# On Detection of Malapropisms by Multistage Collocation Testing*

Igor A. Bolshakov and Alexander Gelbukh

Center for Computing Research (CIC),
National Polytechnic Institute (IPN)
Mexico City, Mexico
{igor,gelbukh}@cic.ipn.mx,
www.gelbukh.com

**Abstract:** Malapropism is a (real-word) error in a text consisting in unintended replacement of one content word by another existing content word similar in sound but semantically incompatible with the context and thus destructing text cohesion, e.g.: *they travel around the word*. We present an algorithm of malapropism detection and correction based on evaluating the cohesion. As a measure of semantic compatibility of words we consider their ability to form syntactically linked and semantically admissible word combinations (collocations), e.g: *travel* (*around the*) *world*. With this, text cohesion at a content word is measured as the number of collocations it forms with the words in its immediate context. We detect malapropisms as words forming no collocations in the context. To test whether two words can form a collocation, we consider two types of resources: a collocation DB and an Internet search engine, e.g., Google. We illustrate the proposed method by classifying, tracing, and evaluating several English malapropisms.

## 1 Introduction

One of the most important applications of computers is the interactive (manual) editing of texts. This obligatorily includes automatic error detection and automated error correction: the computer should be able to detect erroneous or suspiciously looking words and phrases and propose their possible corrections. The author of the text can accept one of these corrections or make another correction.

For example, when one types a phrase like **they travel around the workd* in Microsoft Word 2000, the program underlines the erroneous word and suggests the user to change it to one of the following words, in this order: *worked*, *work*, *world*, *word*, or *works*. In this case, the error resulted in a chain that does not exist in English. The problem of out-of-context correction of such errors—separate words converted while typing into a letter chain not existing in the language (i.e., non-word) due to an orthographic error or a typing slip—is practically solved. Spell-checkers detect such errors and suggests those

* Work done under partial support of Mexican Government (CONACyT, SNI), CGEPI-IPN, Mexico, and RITOS-2. We thank Prof. G. Hirst for useful discussion and criticism.

correction candidates existing in the language that are similar to the textual non-words. However, in other cases, e.g., **they travel around the word*, current spell-checkers fail to detect (and thus to correct) the error. In this paper, we address such a problem.

One of the types of such errors are syntactic errors, which leave correct all separate words but violate the sentence structure by incorrect word agreement in grammatical number, gender, case or person, or by replacement of one part of speech to another, or by violation of habitual word order, or by the like. Eventually, the parser cannot analyze a sentence of all words correct. Grammar checkers detect such errors, but they are rather deficient because of limited capabilities of underlying parsers.

In raw texts, semantic errors also occur. They are of various types and usually violate neither orthography nor grammar. Being expressed by correct words inappropriate in a given context or by grammatically correct phrases contradicting to common sense, these errors break text understanding. A particular type of semantic errors is malapropism.

Encyclopædia Britannica [10] defines malapropism as a verbal blunder in which one word is replaced by another similar in sound but different in meaning, e.g., *travel around the* ***word*** (stands for *world*). The similarity in sound singles out malapropisms from the global class of real-word errors consisting of replacement of one word existing in the language by another existing one.

It is impossible nowadays automatically detect real-word errors by the total morphological, syntactic, semantic, and pragmatic analysis of text. Within the limitations of the state of the art, we are aware of rather few papers on the problem of malapropism detection and correction [11], [14], [16].

Hirst and Budanitsky [12], [13] have presented a model of the detection and correction of malapropisms that relies on semantic anomaly or, equivalently, perturbations of coherence in a text. For them, semantic anomaly is indicated by words that are distant in WordNet from all others in the context; and corrections are spelling variants of the anomalous word that are much closer, in WordNet terms, to the contextual words. This distance is determined on the base of paradigmatic relations (synonyms, hyponyms, hyperonyms) practically only between nouns. The syntactic relations between the words in a matched pair are ignored—the words from different sentences or even paragraphs usually have to be considered—so that this method relies on a rather wide text window.

We follow the same general framework: an anomaly is a word that do not match (given a certain comparison procedure) with any context word, and for correction considered are the spelling variants of the anomalous word that do match with some context word. However, we match the words basing on the syntagmatic (rather than paradigmatic) relations. This leads to a much smaller context window—one sentence,—which is possible because such relations involve content words of all parts of speech: nouns, verbs, adjective, and adverbs, as opposed to only nouns in WordNet.

The paper is organized as follows. Section 2 explains the main idea underlying our approach to measuring text cohesion. Section 3 introduces the basic concept—collocation—used in our approach. Section 4 defines what types of malapropisms we

deal with and what types we do not, illustrating this on numerous examples. In Section 5 we give our algorithm, which relies on the possibility to test, for a given pair of words, whether they can form a collocation. In Sections 6 and 7 we introduce two resources used for such a test, and in Section 8 explain how they are used. In Section 9 we give some preliminary experimental results. Finally, Section 10 draws the conclusions.

## 2 Main Idea of Syntagmatic Approach

We consider a text as a set of syntactically linked pairs of content words, and the syntactic links are treated like in dependency grammars [17]. Each word can participate in several pairs, and the linear distance between components of the pairs within the same sentence is not relevant. In correct texts, the components of any such pair are semantically compatible, forming word combinations of various grades of stability, stable word combinations being referred to as collocations. We consider collocations of content words only, and they are usually of different parts of speech.

A text with each content word participating at least in one collocation is considered cohesive. For example, in the sentence

*A* word(1) *form*(2) *is* *augmented*(2) *with* *information*(2) *necessary*(2) *for the* *correct*(1) *processing*(3) *of the* *data*(1)

there are the following collocations: *word* *form*; *form* —*is*— *augment*(*ed*); *augment*(*ed*) —*with*— *information*; *information* *necessary*; *necessary* —*for*— *processing*; *correct* *processing*; *processing* —*of*— *data*, with underlined words as participants. A syntactic link within a pair can be immediate (*word* *form*) or realized through a functional word(s): a preposition, an auxiliary verb(s), a conjunctive, etc. (*necessary* —*for*— *processing*). Above, the numbers in parentheses next to each content word indicate how many collocations it participates in. These are values of text cohesion counters, see below.

When a content word is replaced by another one inappropriate in the context (for example, *word* is replaced by *world*), some syntactic links, even if remaining grammatically correct, could contradict collocation formation (?*world* *form*), and hence some content words could become semantically isolated, just as in [12], [13]. We consider semantically isolated words (with the cohesion counter containing zero) as indicators of possible malapropisms, thus obtaining a tool for their detection.

The same cohesion notion serves for selecting candidates to correct malapropisms. A special generator outputs candidates similar to the suspicious word, and each is evaluated as to its collocation cohesion within the sentence. All candidates restoring the positive cohesion value are shown to the user for the final decision.

The crucial problem is how to test whether a given pair of content words in a text forms a collocation. The best resource for collocation testing is a collocation DB, i.e. a vocabulary of content words and a collection of syntactical links between them, each corresponding to a specific collocation. A collocation DB replies whether the given pair

sponding to a specific collocation. A collocation DB replies whether the given pair of components can form a collocation and whether morphological features of components, the intermediate functional words, and the word order correspond to the type of the syntactical link recorded in the DB.

The collocation DBs already exist (see below), but for English they are in an initial stage of development, in both structural and capacity aspects. Thus, in autonomous use, these DBs poorly cover open texts and can be scarcely used for the application under consideration.

The alternative resource for collocation testing are Internet search engines. For example, Google engine manages a few billions of web-pages in English, so that among them nearly all thinkable English word and their combinations can be found. However, the engines have several disadvantages, among them weak searching tools, tremendous informational noise, and augmented reaction time. To diminish the number of accesses, a cache deposit based on searches of potential collocation through Internet can be created at runtime, for each given working session. Trying to combine advantages of all resources mentioned, we propose to unite them into a multistage system.

This paper reflects on-going research on this multistage system. It proposes a draft algorithm for malapropism detection and correction based on the collocation cohesion counters. We describe the collocation testing resources mentioned above. Several examples of real-word errors usually considered as English malapropisms permits, firstly, to determine the notion of malapropism more strictly as violation of linguistic knowledge, i.e. words compatibility in formation of collocations or WordNet-like semantic links, and secondly, to mentally trace them with real accesses to Google, thus illustrating and roughly evaluating our method.

## 3 Collocations

To clarify what we mean by a collocation, let us recall that each sentence in natural language is a sequence of *word forms*. Commonly, these are strings of letters from one delimiter to the next (e.g., *links*, *are*, *very*, *short*), but we also call word forms short chains of strings used in a given sense only together (*point of view*, *fountain pen*, *because of*, *of course*, *a fortiori*, etc.).

We divide all word forms into three categories:

***Content words***: nouns; adjectives; adverbs and adverbial expression except of parenthetic expressions (see below); and verbs except auxiliary, modal, and phasal ones (see below).

***Functional words***: prepositions including those coming after verbs (*come* ***in***); coordinative conjunctions (*and*, *but*, *or*); auxiliary verbs (*to be*, *to have*); modal verbs (*can*, *may*, *must*...); phasal verbs (*to begin*, *to continue*, *to stop*...), and articles.

***Stop words***: pronouns; personal names except of the well known geographic and economic objects included in thesauri and encyclopedias; parenthetical expressions (*by contrast, of course, a fortiori, as the matter of fact...*), all other conjunctions, and other parts of speech.

According to dependency grammars [17], each sentence can be represented at the syntactic level as a labeled dependency tree with directed links "governor → its dependent" between word-form nodes. Following along these links in the same direction of the arrows from one content node through any linking functional nodes up to another content node, we obtain labeled subtree structure corresponding to a word combination. If this is a sensible text, the revealed combination is a collocation. For example, in the sentence *she hurriedly went through the big forest*, there are collocations ***went → through → forest, hurriedly → went*** and ***big → forest,*** but not *she → **went*** and *the → **forest***, since these contain non-content words at extreme nodes.

Such operational definition of collocation guarantees that content nodes at the extremes are syntactically linked, whereas their belonging to a semantically correct (i.e., sensible, conceptualized) text guarantees that they are semantically compatible. The compatibility is observed in (purely or partially) idiomatic expressions or in free combinations [18]. As to the stability of collocations usually evaluated by statistics, it is automatically implied by the two mentioned features. Internet shows that any semantically correct word combination eventually realizes several times.

We maintain the same term *collocations* in the situations when content nodes are linked through a more complicated tree-like structure, e.g. ***conversation → was → short***, ***error → will → cause*** or ***soldier → might → begin → (to) → make***. Here all functional words are verbs, and one among them is the root of the subtree (= predicate of the clause), with a noun (= subject of the clause) as one dependent and a verb in infinitive as another. The verb can finish a rather long chain like *might have to begin to want*. The only exception of the given definition is for the combinations like *soldier might to begin the job*. Hereby the two collocations are to be singled out: ***soldier → begin*** and ***begin → job***, with only one extreme pertaining to content words.

To fully describe a collocation, it necessary to give the content words and to specify the link, i.e. intermediate functional word(s), morphological features of content words, and preferable linear orders of the participants. A specific link depends on syntactical type of the collocation. The most frequently used types in European languages are "the modified → its modifier" (*strong tea*); "the verb → its noun complement" (*go to cinema*); "the verb predicate → its subject" (*light failed*); and "the noun → its noun complement" (*struggle against terrorism*) [5], [6].

In texts, collocation components can be linearly separated not only by their own functional word(s) but by many other words usually dependent on the same governor. To put it otherwise, the close context in a dependency tree is in no way equal to the close linear context.

Any collocation has its normalized (= dictionary) and textual forms, the latter composing the so-called morphological paradigm of collocation. For example, the collocation *go to cinema* comprises the four member paradigm: *go to cinema, going to cinema, gone to cinema, went to cinema.* In languages with rich morphology (e.g., of Romance or Slavic groups), these paradigms can be much broader.

## 4 The Type of Malapropisms We Consider

To delimit the scope of the proposed method, let us discuss types of real-word errors we consider malapropisms (cf. also [14]). Our examples are partially taken from [19].

**Violation of linguistic knowledge**. The word used in text is semantically distant from the intended one, so it becomes semantically incompatible with context, violating linguistic knowledge of two types. The first type of errors concerns the collocation compatibility:

1. *They travel around the* ***word*** (stands for *world*).
2. *The salmon swims upstream to* ***spoon*** (for *spawn*).
3. *Take it for* ***granite*** (for *granted*).
4. *The* ***bowels*** (for *vowels*) *are pronounced distinctly*.
5. *She has very loose* ***vowels*** (for *bowels*).
6. *They wear* ***turbines*** (for *turbans*) *on the heads*.
7. *This is an* ***ingenuous*** (for *ingenious*) *machine for peeling bananas.*
8. *Quite affordable* ***germs*** (for *terms*) *were proposed.*
9. *We study* ***dielectric*** (for *dialectic*) *materialism.*
10. *Children have equal* ***excess*** (for *access*) *to school*.
11. *The kinds of Greek columns are Corinthian, Doric, and* ***Ironic*** (for *Ionic*).
12. *The desert was activated by* ***irritation*** (for *irrigation*).
13. *This is the hysterical* (for *historical*) *center.*
14. *This is only a scientific* ***hypotenuse*** (for *hypothesis*).

The second type concerns WordNet-type semantic links:

15. *The four* ***seasons*** (for *seasonings*) *are salt, pepper, mustard, and vinegar.*
16. *The habitants of Moscow are called* ***Mosquitoes*** (for *Muscovites*).

The intended word *seasoning* is hyperonym to *salt, pepper, mustard,* and *vinegar*; whereas *Muscovites* is semantic derivative of *Moscow*.

**Violation of extra-linguistic knowledge**. In the first case, the substitute contradicts general knowledge about the world:

1. *This man is a real knight, a regular* ***Don Coyote*** (for *Don Quixote*).
2. *He studies in Toronto,* ***England*** (for *Canada*).
3. *In nineteenth century* ***pheasants*** (for *peasants*) *led a terrible life*.

In the second case, the substitute contradicts common sense reasoning:

4. *Lead the way and we will* ***precede*** (for *proceed*).
5. *His mother died in* ***infancy*** (for *youth*).
6. *Händel was* ***half*** (for *partially*?) *German,* ***half*** *Italian, and* ***half*** *English*.

Every human (but not a computer without a developed reasoning ability) knows that if somebody leads his/her way, then the follower can only *proceed* the leader (example 4); if a female died in infancy she had no children (example 5); no dividable entity can have three halves (example 6), etc.

We consider below malapropisms only of the first category—with the violation of linguistic knowledge.

## 5 Text Cohesion and Algorithm of Malapropism Detection and Correction

A text is a sequence of content words alternating with functional or stop words. For the *i*-th content word *W*(*i*), the cohesion counter *CC*(*i*) is the number of collocations it is in.

At the start, all *CC*(*i*) are zeroes. The algorithm scans the text sentence by sentence and tests all pairs consisting of a current content word and one of already scanned. The collocation test includes matching against available collocation testing resources. Revealing a collocation implies the increment by 1 of the cohesion counter *CC* for both components of the pair.

The algorithm relies on a Boolean function **Combinable**(*V*,*W*). For a pair of words *V* and *W*, it defines their combinability to a collocation in a manner heavily dependent on the resource used.

When the resource is a collocation DB, **Combinable**(V,W) is admitted true if the corresponding collocation is present in this DB, i.e. both components are in its dictionary, and potential syntactical link between them corresponds to the features recorded in the DB. Hence a local syntactic analysis is necessary.

When the resource is Google, **Combinable**(*V*,*W*) uses a simpler criterion of statistical nature. The words are admitted combinable if the mutual information inequality is satisfied [15]:

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

where $N(V,W)$ is the number of web-pages where $V$ and $W$ co-occur, $N(V)$ and $N(W)$ are numbers of the web-pages evaluated separately, $N_{\max}$ is the total web-page number managed by Google for the given language. As a bottom approximation to $N_{\max}$, the number $N(MFW)$ of the pages containing the most frequent word *MFW* can be taken. For English, *MFW* is '*the*,' and $N$('*the*') is evaluated to 2.5 billions of pages. This inequality rejects a potential collocation, if its components co-occur in a statistically insignificant quantity.

The high level procedure for detection and correction of malapropisms processes sentences of a text one by one:

```
DetectCorrectMalapropisms
     for each sentence p repeat
           EvaluateSentence(p);
           CorrectSentence(p);
```

The procedure **EvaluateSentence**(p) includes collocation tests:

```
EvaluateSentence(p)
     for each word i in the sentence p repeat
           CC(i)=0;
           for each word j in p such that j < i repeat
                 if ContentWord(W(j)) & Conten-
tWord(W(i))
                   & Combinable(W(j),W(i)) then
                       CC(j)+=1, CC(i)+=1;
```

The procedure **CorrectSentence**(p) includes procedure **ReevaluateSentence**(*p*,*i*). It is similar to **EvaluateSentence**(*p*) but revises the cohesion counter only for W(*i*):

```
CorrectSentence(p)
     for each word i in the sentence p repeat
           if CC(i)=0 & ContentWord(W(i)) then
                 ListOfCandidates = Ø;
                 repeat
                       NewCandidate = SearchCandi-
date(i);
                       InsertTemporally(NewCandidate,i);
                       ReevaluateSentence(p,i);
                       if CC(i)>0 then
                             Insert(NewCandidate,
                                   ListOfCandidates);
                 until not NewCandidate;
                 UserTests(ListOfCandidates);
```

All candidates are generated, ordered against the cohesion counter values at the point of suspicion, and then shown to the user's test. Note that the zero cohesion value can be revealed either for the erroneous word, or for another word syntactically linked with the erroneous one before making error, or in both places. In the given version, **CorrectSentence**(*p*) does not consider this complication.

## 6 Collocation Database

The optimal resource for testing collocations collocation databases already exist, both in scientific practice and in the market. For English, the BBI dictionary [3] in printed form contains several thousands of collocations, mainly semiphrasemes. This is about ten times less than necessary for stable malapropism processing. The Oxford Collocations Dictionary for Students of English [20] contains 170,000 collocations for nearly 12,000 headwords. This is about three time less than it is necessary for autonomous use, but is practically sufficient for the use within the multistage system, after reformatting its electronic DB. The Advanced Reader's Collocation Searcher (ARCS) [4] a commercial Internet-based system is announced with a million of collocations. However, judging from its public domain samples, real number is much less, and functional words are not attributed properly to collocations. Thus, ARCS seems acceptable but only for a simplified collocation tests.

To the date, we are not aware of collocation DBs for other European languages except for Russian. Compiled in the 90s, CrossLexica collocation DB [5], [6], [7] contains now more that 800,000 collocations of various semantic and syntactic types. It also contains thousands of WordNet-like semantic links, among which synonymous and hyperonymous ones are especially important, permitting to enrich the collocation collection at runtime for rare or trivial cases [7].

While testing a pair of content words against a collocation DB, both of them are passed to the DB, which normalizes them into their dictionary forms and searches if any collocation with these components is already recorded. If so, it tests functional words, morphologic features, and the word order.

Several links of the same type between the two dictionary entries can exist, e.g., ***go** to the **country*** and ***go** through the **country***, and several word senses for each component can be shown. We consider the test successful if at least one link option is recorded in the DB.

## 7 Google Search Engine

To piece out the absence or the limited size of available DBs for the majority of European languages, an alternative option for collocation testing is necessary, even if not so trustworthy. Internet search engines permit to retrieve thousands or even millions of

web-pages containing a queried word or a word combination in a moment. Each delivery is supplied with approximate statistics: how many pages correspond to the query (cf. [1], [2] on Internet applications for other needs of computational linguistics).

Being swift and efficient, Google engine allows only two search options:

1. The ultimately strict option, when the exact contiguous co-occurrences of the queried components are searched, and
2. The most loose option, when any co-occurrences of the components within a web-page are searched, irrespectively to their order and the linear distance within the page.

The first option roughly minifies the amount of occurrences for the given paradigm member, whereas the second option majorizes this number even more roughly. Both options permits to retrieve entire morphological paradigms of not more than ten members. What is more, Google prohibits to see headers of more than a thousand pages for any query. Hence, it is necessary to develop an approximate method to evaluate the co-occurrence number of any two components as a real collocation, based only on the data mentioned above.

Another deficiency of Internet is its immense informational noise. There are many foreigners, as well as ignorant and negligent persons, among web authors. Thus, diversified errors are quite numerous, including malapropisms.

Last but not least deficiency of Internet is its tremendous heaviness. Indeed, testing a collocation by means of a DB requires one access to the disk (several milliseconds), whereas one access to Goggle requires by hundreds more, mainly because of the network access delay.

## 8 Multistage Collocation Testing System

To minimize the number of accesses to Internet, it is reasonable to create a runtime cache depository in each text elaborating session. At the beginning, the cache is empty, but with each access for a potential collocation it broadens, if the pair satisfies the statistical criterion.

The function **Combinable**(*V*,*W*) uniting all resources mentioned could work as follows:

> If the given word pair in the text is recorded in the collocation DB, it is recognized as a collocation and further tests stop, else
>
> If this pair is already recorded in the cache created in this session, the pair is admitted as a collocation and further tests stop, else
>
> The pair is tested immediately through Internet that gives the ultimate decision. If it is positive, the pair is inserted to the cache.

## 9 Some Experimental Results

Now we will semi-automatically trace our small malapropism collection (cf. sentences in Section 2) using the proposed procedures. Suppose that the relevant word pairs in their malapropos version are not in the collocation DB, so we have to access Internet. For comparison of availability in Google of both correct and malapropos versions of the sentences, we extracted from each of them only relevant word pairs, measuring corresponding statistics with the strict search option. The following table contains the statistics for the first 14 examples:

| Example | Possible collocation | Correct version | Malapropos version |
|---|---|---|---|
| 1 | *travel around the* ***word*** | 55400 | 20 |
| 2 | *swim to* ***spoon*** | 23 | 0 |
| 3 | *take for* ***granite*** | 340000 | 15 |
| 4 | ***bowels*** *are pronounced* | 767 | 0 |
| 5 | *loose* ***vowels*** | 2750 | 1320 |
| 6 (a) | *wear* ***turbines*** | 3640 | 30 |
| 6 (b) | ***turbines*** *on the heads* | 25 | 0 |
| 7 | ***ingenuous*** *machine* | 805 | 6 |
| 8 | *affordable* ***germs*** | 1840 | 9 |
| 9 | ***dielectric*** *materialism* | 1080 | 4 |
| 10 (a) | *equal* ***excess*** | 457000 | 990 |
| 10 (b) | ***excess*** *to school* | 19100 | 4 |
| 11 | ***Ironic*** *columns* | 5560 | 28 |
| 12 | *activated by* ***irritation*** | 22 | 10 |
| 13 | ***histerical*** *center* | 90000 | 7 |
| 14 | *scientific* ***hypotenuse*** | 7050 | 0 |

One can notice that:

In four pairs (2, 4, 6b, 14) of 16 the wrong combinations are not in Google, so errors are easily detectable and correctable.

In seven pairs (1, 3, 7, 8, 9, 10b, 13) the malapropisms are met in small quantities and are acknowledged as wrong by the threshold procedure, so they can be processed too.

In five pairs (5, 6a, 10a, 11, 12) the unintended combinations do exist in English, so Google does not acknowledge them as malapropisms on a fair ground.

Though the combinations *equal excess* (10a) and *ironic columns* (11) do exist in English, rather big part of their occurrences in Google are malapropisms creating evident informational noise.

In the examples 6 and 10, two pairs are relevant for malapropism processing, and only one pair in each (6b and 10b respectively) ensures detection of the errors.

The malapropism in the example 15 is detectable but not correctable, since all its pairs are not collocations either in intended or in malapropos version. In the example 16, the pairs {*habitants, Moscow*} and {*called, Mosquitoes*} are true collocations, so the malapropism is left undetected and thus uncorrected. Note that purely semantic methods of [12, 13] easily detect and correct such errors.

Hence, based on our small collection, we can estimate the recall of the method as 11/16 68%. Compared to [13], it seems high.

As to the precision, we may not come to any conclusions based only on malapropos sentences. However, we can estimate false alarm frequency looking through several arbitrary chosen correct sentences. For our purposes, we considered two initial pages of [13] as a text. By manual scanning we obtained no false alarms at all.

To illustrate that Google does contain occurrences of nearly each collocation, we took the first sentence of the mentioned text and found 16 collocations with the following statistics of their strict option occurrences:

| Pair tested | Occurrences in Google | Pair tested | Occurrences in Google |
|---|---|---|---|
| *conventional checkers* | 13 | *each token* | 14500 |
| *spelling checkers* | 5830 | *token of a text* | 10 |
| *checkers detect* | 56 | *comparing/-ed/-es against a dictionary* | 14 |
| *detect errors* | 21800 | *dictionary of words* | 12900 |
| *typing errors* | 35800 | *words are known* | 990 |
| *detect simply* | 115 | *known to be spelled* | 31 |
| *detect by comparing* | 64 | *spelled correctly* | 50400 |
| *comparing tokens* | 10 | | |

Thus, the level of false alarms in our method is expectedly very low.

## 10 Conclusions and Future Work

We have defined malapropisms in a specific manner: they violate purely linguistic knowledge, mainly that of the semantic compatibility for collocation composing. The collocation components are content words syntactically linked in a sentence, directly of through functional words. An algorithm is proposed for their detection and correction based on collocation cohesion destructible by malapropisms. The roughly evaluated levels of recall and false alarms proved to be competitive and thus justify the on-going research.

The efficiency of the detection and correction heavily depends on the resources used for collocation testing. A large and accurately revised collocation DB with ca. million collocations of various types, being used autonomously, will achieve the best results. How-

ever, till now collocation DBs have limited sizes for English and are unknown for other European languages (except for Russian).

In such a situation, a good alternative is a multistage testing system containing a collocation DB of a modest size, a cache accumulating unknown collocations already met in the given editing session, and runtime accesses to Internet for statistical estimations of the newcomers.

It seems topical to continue this study in order to answer:

How to evaluate, with acceptable precision, the number of web-pages with a given content word pair, in which they really form collocations?

How to rationalize correction of malapropisms that are detected with our method? The nucleus of the correction subsystem is a candidate generator. It is necessary to take as candidates not only spelling variation options [13], but also words similar to the suspicious word in a different sense. Such words are referred to as paronyms. Beforehand compilation of candidates of both types can significantly minimize correction time [9].

Is it worthy to use an intermediate stage in the testing process, i.e. to use the results of a forehand automatic search of collocations through a large corpus of texts? The text corpora are usually much less noisy linguistic data than Internet, and the access time to the data collected beforehand is comparable to that of collocation DBs.

How to improve the multistage testing system by organizing, with minimal expenses of intellectual labor, an information feedback on collocations from the Internet-based cache to the corpus-based data and then to the collocation DB?

How to distribute confidence between various testing resources, in order to minimize both decision errors — to omit a malapropism *vs*. to exert a false alarm — in the situation, where more information-rich resources are at the same time more noisy? To put it differently, when is it worthy to continue the tedious accesses to Internet, if the false alarm can be treated by the user in a shorter time?

How to combine advantages of our method and the WordNet-based method proposed earlier [13]?

## References


[1] Agirre, E., D. Martinez. *Exploring automatic word sense disambiguation with decision lists and the Web*. Proceedings of the Semantic Annotation And Intelligent Annotation Workshop. Organized by COLING, Luxembourg, 2000.

[2] Agirre E., Martinez D. *Integrating selectional preferences in WordNet*. Proceedings of the first International WordNet Conference in Mysore, India, 21-25 January 2002.

[3] Benson, M., E. Benson, R. Ilson. *The BBI Combinatory Dictionary of English*. John Benjamin, 1989.

[4] Bogatz, H. *The Advanced Reader's Collocation Searcher* (*ARCS*). ISBN 09709341-4-9, 1997. http:www.asksam.com/web/bogatz

[5] Bolshakov, I. A. *Multifunction thesaurus for Russian word processing.* Proceedings of 4th Conference on Applied Natural Language Processing, Stuttgart, Germany, October 13-15, 1994, p. 200-202.

[6] Bolshakov I. A., A. F. Gelbukh. *A Very Large Database of Collocations and Semantic Links.* In: Bouzeghoub *et al.* (eds.) *Natural Language Processing and Information Systems.* Natural Language Applications to Information Systems. Lecture Notes in Computer Science 1959, Springer, 2001, p. 103-114.

[7] Bolshakov, I. A., A. Gelbukh. *Heuristics-Based Replenishment of Collocation Databases*. In: E. Ranchhold, N. J. Mamede (Eds.) *Advances in Natural Language Processing*. Lecture Notes in Artificial Intelligence 2379, Springer Verlag, 2002, p. 25-32.

[8] Bolshakov, I. A., A. Gelbukh. *Word Combinations as an Important Part of Modern Electronic Dictionaries*. Procesamiento del Lenguaje Natural No. 29, September, 2002, p. 47-54.

[9] Bolshakov, I. A., A. Gelbukh. *Paronyms for Accelerated Correction of Semantic Errors*. (to appear).

[10] *The New Encyclopædia Britannica*. Micropædia Vol. 7. Encyclopædia Britannica, Inc., 1998

[11] Golding, A.R., Y. Shabes. *Combining trigram-based and feature-based method for context-sensitive spelling correction.* Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, 1996, Santa Cruz, CA, p. 71-78.

[12] 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.

[13] Hirst, G., A. Budanitsky. *Correcting Real-Word Spelling Errors by Restoring Lexical Cohesion.* Computational Linguistics (to appear).

[14] Kukich, K. *Techniques for automatically correcting words in texts*. ACM Computing Surveys. Vol. 24, 1992, p. 377-439.

[15] Manning, Ch. D., H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999. 680 P.

[16] Mays, E., F.J. Damerau, R.L. Mercer. *Context-based spelling correction*. Information Processing and Management. V. 27, No. 5, p. 517-522.

[17] Mel'cuk, I. *Dependency Syntax: Theory and Practice.* SONY Press, NY, 1988.

[18] Mel'cuk, I. *Phrasemes in Language and Phraseology in Linguistics*. In: M. Everaert et al. (Eds.) *Structural and Psychological Perspectives*. Lawrence Erlbaum Associates Publ., Hillsdale, NJ / Hove, UK, p. 169-252.

[19] *Miscellaneous Malapropisms. Quotations*. http://www.geocities.com/~spanoudi/quote-08.html

[20] *Oxford Collocations Dictionary for Students of English.* Oxford University Press. 2003.