[5] Wolfgang Schade, Karola Witschurke, Cornelia Rataj; Improved character recognition of typed documents from middling and lower quality based on application depending tools Processes, results, comparison; Electronic Imaging Events in the Visual Arts - EVA 2003, Berlin 2003

[6] Alexander Geschke, Eva Fischer; Memorial Project - A complex approach to digitisation of personal records; Electronic Imaging Events in the Visual Arts - EVA 2003, Berlin 2003

[7] Henryk Krawczyk, Bogdan Wiszniewski; Definition gleichartiger Dokumententypen zur Verbesserung der Erkennbarkeit und ihre XML-Beschreibung; Electronic Imaging Events in the Visual Arts - EVA 2003, Berlin 2003

[8] Henryk Krawczyk, Bogdan Wiszniewski; Digital Document Life Cycle Development; International Symposium on Information and Communication Technologies, ISICT 2003, Dublin, Ireland

[9] Henryk Krawczyk, Bogdan Wiszniewski; Visual GQM approach to quality-driven development of electronic documents; Second International Workshop on Web Document Analysis, WDA2003, Edinburgh, UK

[10] Bogdan Wiszniewski; Projekt IST-2001-33441-MEMORIAL: Zestaw narzędziowy do tworzenia dokumentów cyfrowych z zapisów osobowych; I Krajowa Konferencja Technologii Informacyjnych 2003 TUG

[11] Alexander Geschke; MEMORIAL Project Overview; Proc. EVA Harvard, Symposium about Collaboration of Europe, Israel and USA, Harvard Library 1-2.10.2003

[12] Jacek Lebiedź, Arkadiusz Podgórski, Mariusz Szwoch; Quality Evaluation Of Computer Aided Information Retrieval From Machine Typed Paper Documents; Third conference on Computer Recognition Systems KOSYR'2003

[13] Witold Malina, Bogdan Wiszniewski; Multimedialne biblioteki cyfrowe; Sesja 50-lecia WETI-PG

[14] Dr. Alexander Geschke, Dr. Wolfgang Schade; The EU Project Memorial - Digitisation, Access, Preservation; Electronic Imaging Events in the Visual Arts - EVA 2002, Berlin 2002

[15] S. Rogerson, B. Wiszniewski; Legislation and regulation: emphasis on European approach to Data Protection, Human Rights, Freedom of Information, Intellectual Property, and Computer Abuse; PROFESSIONALISM IN SOFTWARE ENGINEERING PSE'03

Author's Information

Karola Witschurke - GFaI, Rudower Chaussee 30, 12489 Berlin, Germany; e-mail: witschurke@gfai.de

EXPERIMENTS IN DETECTION AND CORRECTION OF RUSSIAN MALAPROPISMS BY MEANS OF THE WEB

Elena Bolshakova, Igor Bolshakov, Alexey Kotlyarov

Abstract: Malapropism is a semantic error that is hardly detectable because it usually retains syntactical links between words in the sentence but replaces one content word by a similar word with quite different meaning. A method of automatic detection of malapropisms is described, based on Web statistics and a specially defined Semantic Compatibility Index (SCI). For correction of the detected errors, special dictionaries and heuristic rules are proposed, which retains only a few highly SCI-ranked correction candidates for the user's selection. Experiments on Web-assisted detection and correction of Russian malapropisms are reported, demonstrating efficacy of the described method.

Keywords: semantic error, malapropism, error correction, Web-assisted error detection, paronymy dictionaries, correction candidates, Semantic Compatibility Index.

ACM Classification Keywords: I.2.7 [Artificial Intelligence]: Natural language processing – Text analysis

Introduction

Modern computer text editors and spellers readily detect spelling errors and some syntactic errors, primarily, mistakes in word agreement. Step by step, editing facilities of computers are being extended, in particular, by taking into account specificity of the particular text style and genre [4]. The topical problem is now semantic mistakes, which are hardly detectable because they violate neither orthography nor grammar of the text.

Malapropism is a particular type of semantic mistakes, which replace one content word by another similar word. The latter has the same morpho-syntactic form but different meaning, which is inappropriate in the given context, e.g., *animal word* or *massy migration* given instead of *animal world* and *massive migration*. To correct such mistakes, computer procedures are required that reveal erroneous words and supply the user (human editor) with selected candidates for their correction. However, only few papers (cf. [3, 5]) are devoted to the problem of malapropism detection and correction.

A method for malapropism detection proposed in [5] relies on recognition in the text of words (mainly nouns) distant from all contextual ones in terms of WordNet semantic relations (of synonymy, hyponymy, hyperonymy, etc.). Syntactic relations between words are ignored, and words from different sentences or even paragraphs are analyzed.

In the paper [3] malapropism detection is based on syntactico-semantic relations between content words, thereby much smaller context – only one sentence – is needed for error detection. Specifically, sentences are considered consisting of syntactically related and semantically compatible combinations of content word, the so-called *collocations*. It is presumed that malapropisms destroy collocations they are in: they violate semantic compatibility of word combinations while retaining their syntactic correctness.

In order to detect errors in a sentence, all pairs of syntactically linked content words in it are verified as collocations: their semantic compatibility is tested. Words of four principal parts of speech (POS) – nouns, verbs, adjective, and adverb – are considered as collocation components. To test whether a word pair is a collocation, three types of linguistic resources are proposed: a precompiled collocation database like CrossLexica [1], a text corpus, or a Web search engine like Google or Yandex.

This paper develops the latter method on the basis of experiments with Yandex as a resource for collocation testing. The Web is widely considered now as a huge, but noisy linguistic resource [6]. For the Web, it proved necessary to revise heuristic rules for malapropism detection and correction.

Following [3] we consider only malapropisms that destroy collocations. A malapropism is detected if a pair of syntactically linked content words in a sentence exhibits the value of a specially defined *Semantic Compatibility Index* (SCI) lower than a predetermined threshold. Below we call malapropism the whole pair detected as erroneous.

For malapropism correction, in contrast with the blind search of editing variants used in [3, 5], we propose to use beforehand compiled dictionaries of paronyms, i.e. words differing in some letters or in some morphs. The dictionaries provide all possible candidates for correction of a malapropos word, and the candidates are then tested in order to select several highly SCI-ranked correction candidates for ultimate decision by the user.

The proposed method was examined on two sets of Russian malapropisms. The first set of a hundred of samples was used to adjust heuristic threshold values, whereas the second justified these values. Since collocation components (hereafter *collocatives*) may be adjacent or separated by other words in a sentence, in the experiment we took into account the most probable distances between collocatives, which were previously determined through Yandex statistics.

Dictionaries of Paronyms

For correction of malapropos words, quick search of similar words are required. Words similar in letters, sounds or morphs, are usually called paronyms. In any language, only a limited portion of words has paronyms, and paronymy groups are on an average rather small. Hence, it is reasonable to gather paronyms before their use.

For our purposes, we consider only ***literal*** (e.g., Eng. *pace* Vs. *pact*, or Rus. *краска* Vs. *каска*) and ***morphemic*** paronyms (e.g., Eng. *sensible* Vs. *sensitive*, or Rus. *человечный* Vs. *человеческий*). Russian paronyms of these two types were compiled in corresponding dictionaries, which were preliminary described in [2].

The dictionary of Russian literal paronyms consists of word groups. Each group includes an entry word and its one-letter paronyms. Such paronyms are obtained through applying to the entry word of an elementary editing operation: insertion of a letter in any position, omission of a letter, replacing of a letter by another one, and permutation of two adjacent letters. For example, Russian word *белка* has one-letter paronymy group {*булка*, *елка*, *телка*, *челка*, *щелка*}.

The paronymy groups include words of the same part of speech (POS). Moreover, nouns of singular number and nouns of plural number, as well as nouns for different genders of singular have separated groups. Similar division is done for personal and other forms of verbs. Such measure is necessary for retaining syntactic links between words in the sentence while correcting an erroneous word. For this purpose, we extract malapropos word from the text (e.g., Rus. *белкой*), reconstruct its dictionary morphological form (*белка*), take from the dictionary corresponding paronym (e.g., *булка*), change its morphological form (taking it the same as for sourse malapropos word), and, finally, replace the erroneous word by the resulted word (*булкой*).

An entry of dictionary of Russian morphemic paronyms presents a group of words of the same POS that have the same root morph but differ in auxiliary morphs (prefixes or suffixes), e.g. Rus. {*бегающий*, *беглый*, *беговой*, *бегущий*}.

By now, the developed dictionary of literal paronyms comprises 17,4 thousands of paronymy groups with the mean size 2,65, while the dictionary of morphemic paronyms contains 1310 groups with the mean size 7,1.

Method of Malapropism Detection and Correction

To facilitate understanding of key ideas of the method we should first clarify the notion of collocation adopted in the paper. Collocation is a combination of two syntactically linked and semantically compatible content words, such as the pair's *main goal* and *moved with grace*. Syntactic links are realized directly or through an auxiliary word (usually a preposition). If any of conditions indicated above does not hold, the corresponding word combination is not collocation, for example, *the forest*, *river slowly*, *boiling goal*.

There are several syntactic types of collocations in each language. The most frequent types in European languages are: "the modified word → its modifier"; "noun → its noun complement"; "verb → its noun complement"; "verb predicate → its subject"; and "adjective → its noun complement". Directed links reflect syntactic dependency "head → its dependent".

The most frequent types and subtypes of Russian collocations are given in Table 1. They are determined by POS of collocatives and their order in texts; **N** symbolizes noun, **Adj** is adjective or participle, **V** and **Adv** are verb and adverb correspondingly, and **Pr** is preposition. Subindex *comp* means noun complement, while subindex *sub* means the noun subject in nominative case. Subindex *pred* symbolizes specifically Russian predicative short form of adjectival.

Table 1. Frequent types and structures of Russian collocations

Type title	Type code	Type structure	English example	Russian example
modified → its modifier	1.1	Adj ← N	*strong tea*	*крепкий чай*
	1.2	Adv ← Adj	*very good*	*очень хороший*
noun → its noun complement	2.1	$N \rightarrow N_{comp}$	n/a	*огни города*
	2.2	$N \rightarrow Prep \rightarrow N_{comp}$	*signs of life*	*вызов в суд*
verb → its noun complement	3.1	$V \rightarrow N_{comp}$	*give message*	*искать решение*
	3.2	$V \rightarrow Prep \rightarrow N_{comp}$	*go to cinema*	*идти в кино*
	3.3	$N_{comp} \leftarrow V$	n/a	*здание затушили*
verb predicate → its subject	4.1	$N_{sub} \leftarrow V$	*light failed*	*мальчик пел*
	4.2	$V \rightarrow N_{sub}$	*(there) exist people*	*пропали письма*
	4.3	$Adj_{pred} \rightarrow N_{sub}$	n/a	*отправлен груз*
	4.4	$N_{sub} \leftarrow Adj_{pred}$	n/a	*порт открыт*
adjective → its noun complement	5.1	$Adj \rightarrow Prep \rightarrow N_{comp}$	*easy for girls*	*красный от стыда*
	5.3	$Adj \rightarrow N_{comp}$	n/a	*занятый трудом*

Within a sentence, collocatives may be adjacent either distant from each other. The distribution of possible distances depends on the collocation type and specific collocatives. For example, collocatives of subtypes 2.1 and 2.2 are usually adjacent, whereas the 3.1-collocation such as *give* → *message* can contain intermediate contexts of lengths 0 to 4 and even longer, e.g. *give her a short personalized message*.

Our definition of collocations ignores their frequencies and idiomatically. As for frequencies, the advance of the Web shows that any semantically compatible word combination eventually realizes several times, thus we can consider as collocations all those exceeding a rather low threshold.

The main idea of our method of malapropism detection is to look through all pairs of content words within the sentence under revision, testing its syntactic links and its semantic admissibility. If the pair (*V, W*) is syntactically connected but semantically incompatible, a malapropism is signaled.

When a malapropism is detected, it is not known which collocative is erroneous, so we should try to correct both of them. The situation is clarified in Fig. 1. The upper two collocative nodes form the malapropism. The nodes going left-and-down and right-and-down are corresponding paronyms for malapropism's nodes. Each paronym should be matched against the opposite malapropism's node, and any pair may be admissible, but only one combination corresponds to the intended collocation; we call it *true correction*.

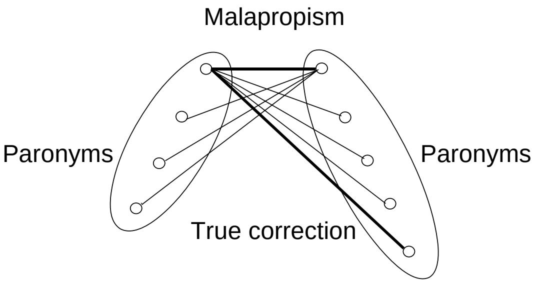

Fig. 1. Correction candidates and true correction

In such a way, all possible pairs of a collocative and its counterpart's paronym are formed, and we call them *primary candidates* for correction. The candidates are tested on semantic compatibility. If a pair fails, it is discarded; otherwise it is included into a list of *secondary candidates*. Then this list is ranked and only the best candidates are kept.

Obviously, for testing pairs (*V, W*) on semantic compatibility, using the Web as a text corpus, a statistical criterion is needed. According to one criterion, the pair is compatible if the relative frequency $N(V,W)/S$ of the co-occurrence of its words in a short distance in the whole corpus is greater than the product of relative frequencies $N(V)/S$ and $N(W)/S$ of occurrences of *V* and *W* taken separately (*N* means frequency; *S* is the size of the corpus). Using logarithms, we have the following threshold rule of pair compatibility:

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

where MII(*V, W*) is the mutual information index [7].

Since any search engine automatically delivers statistics about the queried word or the word combination measured in numbers of pages, to heuristically estimate the pair compatibility we propose a Semantic Compatibility Index (SCI) similar to MII:

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

where *N* is the number of relevant pages; *P* is a positive constant to be chosen experimentally; and *NEG* is a negative constant. A merit of SCI as compared to MII is that the total number of pages is not to be estimated. Similarly to MII, SCI does not depend on monotonic or oscillating variations of all statistical data in the search engine because of the divisor 2.

If SCI(V_m,W_m) < 0, the pair (V_m,W_m) is malapropism, whereas the primary candidate (V,W) is selected as a secondary one according to the following threshold rule:

(SCI(V_m,W_m) = *NEG*) **and** (SCI(V,W) > Q) **or** (SCI(V_m,W_m) > *NEG*) **and** (SCI(V,W) > SCI(V_m,W_m))

where Q (*NEG* < Q < *0*) is a constant to be chosen experimentally.

The resulted set of secondary candidates is ranked by SCI values. The *best candidates* are all with positive SCI (let be n of them), whereas only one candidate with a negative SCI value is admitted, if n=1, and two candidates, if n=0.

Experimental Sets of Malapropisms

For experiments, we use two experimental sets – both of them consist of hundred of Russian sample malapropisms, which were mainly formed with the aid of the Web newswire. Specifically, we extracted collocations from the news messages, and one collocative of each collocation was then falsified using one of paronymy dictionaries, as a rule, the dictionary of literal paronyms (since literal errors are much more frequent in any language than morphemic ones). While falsifying, the morphological features of the word being changed (number, gender, person, case, etc.) were retained.

Then we again used paronymy dictionaries to make all possible correction candidates for each formed malapropism: through replacing of one word of the malapropism by its paronym we obtained the corresponding primary candidate. Among primary correction candidates, the before mentioned true correction (identical with intended collocation) was necessarily appeared.

1)1L 1.1 (проявил) кассовое сознание	'cash consciousness'
1L массовое сознание	'mass consciousness'
1L!! классовое сознание	'class consciousness'
1L кастовое сознание	'caste consciousness'
2L кассовое создание	'cash creature'
2M кассовое знание	'cash knowledge'
2M кассовое признание	'cash confession'
2M кассовое осознание	'cash perception'
2L кассовое познание	'cash cognition'
2)2L! 1.3 (песня)явно сдалась	'evidently capitulated'
2L!! явно удалась	'evidently succeed'
2M явно задалась	'evidently preset'
2M явно далась	'evidently given'
2M явно продалась	'evidently sold'
1L ясно сдалась	'clearly capitulated'
2M явно подалась	'evidently gone'
3)1L 2.1 (занят)смирением террористов	'by submission of terrorists'
1L!! усмирением террористов	'by pacification of terrorists'
1M примирением террористов	'by reconciliation of terrorists'
4)1L 2.2 кастеты с кадрами	'knuckledusters with frames'
1L!! кассеты с кадрами	'cassettes with frames'
2L кастеты с карами	'knuckledusters with retributions'
2L кастеты с кедрами	'knuckledusters with cedars'
1L катеты с кадрами	'legs with frames'
5)2L 3.3 протокол подманили	'protocol is dangled'
2L!! протокол подменили	'protocol is replaced'
2L протокол поманили	'protocol is drown on'
2L протокол подранили	'protocol is injured'

Fig. 2. Several malapropisms and their correction candidates with translations

The resulted sets consist of enumerated sample groups, each group corresponding to a malapropism and its primary candidates. Several sample groups are given in the first column of Fig. 2. Headlines of groups begin with the number of the changed collocative (1 or 2) and the symbol of the used paronymy dictionary (**L**iteral or

Morphemic). The next is code $n_1.n_2$ of syntactic type of the collocation (cf. Table 1), and then goes the malapropism string, may be with a short context given in parentheses. Lines with correction candidates begin with the number of the changed word (1 or 2) and the symbol of the used paronymy dictionary; true corrections are marked with '**!!**'. The translation of malapropisms and their correction variants in the second column of Fig. 2. exhibits the nonsense of wrong corrections.

In total, the first malapropism set includes 648 primary correction candidates, and the second, 737 candidates. So the mean number of primary correction candidates is ≈ 7 candidates per error.

Among the samples, the sets include also errors named *quasi-malapropisms* (their total number equals 16). A quasi-malapropism transforms one collocation to another semantically legal collocation, which can be rarer and contradict to the outer context, e.g., *normal manner* changed to *normal banner* or *give message* changed to *give massage*. An example of Russian *quasi-malapropism* is presented in Fig. 2, it is marked with '**!**' (cf. the second sample group). The detection of quasi-malapropisms (if possible) sometimes permits one to restore the intended words, just as for malapropisms proper.

Experiments with Yandex and Their Results

A specific collocation or malapropism met in a text has its certain distance between collocatives. However, to reliably detect malapropisms by means of the Web and the selected statistic criterion, we should put each word pair being tested in its most probable distance.

For this reason, we initially explored frequencies of various Russian collocative co-occurrences against the distance between them on the base of Yandex statistics, cf. Table 2 and Table 3. The statistics of co-occurrence frequencies (measured in the number of relevant pages) were accumulated for twelve collocations of various frequent types. The used queries contained collocatives in quotation marks separated with */n* indicating the distance *n* between the given words, for example, *+"столбы"/2+"дыма"*. Such queries give frequencies of the words encountered within the same sentence with distance between them equal to *n* (or the number of intermediate words equals to *n-1*).

The statistics show that, for all collocations, frequency maximums correspond to the numbers 0 or 1 of intermediate words, and such cases cover more than 60% of encountered word pairs (cf. the last column of Table 2). Since we cannot determine automatically whether counted Web co-occurrences are real collocations or mere encounters of words without direct syntactic links between, we look through the first fifty page headers, mentally analyzing their syntax. Thereby we ascertained that the most of the co-occurrences with adjacent collocatives or those separated by one word are real collocations.

Table 2. Yandex statistics of collocative co-occurrences

Collocation	Type	Number of intermediate words:					Percents in 0 and 1
		0	1	2	3	4	
Уделить внимание	3.1	52248	**72433**	9111	3537	1335	90%
Отправлен груз	4.3	779	**3408**	100	17	8	97%
Сбор информации	2.1	**141395**	32342	54354	31326	13566	64%
Спасатели обнаружили	4.1	**18534**	2440	929	524	740	91%
Здание потушили	3.3	**48**	7	14	10	0	70%
Сроки рассмотрения	2.1	**31517**	2918	2302	891	1075	89%
Затонувшее судно	1.1	**10250**	496	189	642	128	92%
Оценка деятельности	2.1	**29276**	22847	20373	5370	4183	64%
Занятый трудом	5.3	40	**413**	215	16	11	65%
Столбы дыма	2.1	**4382**	1420	507	79	93	90%
Сделать оговорки	3.1	355	**660**	269	44	14	76%
Приведем пример	3.1	**30665**	13106	6343	1376	580	84%

Thus, we can deduce that for frequent types of Russian collocations the most probable distance between collocatives (measured in the number of intermediate words) equals 0 or 1. As to collocatives linked through prepositions, the most probable distances at both interval are equal to 0, cf. Table 3. The table shows the distribution of frequencies for possible combinations of two distances: between the first collocative and the preposition and the preposition and the second collocative (e.g., combination 0-1 means that the first collocative and the preposition are adjacent, whereas the preposition and the second collocative are separated by one word).

Table 3. Yandex statistics of co-occurrences for collocations with prepositions

Collocation	Number of intermediate words									
	0 –0	1–0	0–1	2–0	1–1	0–2	3–0	2–1	1–2	0–3
ворвались в здание	**9869**	123	156	69	0	74	24	1	9	2
тайники с оружием	**4775**	1	43	10	1	29	3	0	0	38
вызов в суд	**3633**	281	91	90	4	15	120	6	0	15
справиться с управлением	**5744**	16	177	2	0	11	2	0	0	12

Then we have applied our method to the both experimental sets by means of the computer program that gathers statistics of word pairs co-occurrences with the distance between them (measured in the number of intermediate words) equal to 0 or 1 (collocatives linked through a preposition were tested as adjacent triples). This in no way means that collocations cannot have more distant collocatives, but the Web is not suited for collocation testing at greater distances. The frequencies of word occurrences and co-occurrences gathered for several malapropisms are given in the second column of Fig. 3 (the repeating data for the collocatives are omitted).

```
1)1L 1.1 кассовое сознание       2, кассовое:354955,сознание:4770500
  1L массовое сознание           32973, массовое:916455
  1L!! классовое сознание        2927, классовое:38924
  1L кастовое сознание           56, кастовое:11799
  2L кассовое создание           10, создание:32199807
  2M кассовое знание             1, знание:7120311
  2M кассовое признание          0, признание:2437390
  2M кассовое осознание          0, осознание:823650
  2L кассовое познание           0, познание:605134
2)2L! 1.3 явно сдалась           13, явно:9871866, сдалась:198061
  2L!! явно удалась              6703, удалась:610646
  2M явно задалась               386, задалась:46599
  2M явно далась                 38, далась:88177
  2M явно продалась              2, продалась:24594
  1L ясно сдалась                2, ясно:10816398
  2M явно подалась               0, подалась:298216
3)1L 2.1 смирением террористов   0, смирением:79063,террористов:2762914
  1L!! усмирением террористов    3, усмирением:1787
  1M примирением террористов     0, примирением:17515
4)1L 2.2 кастеты с кадрами       0, кастеты:42266,кадрами:481878
  1L!! кассеты с кадрами         21, кассеты:2923258
  2L кастеты с карами            0, карами:19351
  2L кастеты с кедрами           0, кедрами:5666
  1L катеты с кадрами            0, катеты: 3151
5)2L 3.3 протокол подманили      0, протокол:7635243, подманили:3521
  2L!! протокол подменили        36, подменили:86957
  2L протокол поманили           0, поманили:7545
  2L протокол подранили          0, подранили:946
```

Fig. 3. Several malapropisms and their correction candidates with Yandex statistics

We used the first experimental set to adjust the necessary constants of our method. To obtain all negative SCI values for all proper malapropisms from the first set, we take *P* = 1200. The constant *NEG* = –100 is taken lower than SCI values of all occurrences counted as non-zero events. The constant *Q* = –7.5 is adjusted so that all candidates with non-zero occurrences have SCI values greater then this threshold.

Though all eight quasi-malapropisms were excluded while selecting the constant *P*, our method detects seven of them as malapropisms proper: their SCI values proved to be too low to be acknowledged as collocations. Our program selects 169 secondary candidates from 648 primary ones and then reduces them to 141 best correction candidates. Among the best candidates for the 99 malapropisms signaled, as many as 98 have true correction options, and only two of them are not first-ranked.

While testing the method and the determined constants on the second experimental malapropisms set, all its malapropisms and even all eight quasi-malapropisms were detected. 165 secondary candidates were selected from 737 primary ones; and the secondary ones were reduced to 138 best candidates. But for five detected malapropisms their true corrections do not enter corresponding lists of the best candidates, and three true corrections among them were not selected as secondary candidates (we admit that these collocations are rather infrequent in texts).

```
1)1L 1.1 кассовое сознание        -6,29  Detected
  1L массовое сознание             2,95  Best
  1L!! классовое сознание          2,11  Best
2)2L! 1.3 явно сдалась          -4,49    Detected
  2L!! явно удалась                1,20  Best
  2M явно задалась                -0,37  Best
3)1L 2.1 смирением террористов  -100,00  Detected
  1L!! усмирением террористов     -2,96  Best
4)1L 2.2 кастеты с кадрами      -100,00  Detected
  1L!! кассеты с кадрами          -6,28  Best
5)2L 3.3 протокол подманили     -100,00  Detected
  2L!! протокол подменили         -2,93  Best
```

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

We should note that the occasional omission of a true correction does not seem too dangerous, since the user can restore it in the case of error detection. Nevertheless, the most commonly used collocations among primary correction candidates always enter into the list of the best candidates, as true corrections or not.

For both experiments, the lists of the best candidates contain 1 to 4 entries, usually 1 or 2 entries; cf. the detected malapropisms with corresponding SCI values and decision qualifications in Fig. 4. The total decrease of correction candidates, from the primary to the best, exceeds 5.

Hence the results of our experiments are rather promising: for our experimental sets, the method of testing semantic compatibility through the Web has the recall 0.995. The proposed SCI is a quite good measure for detecting malapropisms, and the proposed heuristic rule for selection of secondary correction candidates is appropriate, whereas the heuristic rule for selection of best candidates may be slightly improved.

Conclusions and Further Work

A method is proposed for automatic detection and computer-aided correction of malapropisms. Experimental justification of the method was done on two representative sets of Russian malapropisms with the aid of Yandex search engine. While testing word pairs on their semantic compatibility through the Web, the most probable distances between the Russian words were taken into account.

Since the experiments gave good results, the problem of the Web statistics validity for collocation testing deserves to be investigated deeper. It would be worthwhile to extend the results of our study to broader experimental data and to other Web search engines. Of course, it is quite topical to develop a local grammar parser appropriate for malapropism detection, since for our experiments we extracted collocation components manually.

Since the Web proved to be adequate for testing semantic compatibility of collocations, we hope to use the method to develop procedures of automatic acquisition of collocation databases.

Bibliography

1. Bolshakov, I.A. Getting One's First Million…Collocations. In: A. Gelbukh (Ed.). Computational Linguistics and Intelligent Text Processing. Proc. 5th Int. Conf. on Computational Linguistics CICLing-2004, Seoul, Korea, February 2004. LNCS 2945, Springer, 2004, p. 229-242.
2. Bolshakov, I.A., A. Gelbukh. Paronyms for Accelerated Correction of Semantic Errors. International Journal on Information Theories & Applications. V. 10, N 2, 2003, p. 198-204.
3. Bolshakov, I.A., A. Gelbukh. On Detection of Malapropisms by Multistage Collocation Testing. In: A. Düsterhöft, B. Talheim (Eds.) Proc. 8th Int. Conference on Applications of Natural Language to Information Systems NLDB´2003, June 2003, Burg, Germany, GI-Edition, LNI, V. P-29, Bonn, 2003, p. 28-41.
4. Bolshakova, E.I. Towards Computer-aided Editing of Scientific and Technical Texts. International Journal on Information Theories & Applications. V. 10, N 2, 2003, p. 204-210.
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. Kilgarriff, A., G. Grefenstette. Introduction to the Special Issue on the Web as Corpus. Computational linguistics, V. 29, No. 3, 2003, p. 333-347.
7. Manning, Ch. D., H. Schütze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.

Authors' Information

Elena I. Bolshakova – Moscow State Lomonosov University, Faculty of Computational Mathematics and Cybernetic, Algorithmic Language Department; Leninskie Gory, Moscow State University, VMK, Moscow 119899, Russia; e-mail: bolsh@cs.msu.su

Igor A. Bolshakov – Center for Computing Research (CIC), National Polytechnic Institute (IPN); Av. Juan Dios Bátiz esq. Av. Miguel Othon Mendizabal s/n, U.P. Adolfo Lopez Mateos, Col. Zacatenco, C.P. 07738, Mexico D.F., Mexico; e-mail: igor@cic.ipn.mx

Alexey P. Kotlyarov – Moscow State Lomonosov University, Faculty of Computational Mathematics and Cybernetic, Algorithmic Language Department; Leninskie Gory, Moscow State University, VMK, Moscow 119899, Russia; e-mail: koterpillar@yandex.ru

A MATHEMATICAL APPARATUS FOR DOMAIN ONTOLOGY SIMULATION. AN EXTENDABLE LANGUAGE OF APPLIED LOGIC[1]

Alexander Kleshchev, Irene Artemjeva

Abstract: A mathematical apparatus for domain ontology simulation will be described in the series of the articles. This article is the first one of the series. The paper is devoted to means for representation of domain models and domain ontology models, so here a logical language is used only as a means for formalizing ideas. The chief requirement to such a language is that it must have such a semantic basis that would allow us to determine the most exact approximation of a set of intended interpretation functions as often as possible. Another requirement closely connected with the foregoing one is that the awkwardness of expressing ideas in such a language must not considerably exceed the complexity of their expressing in natural language. There are two ways to meet the requirements. The first one is to define and fix a wide semantic basis of the language. In this case the semantic basis nonetheless can be insufficient for some applications of the language. Extending applications of the language can lead from time to time to the necessity of further extending its semantic basis, i.e. to the

[1] This paper was made according to the program of fundamental scientific research of the Presidium of the Russian Academy of Sciences «Mathematical simulation and intellectual systems», the project "Theoretical foundation of the intellectual systems based on ontologies for intellectual support of scientific researches".