocation, if the mutual information inequality is satisfied [7; cf. 10]:

$$\ln\frac{N(V,W)}{N_{\max}} > \ln\frac{N(V)}{N_{\max}} + \ln\frac{N(W)}{N_{\max}}, \qquad (2)$$

where $N\ (V,W)$ is the number of web pages where *V* and *W* co-occur, $N\ (V)$ and $N\ (W)$ are the numbers of the web pages where each words occurs evaluated separately, and $N_{\max}$ is the total number of pages collected by Google for the given language. The latter value can be approximately calculated through the numbers of pages where the most used (functional) words of a given language occur [1]. It was shown [1] that the number of Spanish-language pages in Google was ca. 12.4 million at the end of 2002.

The inequality (2) rejects those potential collocations whose components co-occur in a statistically insignificant number.

## 6 Illustrative Examples

Let us track the occurrences of malapropisms in the following three intended Spanish sentences:

- *Seguimos la cadena causal entera* 'we follow the whole causal chain'.
- *La gente espera una tormenta de granizo* 'people expect a tempest of hail'.
- *Visitaremos el centro histórico* 'we will visit the historical center'.

The intended words are now changed to malapropos: *causal* to *casual, granizo* to *granito*, and *histórico* to *histérico*, given the following erroneous variants:

- *Seguimos la cadena casual entera* 'we follow the whole casual chain'.
- *La gente espera una tormenta de granito* 'people expect a tempest of granite'.
- *Visitaremos el centro histérico* 'we will visit the hysterical center'.

Let is trace the two different recourses for collocation testing.

**Collocation base.** In the base, the following collocations are supposed: *seguir la cadena, cadena causal, cadena entera, gente espera, esperar una tormenta, tormenta de granizo, visitar el centro, centro histórico.* At the same time, the malapropos combinations *cadena casual* 'casual chain', *tormenta de granito* 'tempest of granite', *tormenta de grafito* 'tempest of graphite', and *centro histérico* '*hysterical center'* are not present in the base. Thus the malapropos words *casual, granito,* and *histérico* will be detected immediately. Accessing to dictionary of literal paronyms, we will find the following candidates: *causal* for *casual*; *grafito* and *granizo* for *granito*; *histórico* for *histérico.* The unique candidates among them correspond to collocations recorded in the base, thus indicating the true way for correction. Among the two options to correct the second sentence, only *granizo* restores the collocation, so only this word will be proposed to the user for correction.

**Google.** If only access to Google (but not to a collocation database) is available, malapropism detection and correction is possible on the statistical threshold rule (1). We gathered the necessary statistics of occurrences and co-occurrences in Google (see Table 2).

Table 2. Statistics for detection and correction

| Intended combination | Statistics | | | | |
|---|---|---|---|---|---|
| | Intend. combin. | Malaprop. combin. | 1st word (invariable) | 2nd word (intended) | 2nd word (malapropos) |
| *cadena causal* | 1,080 | 21 | 786,000 | 80,800 | 140,000 |
| *tormenta de granizo* | 802 | 0/0 | 191,000 | 30,900 | 98,100 |
| *centro histórico* | 110,000 | 75 | 403,000 | 883,000 | 10,600 |

One can see that malapropos word combinations occurred in insignificant numbers rejected by the statistical criterion after taking into account that all words under consideration are rather frequent in Internet. For the second sentence, both malapropos combination and the alternative candidate for correction (*grafito*) give zero number of co-occurrences among millions of Spanish web-pages.

Hence, the results are incentive to continue the research.

## 7 Conclusion and Future Work

To drastically speed up the search of candidates for malapropism correction, we have proposed a dictionary of literal paronyms. Such paronyms are real words at the distance 1 between them in the letter string space.

Significant limitations on further experimentations with our dictionary are laid by that

- We possess currently only a small collocation base for Spanish (less that 10,000) and its broadening will take some time.
- In statistical threshold formula, we use the number of strict co-occurrences of the collocation-forming words, while Google does not permit us to estimate more realistically the situations when these words are separated by several other words.

It is worthwhile to gather also literal paronyms distanced by 2. For example, the malapropos words in *materialismo* *dieléctrico* ‘dielectric materialism’ and in *orugas en la piel* ‘tracks on the skin’ are at the distance 2 from intended words *dialéctico* ‘dialectic’ and *arrugas* ‘wrinkles’.

As another tool accelerating the candidate search, a dictionary of sound paronyms can be introduced. In Spanish, the sound distance between *k*, *qu* and *c* (before *a*, *e*, *u*) is equal to zero (all of them are pronounced as [k]), whereas in letter space this distance is 1 or 2; an example of English strings with zero distance in the sound is *right*, *Wright*, *write*. Thus, the transition to the phonological space can sometimes simplify the search.

## References

1. Bolshakov, I. A., S. N. Galicia-Haro. *Can We Correctly Estimate the Total Number of Pages in Google for a Specific Language?* In: A. Gelbukh (Ed.) *CICLing-2003*, *Computational Linguistics and Intelligent Text Processing*. Lecture Notes in Computer Science, No. 2588, Springer-Verlag, 2003, p. 415-419.
2. Bolshakov, I. A., A. Gelbukh. *A Very Large Database of Collocations and Semantic Links*. In: M. Bouzeghoub *et al*. (Eds.) *NLDB-2000, Natural Language Processing and Information Systems*. Lecture Notes in Computer Science, No. 1959, Springer-Verag, 2001, p. 103–114.
3. Bolshakov, I. A., A. Gelbukh. *On Detection of Malapropisms by Multistage Collocation Testing*. In: A. Dusterhoft, B. Talheim (Eds.) *8th Intern. Conference on Applications of Natural Language to Information Systems NLDB-2003*, GI-Edition, Lecture Notes in Informatics, v. P-29, Bonn, 2003, p. 28-41.
4. Hirst, G., D. St-Onge. *Lexical Chains as Representation of Context for Detection and Corrections of Malapropisms*. In: C. Fellbaum (ed.) *WordNet: An Electronic Lexical Database*. The MIT Press, 1998, p. 305-332.
5. Kashyap, R. L., B. I. Oomen. *An effective algorithm for string correction using generalized edit distances. I. Description of the algorithm and its optimality*. *Information Science*, 1981, Vol. 23, No. 2, p. 123-142.
6. Mays, E., F. J. Damerau, R. L. Mercer. *Context-based spelling correction. Information Processing and Management*. 1992, Vol. 27, No. 5, p. 517-522.
7. Manning, Ch. D., H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.
8. *Oxford Collocations Dictionary for Students of English*. Oxford University Press. 2003.
9. Wagner, R.A., M. J. Fisher. *The string-to-string correction problem*. *J. ACM*, Vol. 21, No. 1, 1974, p. 168-173.
10. Biemann, C., S. Bordag, G. Heyer, U. Quasthoff, C. Wolff. *Language-independent Methods for Compiling Monolingual Lexical Data*. In A. Gelbukh (Ed.), *CICLing-2004, Computational Linguistics and Intelligent Text Processing*. Lecture Notes in Computer Science, N 2945, Springer, 2004, pp. 214–225.