Detection and Correction of Malapropisms in Spanish by means of Internet Search*

Igor A. Bolshakov,[1] Sofia N. Galicia-Haro,[2] and Alexander Gelbukh[1]

[1] Center for Computing Research (CIC), National Polytechnic Institute (IPN), Mexico
igor@cic.ipn.mx, gelbukh@gelbukh.com; www.Gelbukh.com

[2] Faculty of Sciences, National Autonomous University of Mexico (UNAM), Mexico
sngh@fciencias.unam.mx

Abstract. Malapropisms are real-word errors that lead to syntactically correct but semantically implausible text. We report an experiment on detection and correction of Spanish malapropisms. Malapropos words semantically destroy collocations (syntactically connected word pairs) they are in. Thus we detect possible malapropisms as words that do not form semantically plausible collocations with neighboring words. As correction candidates, we select words similar to the suspected one but forming plausible collocations with neighboring words. To judge semantic plausibility of a collocation, we use Google statistics of occurrences of the word combination and of the two words taken apart. Since collocation components can be separated by other words in a sentence, Google statistics is gathered for the most probable distance between them. The statistics is recalculated to a specially defined Semantic Compatibility Index (SCI). Heuristic rules are proposed to signal malapropisms when SCI values are lower than a predetermined threshold and to retain a few highly SCI-ranked correction candidates. Our experiments gave promising results.

1 Introduction

Malapropism is a type of semantic error that replaces one content word by another existing word similar in sound or letters but semantically incompatible with the context and thus destroying text cohesion, e.g., Spanish *mañana* <u>*sopeada*</u> 'overridden morning' for the intended *mañana* <u>*soleada*</u> 'sunny morning.' Two interconnected tasks arise: (1) detecting erroneous words and (2) suggesting candidates for their correction.

Hirst & St-Onge [5] proposed detecting suspected malapropisms as words not related to any word in the context and selecting correction candidates as words similar to the suspected ones but related to the words in the context; if such a possible correction is found then the suspected word is signaled to the user along with the proposed correction. As a particular measure of relatedness between words, they used the distance in WordNet graph. In particular, this distance is determined through paradigmatic relations (synonyms, hyponyms, hyperonyms), mainly between nouns. The syn-

* Work done under partial support of Mexican Government (CONACyT, SNI, CGEPI-IPN).

tactic links between words are ignored. The matched words are usually in different sentences or even paragraphs.

Bolshakov & Gelbukh [1] observed that malapropos words destroy collocations the original word was in, where by a collocation we mean a combination of two syntactically linked (maybe through an auxiliary word such as a preposition) and semantically compatible content words. The resulting word combinations usually retain their syntactic type but lose their sense, as in the example above. Thus, the general idea in [1] is similar to [5], but the anomaly detection is based on syntactico-semantic links between content words. A much smaller context—only one sentence—is needed for error detection, and words of all four open POSs—nouns, verbs, adjective, and adverbs—are considered as collocation components (collocatives). To test whether a content word pair is a collocation, three types of linguistic resources were considered: a precompiled collocation dictionary, a large corpus, or a Web search engine. However the experiment described in [1] was limited.

This paper extends [1] by intensive experimentation with Google as a resource for testing Spanish collocations (for English, the results would be even much more statistically significant). The Web is widely considered now as a huge (but noisy) linguistic resource [3, 4]. To use it for malapropism detection and correction, we had to revise the earlier algorithm and to develop new threshold-based procedures. Especially important was to investigate collocations of various syntactical types with collocatives either sequentially adjacent (forming bigrams, which are sufficiently explored [3]) or distant from each other (such collocations are insufficiently explored [8]; they are considered in dependency grammar approaches [7]).

The rest of the paper is organized as follows. In Section 2 we give a classification of syntactical types for collocations frequent in Spanish and demonstrate that some collocations can have rather distant collocatives. In Section 3, we explore frequencies of collocative co-occurrences in relation with the distance between them and discuss what part of the co-occurrences are real collocations. In Section 4, we present a method of malapropism detection and correction using a Semantic Compatibility Index (SCI) as a numeric measure for semantic compatibility of collocatives. In Section 5, we give details on our test collection. In Section 6 we describe our experiments. We dispose the collocatives of 125 rather common Spanish collocations at the most probable distances, convert them to malapropisms, and then gather word combinations for potential correction by replacing one component of the suspected malapropos collocation by its paronyms (e.g., similar words [2]).We use Google statistics to obtain SCI values. Finally, in Section 7 we give conclusions and discuss future work.

2 Collocations in Their Adjacent and Disjoint Forms

We consider syntactico-semantic links between collocatives as in dependency grammars [7]. Each sentence can be syntactically represented as a dependency tree with directed links ‘head → its dependent’ between tree nodes labeled by words of the sentence. Going along these links in the direction of the arrows from one content node through any functional nodes down to another content node, we obtain a labeled substructure corresponding to a word combination. If this is a meaningful text, we call

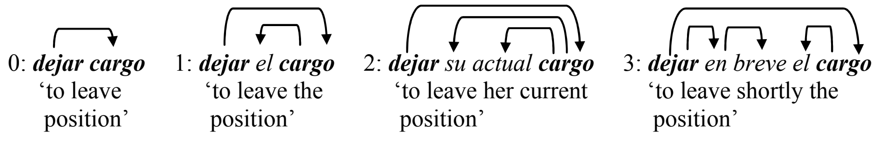
Table 1. Frequent types and structures of Spanish collocations

Type	Code	Depend. subtree	Example	%
Modified → Modifier	1.1	Adj ← N	*vago presentimiento*	2
	1.2	N → Adj	*mañana soleada*	20
	1.3	Adv ← Adj	*moralmente inferior*	1
	1.4	Adj → Adv	*libre forzosamente*	1
	1.5	V→ Adv	*salir botando*	3
	1.6	V → Pr → N	*junta con* (*las*) *manos*	5
	1.7	N → Pr → N	*hijos de familia*	6
	1.8	Adv → Adv	*mirando fijamente*	2
	1.9	Adv → Pr → Adj	*negando en rotundo*	3
Noun → Noun Complement	2.1	N → Pr → N	*mechón de canas*	6
	2.2	N → Pr → V	*goma de mascar*	1
Noun → Noun Attribute	3.1	N → N	*rey mago*	1
Verb → Noun Complement	4.1	V → N	*afilar navajas*	17
	4.2	V → Pr → N	*tener en mente*	9
Verb→ Verbal Complement	5.1	V → Pr → V	*trata de cambiar*	1
	5.2	V → V	*piensa escribir*	1
Verb→ Adjective Complement	6.1	V → Adj	*era liso*	3
	6.2	V → Pr → Adj	*duda de todo*	1
Verb Predicate → Subject	7.1	N ← V	*colección crece*	1
	7.2	V → N	*existe gente*	1
Adjective → Noun Complement	8.1	Adj → Pr → N	*lleno de tierra*	3
Adverb → Noun Complement	9.1	Adv → N	*pateando puertas*	1
	9.2	Adv → Pr → N	*junto con* (*su*) *familia*	1
Coordinated Pair	10.1	N → Cc → N	*ida y vuelta*	4
	10.2	Adj → Cc → Adj	*sano y salvo*	2
	10.3	V → Cc → V	*va y viene*	2
	10.4	Adv→ Cc → Adv	*rápidamente y bien*	1
			Total:	100

such word combination a *collocation*. Such a definition of collocations ignores their frequencies and idiomaticity.

The most frequent types of Spanish collocations (as dependency sub-trees [7]) are given in Table 1. The types and subtypes are determined by POS of collocatives and their order in texts; V stands for verb, N for noun, Adj for adjective or participle, Adv for adverb, Pr for preposition. Usually dependency links reflect subordination between words (1.1 to 9.2). However, there exist coordinate dependency with collocatives of the same POS linked through the coordinating conjunction Cc (10.1 to 10.4).

Though adjacent in the dependency tree, collocatives can be distant in linear word order. The possible distances depend on the collocation type and specific collocatives. E.g., 3.1-collocatives are usually adjacent, whereas the 4.1-collocation *dejar cargo* 'to leave position' can contain intermediate context of 0 to 3 or even more words:



0: ***dejar cargo*** 'to leave position'

1: ***dejar*** *el* ***cargo*** 'to leave the position'

2: ***dejar*** *su actual* ***cargo*** 'to leave her current position'

3: ***dejar*** *en breve el* ***cargo*** 'to leave shortly the position'

3 The Most Probable Distance between Collocatives

A specific (malapropos or not) collocation encountered in a text has a specific distance between collocatives. However, to explore collocations in a general manner, we have to consider each collocative pair in its most probable distance.

Before determining the most probable distances between specific collocatives by means of the Web, it is necessary to clarify correspondences between the Web frequencies of collocative co-occurrences and occurrences of real collocations potentially formed by them. Google statistics of co-occurrences of any two strings with any *N* intermediate words can be gathered by queries in quotation marks containing these strings separated with *N* asterisks, e.g., `"dejar * * cargo"` for $N = 2$. So we intended to compare frequencies of the two kinds of events in the following way.

We took at random ten different commonly used collocations in their arbitrary textual form (maybe inflected) with unknown length of intermediate context. Co-occurrence frequencies for each collocative pair were evaluated with 0 to 5 intermediate asterisks. We cannot determine automatically whether counted co-occurrences are real collocations or merely random coincidences of words, possibly from different sentences. To evaluate the true portion (TP) of collocations in the automatically counted amounts, we looked through the first hundred snippets with co-occurrences for various lengths of intermediate context and manually analyzing their syntax. Multiplying the Google statistics (GS) by TP values, we obtained approximate collocation statistics (CS), see Table 2.

One can see that within 0 to 5 intermediate words, GS has one or more local maxima, whereas the first local maximum of CS is at 0 to 2 and in all cases is unique, coinciding with the first local maximum of GS. So we can believe Google statistics in that the most probable distance between collocatives of real collocations corresponds to the first local maximum of GS. The majority of collocative co-occurrences counted by the Web at the distances 0 to 3 are real collocations, whereas at greater distances they are mostly coincidences of words without direct syntactic relation. This does not mean that collocations cannot be more distant, but the Web is not suited for collocation testing at greater distances, in contrast with collocation databases.

Table 2. Statistics of co-occurrences and collocations

Collocation	Statistics	Number of intermediate words					
		0	1	2	3	4	5
dejar ... cargo 'leave ... position'	GS	301	**17600**	735	973	846	737
	TP	0.93	**0.99**	0.63	0.54	0.11	0.09
	CS	280	**17424**	463	525	93	66
tentar ... suerte 'have ... luck'	GS	516	1230	**2160**	239	70	33
	TP	**1.00**	**1.00**	0.98	0.85	0.83	0.49
	CS	516	1230	**2117**	203	58	16
tener ... mente 'have ... mind'	GS	**22400**	1220	93	97	68	120
	TP	**1.00**	0.99	0.06	0.05	0.10	0.07
	CS	**22400**	1208	6	5	61	8
cortarse ... venas 'cut ... sales'	GS	3	**5500**	47	7	321	9
	TP	**1.00**	**1.00**	0.79	0.86	1.00	1.00
	CS	3	**5500**	37	6	321	9
fijar ... fecha 'define ... date'	GS	4720	**13400**	1320	693	1350	2320
	TP	**0.89**	0.77	0.38	0.28	0.23	0.24
	CS	4201	**10318**	502	194	311	557

4 Algorithm for Malapropism Detection and Correction

The main idea of our algorithm is to look through all pairs of content words W_i within a sentence, testing each pair on syntactic and semantic compatibility. If the pair is syntactically combinable but semantically incompatible, a malapropism is suspected. Then all primary candidates for correction are tested on semantic compatibility with the context. The list of secondary candidates is ranked and only the best ones are kept:

```
for all content words $W_i$ and $W_j$ in sentence such that $j < i$ repeat
  if SyntCombinable($W_j$,$W_i$) & not SemCompatible($W_j$,$W_i$) then
    ListOfPairs = $\varnothing$
    for each paronymy dictionary
      for all paronyms $P$ of the left collocative $W_j$ repeat
        if SemAdmissible($P$,$W_i$) then InsertToListOfPairs($P$,$W_i$)
      for all paronyms $P$ of the right collocative $W_i$ repeat
        if SemAdmissible($W_j$,$P$) then InsertToListOfPairs($W_j$,$P$)
Filter(ListOfPairs), LetUserTest(ListOfPairs)
```

Here, Boolean function **SyntCombinable**(V,W) determines if the word pair (V,W) forms a syntactically correct word combination. It implements a partial dependency parser searching for a conceivable dependency chain with V and W at the extremes that includes their intermediate context, see Table 1. Boolean functions **SemCompatible**(V,W) and **SemAdmissible**(V,W) both check if the pair (V,W) is semantically compatible. The procedure **Filter**(*ListOfPairs*) selects the best candidates. These three procedures heavily depend on the available resources for collocation testing.

When the resource is a text corpus, **SemCompatible**(V,W) determines the number $N(V,W)$ of co-occurrences of V and W within a limited distance one from another in the whole corpus. If $N(V,W) = 0$, it returns *False*. Otherwise, for a definite decision it is necessary to syntactically analyze each co-occurrence, which is considered impractical in a large corpus. In the case of ambiguity of whether the co-occurrences are real collocations or mere coincidences in a text span, only statistical criteria are applicable. According to one criterion, the pair is compatible if the relative frequency $N(V,W) / S$ (empirical probability) of the co-occurrence is greater than the product of relative frequencies $N(V) / S$ and $N(W) / S$ of V and W taken apart (S is the corpus size). Using logarithms, we have the following rule for compatibility of a pair:

$$\mathrm{MII}(V,\ W) \equiv \ \ln(N(V,\ W)) + \ln(S) - \ln(N(V) \times N(W)) > 0,$$

where MII(V, W) is the mutual information index [6].

In the Web searchers, only a statistical approach is possible. Search engines automatically deliver statistics about the queried words and word combinations measured in numbers of pages. We can re-conceptualize MII with all Ns as numbers of relevant pages and S as the page total managed by the searcher. However, now N / S are not the empirical probabilities (but presumably values monotonically connected with them).

To heuristically estimate the collocative pair compatibility, we propose a Semantic Compatibility Index (SCI) value similar to MII:

$$\mathrm{SCI}(V,W) \equiv \begin{cases} \ln N(V,W) - \tfrac{1}{2}\,(\ln N(V) + \ln N(W)) + \ln P, & \text{if } N(V,W) > 0, \\ NEG, & \text{if } N(V,W) = 0, \end{cases}$$

where *NEG* is a negative constant symbolizing $-\infty$; *P* is a positive constant to be chosen experimentally. An advantage of SCI as compared to MII is that the total number of pages is not needed to be known. Because of the factor ½, SCI does not depend on monotonic or oscillating variations of the statistics of the search engine, just as MII.

SemCompatible returns *False* and thus signals the pair (V_m, W_m) as a malapropism if $\text{SCI}(V_m, W_m) < 0$, whereas **SemAdmissible** returns *True* and admits the primary candidate (V,W) as a secondary one if the SCI values for the candidate and the malapropism conform to the following threshold rule:

$$\text{SCI}(V,W) > \text{SCI}(V_m, W_m) > NEG \text{ or } \text{SCI}(V_m, W_m) = NEG \text{ and } \text{SCI}(V,W) > Q,$$

where Q, $NEG < Q < 0$, is a constant to be chosen experimentally.

Filter procedure operates on a whole group of secondary candidates, ranking them by SCI values. The chosen candidates are all *n* those with positive SCI; if $n = 1$ then one more with a negative SCI value is admitted, or two more if $n = 0$.

5 An Experimental Set of Malapropisms

When a malapropism is detected in text, it is not initially known which collocative is erroneous. We try to correct both, but only one combination corresponds to the intended collocation; we call it *true correction*. Sometimes an error transforms one collocation to another semantically plausible collocation, which happens rarer and contradicts the extra-collocational context, e.g., *nueva ola* 'new wave' changed to *nueva ala* 'new wing.' We call such errors quasi-malapropisms. Their detection (if possible) usually permits to restore the intended word.

We have collected our experimental set in the following way. We took at random 125 valid collocations, most of them commonly used. Collocatives in each collocation were then separated to their most probable distance in the way described in Section 3. The number of intermediate asterisks in the search pattern was determined for the whole group. Then one collocative in each pair was changed to another real word of the same POS through an elementary editing operation, thus forming a malapropism. To simulate the detection and correction phase, other editing operations were then applied to the both components of the resulting test malapropism, each change giving a correction candidate. Each resulting word combination was included in the set.

The set (Fig. 2) thus consists of groups with headlines containing malapropisms with their collocation subtype codes (cf. Table 1). The changed word is underlined. The true correction is marked in bold. In total, the set includes 977 correction candidates, i.e. 7.82 primary candidates per error. The number of quasi-malapropisms is 8.

6 An Experiment with Google and its Results

The initial groups of the experimental set supplied with statistics are given in Fig. 2. As many as 71 (56.8%) malapropisms and 662 (67.7%) primary candidates were not met in Google. However we keep hope that further elaboration of the statistics and threshold procedures could give much better results.

Collocation statistics		Word statistics	
mañana sopeada (1.2)	0	*mañana*	3180000
		sopeada	99
macana sopeada	0	*macana*	173000
mañana soleada	3710	*soleada*	48100
mañana topeada	0	*topeada*	117
mañana copeada	0	*copeada*	15
mañana hopeada	0	*hopeada*	4
mañana jopeada	0	*jopeada*	26
ora liso (6.1)	7	*ora*	13100000
		liso	382000
ara liso	0	*ara*	4160000
era liso	922	*era*	37700000
osa liso	0	*osa*	3810000
ova liso	0	*ova*	1880000
ora luso	1	*luso*	579000
ora laso	1	*laso*	144000
ora leso	2	*leso*	247000

Collocation statistics		Word statistics	
rey vago (3.1)	10	*rey*	6330000
		vago	652000
bey vago	0	*bey*	1670000
ley vago	1	*ley*	11800000
reg vago	0	*reg*	17600000
reo vago	7	*reo*	1360000
rey lago	198	*lago*	4280000
rey mago	8320	*mago*	705000
rey pago	88	*pago*	5160000
rey vaho	0	*vaho*	28800
rey vaso	4	*vaso*	882000
rey vado	2	*vado*	659000

Fig. 2. Several malapropisms and their primary candidates with Google statistics.

To obtain all negative SCI values for all true malapropisms, we took $P = 3500$. The value $NEG \approx -9$ is taken lower than SCI values for all events met. The value $Q = -7.5$ is adjusted so that all candidates with non-zero occurrences have SCI values greater then this threshold. The distribution of SCI values rounded to the nearest integers for malapropisms and their true corrections is shown in Fig. 3. The peak for malapropisms is reached at –4, while for their true corrections it is between 2 and 3.

Though none of the eight quasi-malapropisms was taken into account while selecting the constant *P*, our algorithm detected all of them: their SCI values are too low to be admitted as collocations by our algorithm. That is, the algorithm detected all unintended real word errors (in our experimental set).

SemAdmissible function leaves 207 secondary candidates of 977 primary ones (decrease by 4.72), while **Filter** procedure reduces them to 175 best candidates (total decrease is 5.58). Thus the lists of the best candidates contain on an average 1.4 entries, cf. several groups with SCI values and decision qualifications in Fig. 4.

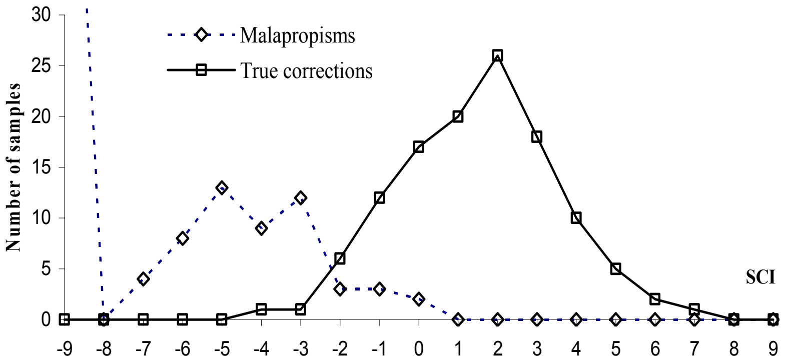


Fig. 3. Distribution of SCI for malapropisms and their true corrections.

Malapropism	Type	SCI	Strength	Candidates	SCI
mañana sopeada	1.2	–9.00	true	***mañana soleada***	1.98
ora liso	6.1	–4.27	true	***era liso***	–0.32
rey vago	3.1	–4.06	true	***rey mago***	2.62
				rey lago	–2.02
pelado venial	1.2	–9.00	true	***pecado venial***	2.67
				pelado genial	–3.06
hombre mosco	1.2	–5.31	true	***hombre hosco***	0.71
				hombre tosco	–0.19
comida sola	1.2	–3.18	quasi	***comida sosa***	–1.34

Fig. 4. Several malapropisms and best candidates with their SCI values.

Among the best candidates always were true corrections, and only four of them were not first-ranked. The most commonly used collocations among primary candidates always enter into the selected list, as true corrections or not.

Hence the results of our experiment are very promising. SCI proved to be an excellent measure for detecting malapropisms and selecting the best correction candidates.

7 Conclusions

A method for detection and correction of malapropisms is proposed. It is based on Google occurrence statistics recalculated as a novel numeric Semantic Compatibility Index for syntactically linked words (collocatives). The experiment was conducted on a test set of 117 malapropisms and 8 so-called quasi-malapropisms (collocations existing in language but used erroneously in a given context). All 125 errors were detected and for all of them their intended correction candidates entered highly ranked into the lists of best correction candidates selected by the algorithm.

References

1. Bolshakov, I.A., A. Gelbukh. On Detection of Malapropisms by Multistage Collocation Testing. In: NLDB´2003, GI-Edition**,** LNI, V. P-29, Bonn, 2003, p. 28-41.
2. Bolshakov, I. A., A. Gelbukh. Paronyms for Accelerated Correction of Semantic Errors. *International Journal on Information Theories and Applications*, Vol.10, 2003, pp. 11–19.
3. Keller, F., M. Lapata. Using the Web to Obtain Frequencies for Unseen Bigram. *Computational linguistics*, V. 29, No. 3, 2003, p. 459-484.
4. Kilgarriff, A., G. Grefenstette. Introduction to the Special Issue on the Web as Corpus. *Computational linguistics*, V. 29, No. 3, 2003, p. 333-347.
5. 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*. MIT Press, 1998, p. 305-332.
6. Manning, Ch., H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.
7. Mel'čuk, I. *Dependency Syntax: Theory and Practice*. SONY Press, NY, 1988.
8. Wermter, J., U. Hahn. Collocation Extraction Based on Modifiability Statistics. Proc. COLING'2004, Geneva, Switzerland, August 2004, p. 980–986.