PARONYMS FOR ACCELERATED CORRECTION OF SEMANTIC ERRORS *

I. A. Bolshakov, A. Gelbukh

***Abstract**: The errors usually made by authors during text preparation are classified. The notion of semantic errors is elaborated, and malapropisms are pointed among them as “similar” to the intended word but essentially distorting the meaning of the text. For whatever method of malapropism correction, we propose to beforehand compile dictionaries of paronyms, i.e. of words similar to each other in letters, sounds or morphs. The proposed classification of errors and paronyms is illustrated by English and Russian examples being valid for many languages. Specific dictionaries of literal and morphemic paronyms are compiled for Russian. It is shown that literal paronyms drastically cut down (up to 360 times) the search of correction candidates, while morphemic paronyms permit to correct errors not studied so far and characteristic for foreigners.*

***Keywords:** error correction, correction candidates, semantic errors, malapropisms, paronyms, literal paronyms, morphemic paronyms, paronymy dictionaries.*

Introduction

Various errors made by authors in theirs natural language texts can be categorized as follows:

- Orthographic errors transform a correct word into a senseless letter string, e.g., *interesting vook* (instead of *book*);
- Syntactic errors transform one real word to another, thus violating syntactic correctness of the texts concerning agreement of adjectives with their ruling nouns in gender and/or number in Slavic or Romance languages, e.g. Rus. *маленький мальчики* ‘little$_{SG}$ boys$_{PL}$’ instead of *маленькие*$_{PL}$; grammatical cases of the valence dependent noun in Slavic languages (Rus. *довольный правительству*$_{DAT}$ lit. ‘content to the government’ instead of *правительством*$_{INS}$), personal verb forms (*he go* for *goes*) (SG, PL are singular and plural; DAT, INS are dative and instrumental case), etc.
- Semantic errors leave the text orthographically and syntactically faultless, but make it senseless or absurd (*inculpation period* for *incubation period, massy migration* for *massive migration,* etc.).

All modern text editors have tools for error detection. Purely orthographic errors are detected always, and lists of potential correction candidates are given out similar to the suspicious string in letters and/or sounds. Grammatical errors are not always detectable because of deficiencies of modern syntactic analyzers, and variants of syntax corrections are rare so far. Semantic errors are not detected at all.

Meanwhile, methods are already proposed of how to correct one type of semantic errors. For this type, one real word is replaces by another “similar” to the intended one in literal or sound content. If such errors violate semantic correctness of texts, they are referred to as malapropisms.

In [Hirst & St-Onge, 1998; Hirst & Budanitsky, 1998] detection and correction of malapropisms use paradigmatic semantic links between words occurring in adjacent paragraphs and sentences. These are links between direct word repetitions, a word and its hyperonym (*appliance* Vs. *vacuum cleaner*), a part and the whole (*steering wheel* Vs. *car*), etc. For several languages, the links are recorded in thesauri, among which EuroWordNet is well known [Vossen, 2000]. For example, the replacement of *wheel* by *weal* semantically isolates *weal* from words *car*, *brakes* or *gas* within a text.

In [Bolshakov, 2002] malapropism processing uses syntagmatic links between words in a sentence. Malapropisms destroy stable syntactically linked and semantically admissible combinations of content words (=collocations). E.g., *massive migration* is collocation whereas the syntactically correct *massy migration* is not; cf. [Bolshakov & Gelbukh, 2001, 2002]. Thus malapropisms make some content word(s) in a sentence semantically isolated concerning collocations.

For any method of malapropism detection, a generator of correction candidates is necessary. They should be somehow “similar” to the intended words. Such generation is analogous to candidate search for orthographical errors but it differs in the rational search strategy.

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

Indeed, word forms of natural language are rare interspersions in the space of literal strings. For approximate evaluation of this rarefaction, let us take into account that in such highly inflectional language as Russian there exist ca 1.2 million of different word forms, whereas in low inflectional English, say, four times less. To calculate the number of all possible strings above a given alphabet of *A* letters, suppose their length equal to the mean length *L* of real words in a corresponding dictionary. Then the total string number equals A^L, i.e. $32^9 \approx 3.5*10^{13}$ in Russian and $26^8 \approx 2.1*10^{11}$ in English. This means that in Russian a real word form contrasts 29 millions of senseless strings, while in English, contrasts 700,000. The change of the mean length of word form in a dictionary to the mean textual value decreases this contrast, still leaving it striking.

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

Just this circumstance facilitates the candidate search for replacement of one real word by another. Whereas for an orthographical error a wrong string can be arbitrary and the task to gather beforehand, for each string, literally close real words seems impractical, the environments of a real word form, as our evaluations show, contains only few other real words. Hence, the close words can be gathered for each word that has them. Being put in a special dictionary, they could be used for malapropism correction, to cut down the search of candidates. Indeed, for correction of one-letter error in a string with the length *L*, it is necessary $A(2L+1)+L-1$ tries, that for a word of nine letters equals 616. For two-letter errors already ca. 360,000 tries are necessary. In the same time, forehand gathered one-letter-apart candidates are numbered few units, for two-letter-apart ones, numbered few tens. For words that are not in the dictionary of substitutes, the candidate search is unneeded, and this also cuts the search.

This work has the objective to classify semantic errors in some detail and to propose for malapropism correction dictionaries of paronyms, i.e. of words similar to each other in some specific sense. Paronyms can be introduced of the following intersecting types:

- ***Literal paronyms*** [Гусев & Саломатина, 2000, 2001] differ in few letters, so they are within easy distance in the space of letter strings, e.g., Rus. *ожижать* ‘to liquidize’ Vs. *ожидать* ‘to wait,’ *рок* ‘doom/rock’ Vs. *срок* ‘period.’ They are intended for correcting errors characteristic for careless and/or poorly literate persons.
- ***Sound paronyms*** differ in few sounds, so they are within easy distance in the space of phonological records of speech, e.g., Rus. *проектировать* ‘to design’ Vs. *проецировать* ‘to project’). They are indispensable for poorly literate persons.
- ***Morphemic paronyms***, known in Russian lexicography as paronyms proper [Бельчиков & Панюшева, 1994], have the same radix, pertain to the same part of speech (POS), and differ only in prefixes and/or suffixes. E.g., *sens-ible* Vs. *sens-itive* differ in one suffix; *re-volu-tion Vs. in-volu-tion,* in one prefix; *sens-ation-al* Vs. *sens-itive,* in two suffixes. Such paronyms can be close in the space of strings of morphemic symbols. They are important for poorly educated native speakers and for foreigners.

This work reports on compiling Russian dictionaries of one-letter and morphemic paronyms. The dictionaries’ fragments and general statistical parameters are given. Literal paronyms cut the search trials by approximately 360 times, while morphemic paronyms permits to quickly detect the errors not yet discussed anywhere but really occurring in texts and speech.

Sources of semantic errors and their effect

Let us classify semantic errors against their sources, giving minimal contexts.

1. Random error directly giving a real word. This could occur by the following reasons:
    - A writing slip immediately gives another real word, e.g. Rus. *испытательный рок* ‘trial rock’ instead of *срок* ‘period.’
    - A slip gives senseless string that is falsely “corrected” based on a spellchecker menu, since the author took incorrect candidate among those proposed by text editor. If we enter Rus. *испытательный* ***мрок***(?), the menu of spellchecker will contain for the highlighted string the items *мирок, мрак, прок, рок, срок*, *урок,* along with some non-nouns*,* and the careless author can select a wrong item.
    - A correct but very rare word is entered, for which the spellchecker contains one or more alternatives. For example, in the sentence *Ethology of these animals is not studied* spellchecker will propose to

  replace *ethology* by the more known *etiology* or *ethnology,* and the author can hastily accept such corrections.
  - An entered rare word is automatically corrected by a special utility of automatic correction embedded in the text editor. In this case the user transfers a power to the software to make some amendments without any consultations.

2. Ignorance or imprecise knowledge of the intended word, so that instead of it a different word is entered similar to the intended one in sound, e.g. *scientific hypotenuse* instead of *hypothesis*.
3. Imprecise knowledge of meaning for words with the same radix (which really can have the same semantic components), e.g. *sensual news* instead of *sensational news.*
4. Wrong facts or incorrect logic of reasoning transferred in the text. This rarely implies an error in one word, and if so, the resulting word frequently differs from the correct one: *His mother died in* ***infancy*** (for *youth*?); *Hendel was* ***half*** (for *partially*?) *German,* ***half*** *Italian, and* ***half*** *English*. Every human (but not a computer) knows that if a female died in infancy she had no children; that no dividable entity can have three halves, etc.

Hereafter, we deal with the errors of the types 1 to 3. In contrast to errors of the type 4, they violate purely linguistic knowledge on how to commonly use words within the same text. The textual word proved to be:

- similar to the intended one in letters, sounds or morphs,
- preserving syntactic correctness of the utterance, and
- essentially deforming its meaning.

Just such errors are called malapropisms [Encyclopædia, 1998]. Linguistic knowledge is violated by them in the aspects of :

- Syntagmatic semantic links in the texts. The resulted word combinations are not collocations but are syntactically correct. The examples were given above (except of p. 4). More examples are: *polling* ***company*** (for *campaign*); ***hysterical*** (for *historical*) *center;* ***dielectric*** (for *dialectic*) *materialism*; *travel about the* ***word*** (for *world*); *equal* ***excess*** (for *access*) *to school.*
- Paradigmatic semantic links in the texts. Here is an example fit for a single sentence: *Total* ***garniture*** (for *furniture*) *was ruined*: *tables, chairs, armchairs*. Tables, chairs, and armchairs really are related to furniture (not of garniture!), and this is also linguistic knowledge: interrelation of parts and the whole. However, furniture never form collocations with tables, chairs, and armchairs.

The task of candidate search is the same for both type of violation of linguistic knowledge.

Literal paronyms

One literal string of the length *L* can be formed from any other with the series of editing operations [Kashyap & Oomen, 1981; Mays *et al.*, 1992; Wagner & Fisher, 1974]. Let us take strings under an alphabet of *A* letters. Elementary editing operations are: replacement of a letter with any other letter in any place within source string [giving $(A–1)L$ options]; omission of a letter [*L* options]; insertion of a letter [$A(L+1)$ options]; permutation of two adjacent letters [$L–1$ options].

The string obtained with any of $A(2L+1)+L–1$ operations mentioned, is at the distance 1 from the source string, i.e. on the sphere of radius 1 in the string space. Making another elementary step off, we form a string on the sphere with radius 2—with regard to the source one, etc. Points obtained with minimum *R* steps are on *R*-sphere, points of *r*-spheres with $r < R$ and the source point are not here. Among previous examples,

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

Though the mean distance between word forms is large in any language, they proved to be disposed in clusters. Firstly, such clusters contain elements of morphological paradigms of various lexemes, word forms within them being usually distanced 0 to 3 from each other. Just such a cluster is lexeme, and one of the composing forms is its dictionary name. Secondly, paradigms of various lexemes with similar morphs can be close to each other, sometimes even with intersection.

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

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

$$\lambda(\chi) : \lambda(\chi_0) = R_i(\lambda(\chi)) : R_i(\lambda(\chi_0)). \quad (1)$$

The formula (1) means that, for any suspicious form $\lambda(\chi)$ in text, it is necessary to find its dictionary form $\lambda(\chi_0)$, and, if a close parallel $R_i(\lambda(\chi_0))$ for it exists, $R_i(\lambda(\chi))$ should be tried as a correction candidate. For such try, the syntactic correctness pertains as a rule, and some try can correct the error.

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

For example, noun lexemes of European languages have grammemes of singular and plural. They play the same role in a sentence but differ in the sets of collocations they can be in. The division by grammatical number permits to describe easier Slavic declension and well serves for our purposes. E.g., the subparadigms of singular for Russian *метель* ‘blizzard’ and *мотель* ‘motel’ are not parallel, whereas they do—in plural.

Russian verbs have grammemes of personal forms (we join the infinitive to them), of active and passive participles in all grammatical cases, and of gerund. These grammemes differ in their role in a sentence, so that their separate use keeps syntactic correctness of text after the substitution. It is also reasonable to divide each Slavic verb grammeme to its perfect and imperfect aspects, morphonologically rather different.

Each grammeme has its own dictionary name, e.g., a participle is represented by the singular form of nominative case. For the dictionary names and specific forms, the formula (1) pertains. Note that it is not obligatory to require strict parallelism within whole grammemes. E.g., formula (1) applied to Rus. *метры* ‘meters’ и *меры* ‘measures’ fails in genitive case. However such failed tries are not too burdensome.

The idea to divide morpho-paradigms into grammemes is not taken at random. The CrossLexica system elaborated by authors [Bolshakov & Gelbukh, 2001] operates just with grammemes, and paronyms dictionaries under questions are oriented primarily to systems of this kind.

Let us call ***literal paronyms*** any two grammemes that:

- are of the same part of speech;
- concern to the same grammeme type, e.g., both are participles;
- have (close) parallel forms; and, only for nouns,
- have the same gender in singular or are both plural.

With such definition, we have searched close parallel literal paronyms among rather frequent content Russian words. The pairs with at least one member being functional word (pronoun, preposition, conjunctions, etc.) were omitted. A large preliminary version of dictionary was compiled first, and then a special utility proofreads this version for repetitions, omission of inverted pairs, larger distances, wrong orders, etc. Note that in [Гусев & Саломатина, 2000, 2001] the same task has been performed for lexeme names, thus giving less information (see above).

In the current version, there are more than 6,000 paronym groups each having a item-head grammeme to be replaced and the rest grammemes as substitute candidates. The mean number of candidates stably equals 2.25, while the mean name length is 6.75.

Functional words, the shortest in any language, were excluded. Nevertheless, the mean word length in our dictionary proved to be two letters shorter than the mean dictionary length value. So grammemes in our dictionary are seemingly the shortest among the content words, and probably the most frequent among them.

Below, we give a fragment of our dictionary. Note that homonyms like *болеть*1 ‘to be ill’ Vs. *болеть*2 ‘to ache’ or *белки*1 ‘squirrels’ Vs. *белки*2 ‘proteins’ enter separately, but the group for one of them does not include the others. The number of candidates varies from 1 to 12. The maximum number is for the shortest words, i.e. of three letters.

```
бездомный
  бездонный
  бездумный
бездумный
  бездомный
  безумный
безумный
  бездумный
бекон
  бетон
белеть
  белить
  болеть1
  болеть2
  велеть
  мелеть
белея
  болея
  мелея
белить
  белить
  делить
белка
  булка
  елка
  челка
  щелка
белки1
  балки
  бели
  булки
  челки
  щелки
белки2
  балки
  бели
  булки
  челки
  щелки
```

The main gain in candidate search is reached thanks to looking up only the candidates given in our dictionary. Using the total number of tries for a 9-letter Russian word, we get the gain coefficient G_1 = 616/2.25 = 274.

ClossLexica contains ca. 100,000 one-word grammemes. Even if after further replenishments the total number of groups would reach 6500, this will be only 6.5% of the whole systemic dictionary. Nevertheless, the revealed paronyms are supposedly the most frequent among content words. With the reasonable assumption that the rank distribution of all words in systemic dictionary conforms to Zipf law, these paronyms cover approximately 80% of all word occurrences in texts, and we have the additional gain coefficient G_2 = ln 100000 / ln 6500 = 1.31 owing to that all other 93,500 word are ignored in the candidate search. The global gain is $G_1 \times G_2 \approx 360$.

Morphemic paronyms

Several errors of a different nature were demonstrated above: *massy* Vs. *massive, sensible* Vs. *sensitive, revolution* Vs. *involution*. They are of the same POS and have the same radix (*mass-, sens-, -volu-*). In Russian linguistics, only this similarity is called paronymy. Confusions of morphemic paronyms are usual errors, especially for foreigners. For example, it is rather difficult to explain to them how to use Rus. paronyms *вислый* 'slouching', *висящий* 'hanging,' *висячий* 'bangled,' and *повисший* 'flagging' that differ only in one suffix and one prefix.

We have gathered morphemic paronyms into groups with the following additional requisites:

- Grammemes are taken as units of the dictionary, so that, e.g., *бок* 'side' and *бока* 'sides' are put into the same group;
- Grammemes of participles are considered as adjectives;
- All grammemes with homonymous radixes are pu to the same groups, e.g., adjectives *бур-ный* 'roaring,' *бур-овой* 'boring,' and *бур-ый* 'brown';
- Homonymous lexemes are given in the groups separately, however none of them can replace another;
- Two-root words are involved, one radix considered as the radix proper and another as the so-called suffixoid or a prefixoid. The negation *не* is a common prefix, the inseparable reflexive particle *–ся* is considered as suffix after the ending.

All in all, a morphemic paronym can be represented as a string $P_1...P_mRS_1...S_nE,$ where $P_1,...,P_m$, m = 0, 1..., are symbols of prefixes; R is radix: $S_1,...,S_n$, n = 0, 1..., are suffixes; E is ending. The distance between paronyms within a group is measured by the number of elementary editing operations in the space of morphemic symbol strings. For example, *+отеч-еств*о* 'homeland' Vs. *отч-еств*о* 'patronym' and *+бед*а* Vs. *+бед*ы* are at the distance 0, *+волос-ат*ый* Vs. *волос-ист*ый*; *вы-нос* Vs. *из-нос*; *эффект-ив-н-ост*ь* Vs *эффект-н-ост*ь* are at the distance 1, *юнош-еск*ий* Vs. *юн*ый*, *гриб-н*ой* Vs. *гриб-к-ов*ый* are at the distance 2. Here the sign '+' initiates a radix; '–' a prefix or a suffix, '*' an ending. The differences in endings are ignored, since inflexional class is implied by POS and the previous suffix, and specific ending is different for each element of a grammeme.

Our dictionary of morphemic paronyms contains now 1120 paronymy groups with the mean length 5.65. A group element has on an average 1.4 paronyms at the distance 0 or 1. Summarize 'all-to-all' links in all groups at any distances, the total link number is up to 55,000, i.e. approximately 49 links within each group. Following is a fragment of the morphemic dictionary:

```
+бег*
  +бег*а
  +бег-л-ост*ь
  +бег-ств*о
  +бег-ун*
  +бег-ун-ок*
  +бег-ун*ья
  -на+бег*
  -при+беж-щ*е
  -про+бег*
  -про+беж-к*а
  -раз+бег*
  -у+беж-ищ*е
+бег-ающ*ий
  +бег-л*ый
  +бег-ов*ой
  +бег-ущ*ий
  -при+бег+ающ*ий
  -при+бег+ну-вш*ий
  -раз+беж-авш*ий-ся
  -с+бег-ающ*ий
  -с+беж-авш*ий
  -у+бег-ающ*ий
  -у+беж-авш*ий
+бед*а
  +бед-н-ост*ь
  +бед-н-от*а
  +бед-ств-енн-ост*ь
  +бед-ств-и*е
  +бед-ств-и*я
  +бед-ств-ован-и*е
  +бед*ы
  -о+бед-н-ени*е
+бед-н*еть
  +бед-овать
  +бед-ств*овать
  -о+бед-н*еть
+бед-н-еющ*ий
  +бед-н*ый
  +бед-ов*ый
  +бед-ств-енн*ый
  +бед-ств-ующ*ий
  -о+бед-н-евш*ый
  -о+бед-н-енн*ый
```

The search of morphemic errors is cut down by the same ways as for literal errors. If the suspicious word is in the dictionary, only its co-members are taken at the distant 0 or 1 to match. If the textual word is not available in the dictionary, no candidate of morphemic type is searched. We cannot compare our method with others quantitatively, since the latter do not exist. Indeed, the letter distance between morphemic paronyms is usually so high that their direct search in the literal space is absolutely impractical.

Conclusion

It is argued that correction of some semantic errors (namely, malapropisms) is possible by the use of paronyms, i.e. of words similar to each other in letters, sounds or morphs. It is proposed to compile paronymy dictionaries of three types beforehand. Literal paronyms essentially cut the search of correction candidates. Morphemic paronyms permit to quickly correct errors not studied so far and specific for foreigners. Russian dictionaries are already created—for literal and morphemic paronyms. The compiling of sound paronyms is the task for the future.

Bibliography

[Бельчиков & Панюшева, 1994] Бельчиков, Ю. А., М. С. Панюшева. Словарь паронимов современного русского языка. М.: Русский Язык, 1994.

[Bolshakov & Gelbukh, 2001] 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 No. 1959, Springer, 2001, p. 103-114.

[Bolshakov & Gelbukh, 2002] Bolshakov, I. A., A. Gelbukh. Word Combinations as an Important Part of Modern Electronic Dictionaries. Procesamiento del Lenguaje Natural (Spain), No. 29, Sept. 2002, p. 47-54.

[Bolshakov, 2002] Bolshakov, I. A. Detección y Corrección de Malapropismos en Español mediante un Sistema Bietapa para Comprobar Colocaciones. Memorias del XI Congreso Internacional de Computación "Avances en Ciencias de la Computación e Ingeniería de Cómputo" CIC'2002, noviembre 2002, CIC-IPN, México, v. II, p. 303-313.

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

[Гусев & Саломатина, 2000] Гусев, В. Д., Н. В. Саломатина. Электронный словарь паронимов: версия 1. Научно-Техническая Информация (НТИ), Сер. 2, № 6, 2000, с. 34-41.

[Гусев & Саломатина, 2001] Гусев, В. Д., Н. В. Саломатина. Электронный словарь паронимов: версия 2. Научно-Техническая Информация (НТИ), Сер. 2, № 7, 2001, с. 26-33.

[Hirst & St-Onge, 1998] 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.

[Hirst & Budanitsky, 1998] Hirst, G., A. Budanitsky. Correcting Real-Word Spelling Errors by Restoring Lexical Cohesion. Computational Linguistics (to be published).

[Kashyap & Oomen, 1981] Kashyap, R.L., B.I. Oomen. An effective algorithm for string correction using generalized edit distances. I. Description of the algorithm and its optimality. Information Science, 1981, Vol. 23, No. 2, p. 123-142.

[Mays et al., 1992] Mays, E., F.J. Damerau, R.L. Mercer. Context-based spelling correction. Information Processing and Management. 1992, Vol. 27, No. 5, p. 517-522.

[Vossen, 2000] Vossen, P. (ed.). 2000. EuroWordNet General Document. Vers. 3 final. www.hum.uva.nl/~ewn.

[Wagner & Fisher, 1974] Wagner, R.A., M. J. Fisher. The string-to-string correction problem. J. ACM, Vol. 21, No. 1, 1974, p. 168-173.

Author information

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

Alexander Gelbukh – CIC-IPN, Research Professor and Chung-Ang University, Visiting Professor; Center for Computing Research (CIC), National Polytechnic Institute (IPN), Av. Juan Dios Bátiz s/n esq. Av. Miguel Othon Mendizabal, Unidad Profesional “Adolfo Lopez Mateos”, Col. Zacatenco, C.P. 07738, D.F., Mexico; e-mail: gelbukh@cic.ipn.mx, Internet: www.gelbukh.com

TOWARDS COMPUTER-AIDED EDITING OF SCIENTIFIC AND TECHNICAL TEXTS

E. I. Bolshakova

***Abstract**: The paper discusses facilities of computer systems for editing scientific and technical texts, which partially automate functions of human editor and thus help the writer to improve text quality. Two experimental systems LINAR and CONUT developed in 90s to control the quality of Russian scientific and technical texts are briefly described; and general principles for designing more powerful editing systems are pointed out. Features of an editing system being now under development are outlined, primarily the underlying linguistic knowledge base and procedures controlling the text.*

***Keywords:** scientific and technical texts, automatic editing, linguistic knowledge base.*

Introduction

Scientific and technical writing is by no means easy, even for skilled and experienced authors. Usually, the elaboration of a good scientific or technical (sci-tech) text is iterative and time-consuming process, with several persons taking part in it. Besides an author of the document, colleagues, reviewers, and an editor participate in the process, helping the author to improve the text.

Scientific papers and technical documents are essential means of communication between scientists and engineers; therefore the efficacy of the communication depends on the quality of texts. A professional editor of sci-tech texts not only looks for grammar and spelling mistakes, but also accomplishes editing specific for functional style of scientific and technical prose: controlling word usage, revealing drawbacks in logic of reasoning, judging text organization, etc. [10]. The editor explains revealed defects and drawbacks, as well as proposes possible ways of how to overcome them, thereby helping the author to improve the text and to enhance its stylistic uniformity. Almost all sci-tech writers need some aid of professional editor, and without it they lack computer systems automating certain editor functions.

Of course, well-known universal computer text editors and spellers (e.g., MS Word) are widely used for preparing texts. These systems reveal many mistakes, including spelling and simple syntactic mistakes, and their facilities are permanently extended. But the universality of these systems means that they do not account for specificity of the particular text style and genre, in particular, sci-tech prose with its intensive usage of