Synonymous Paraphrasing Using WordNet and Internet*

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

[1] Center for Computing Research, National Polytechnic Institute, 07738, Mexico
{igor,gelbukh}@cic.ipn.mx; www.gelbukh.com

[2] Department of Computer Science and Engineering,
Chung-Ang University, Seoul, 156-756, Korea

Abstract. We propose a method of synonymous paraphrasing of a text based on WordNet synonymy data and Internet statistics of stable word combinations (collocations). Given a text, we look for words or expressions in it for which WordNet provides synonyms, and substitute them with such synonyms only if the latter form valid collocations with the surrounding words according to the statistics gathered from Internet. We present two important applications of such synonymous paraphrasing: (1) style-checking and correction: automatic evaluation and computer-aided improvement of writing style with regard to various aspects (increasing *vs.* decreasing synonymous variation, conformistic *vs.* individualistic selection of synonyms, etc.) and (2) steganography: hiding of additional information in the text by special selection of synonyms. A basic interactive algorithm of style improvement is outlined and an example of its application to editing of newswire text fragment in English is traced. Algorithms of style evaluation and information hiding are also proposed.

1 Introduction

Synonymous paraphrasing (SP) is such change of natural language (NL) text or of its fragments that preserves the meaning of the text as a whole. Nearly every plain text admits SP (in contrast to lists of names, numerical data, poetry, and the like). Computational linguistics has always considered SP an important and difficult problem. The ability of a system to generate good synonymous variations was even considered an indicator of "comprehension" of natural language by computer. Currently, most important applications of SP are text generation and computer-aided style improvement.

There exists a well developed linguistic theory—Meaning–Text Theory by I. Mel'čuk [9]—that takes SP as one of its basic principles, considering NL as something like a calculus of synonymous paraphrasings. A set of meaning-conserving rules for restructuring of sentences was developed in frame of MTT. In the process of paraphrasing both words and word order significantly change. The changes in words can touch upon their part of speech and number, for example, *to help tremendously* vs. *to*

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

give tremendous help. However, existing so far special dictionaries and software for paraphrasing based on MTT [1] cover a rather limited fragment of natural language.

Without full-fledged realization of a comprehensive theory of paraphrasing, we nevertheless already possess large linguistic resources—WordNet [8] (EuroWordNet [12]) and Internet—that can help resolving the problem of local paraphrasing with acceptable performance. By local paraphrasing we mean those SP techniques that conserve the structure and word order of a sentence, as well as the number of words (counting stable multiword expressions—or multiwords—like *hot dog* as one unit).

SP is especially important for English—lingua franca of modern sci-tech world. Fortunately, just for English the mentioned resources are highly developed.

In this paper we propose a method of local SP of NL texts based on WordNet synonymy information (synsets) and Internet-based statistics on stable word combinations (collocations) the members of a synset are in. To paraphrase a text, we look for words or multiwords in it that are members of a WordNet synset, and substitute them with other members of the same synset only if they are feasible components of collocations with the surrounding words according to statistical evaluation through the an Internet search engine, such as Google.

More specifically, the objectives of our paper are:

- To touch upon the notion of synonymy in order to make it clear that we consider as synonyms not only separate words but also multiwords and that we divide all synonyms into absolute and non-absolute ones, which are used in a different manner for the purposes o SP;
- To describe relevant features of collocations and to explain how Internet statistics can be used to test whether two given small text fragments can form a collocation;
- To enumerate and to formalize various types of SP. Some text authors always use only one, the most frequent, synonym for the given concept—unintentionally or to be intelligible for foreigners, children, etc. Other authors widely use synonymous variation to increase literary merits of their texts. So, at least two options are possible for such variation: conformistic (like others) or individualistic. The use of various abbreviations can also be considered a means for SP: some authors and editors prefer concise style, whereas others prefer more verbose one;
- To outline an algorithm realizing interactive SP;
- To present the results of interactive SP applied to a fragment of a newswire text;
- To describe another application of SP methods: given a text, the algorithm analyzes the author's usage of synonyms in it;
- Finally, to develop yet another, totally new application of SP: steganography, i.e. hiding arbitrarily information in a text by an appropriate choice of synonyms. Namely, each word having synonyms can be replaced by its synonym, depending on the current bit in the bit sequence to be hidden in the text.

2 Absolute and Non-Absolute Synonyms

In a simplest definition, synonyms are words that can replace each other in some class of contexts with insignificant change of the whole text's meaning. The references to "some class" and to "insignificant change" make this definition rather vague, but we are not aware of any significantly stricter definition. Hence the creation of synonymy dictionaries, which are known to be quite large, is rather a matter of art and insight.

A synonymy dictionary consists of groups of words considered synonymous. However, a word can be similar to members of one group in some semantic elements and of another group in other semantic elements. Hence generally speaking a word can belong to more than one synonymy group (if any).

It proved to be insufficient to include to a synonymy groups only separate words: sometimes multiword expressions referring to a given concepts are included. Attempts to translate any dictionary from one language to another always results in the use of such multiwords. For example, the English synonymy group {*rollercoaster, big dipper, Russian mountains*} contain only one single word. Thus we consider multiwords as possible members of synonymy groups.

The only mathematically formal type of linguistic synonymy is when the compared words can replace each other in any context without any change in meaning. These are absolute synonyms, e.g., English {*sofa, settee*}. Absolute synonyms can be formalized as connected by the mathematical relation of equivalence. In the dictionary, such absolute synonyms should be specially marked within their synonymy group.

Note that absolute synonyms are extremely rare in any language. However, there exists much more numerous type of linguistic equivalence—equivalence between various abbreviations and the complete expression. E.g., we can consider as a group of equivalence {*United States of America*, *United States*, *USA*, *US*}. Such equivalents can occur in the same text without any violation of its style. In fact, admission of multiword synonyms brings in a lot of new absolute synonyms like {*former president, ex-president*} or {*comical actor, comic*}.

In many synonymy dictionaries, in each group one member is selected that expresses the common meaning of the words in the group in the most general and neutral way. This, however, is not the case with WordNet [8] and its follower EuroWordNet [12], where all synset members are considered equivalent and corresponding to the common interpretation formula (gloss).

If a word (letter string) enters several synsets, it is always considered homonymous in WordNet, individual homonyms being labeled by different numbers (sense numbers). Homonyms exist in all dictionaries, but in WordNet their quantity is, according to many opinions, exaggerated. In fact, not admitting the same word to enter different synsets does not make all members of each synset absolute synonyms, since there is no guaranty that all of them form the same collocations. Hence, in contrast to absolute synonyms, collocational compatibility with the context should be tested for each member of a synset individually.

As to words commonly recognized as clear homonyms (like $bank_1$ 'shore' vs. $bank_2$ 'organization'), they very rarely enter into the same collocations. So in the task of collocational compatibility of synonyms for a given word, not only all members of its synset but also all members of synsets of its homonyms should be always tested.

Hereafter we assume that a set of synonymy tools is available that includes:

- Synonymy dictionary such as WordNet (or EuroWordNet);
- A specially compiled dictionary of absolute synonyms that contain all abovementioned types of English equivalents. The synsets of such a dictionary can be sub-

sets of corresponding WordNet synsets, which does not cause any problem since our algorithms look up first absolute synonyms.

So far, WordNet contains rather small number of multiwords, but this number grows from version to version. The discussion in [3] shows that the problem of multiword synonym gathering is fully recognized.

3 Collocations

By a *collocation* we mean a syntactically connected and semantically compatible pair of content words, e.g. *full-length dress, well expressed, to briefly expose, to pick up the knife* or *to listen to the radio*; the components of collocations are underlined.

Syntactical connectedness is understood as in dependency grammars [9] and it is in no way merely a co-occurrence of the collocatives in a short span of a text [11]. The head component syntactically governs the dependent component, being adjoined to it directly or through an auxiliary word (usually a preposition). In the linear order of words in the sentence the collocatives can be at any distance from each other, though they are close to each other in the dependency tree.

For a long time collocations were studied in lexicography rather than in computational linguistics. Till now collocations are often treated as series of two or more words occurring together in a narrow window moving along a text [11] or in a specific unit of text [13].

At the same time WordNet includes only semantic links of the paradigmatic type. Their related terms usually include semantically associated components but do not co-occur in close contexts. However, lexicographers have always considered collocations as semantic connections of syntagmatic type with the components usually co-occurring in texts. A comprehensive part of English collocations is now collected in the Oxford Collocations dictionary [10].

To our knowledge, publicly available electronic databases of English collocations did not exist until 1997, when the Advanced Reader's Collocation Searcher (ARCS) for English appeared [2]; however, its deficiencies are too severe for indulgent criticism. The only project in the last decade of a very large collocation database was dedicated to Russian [4]. Thus there is no collocation database for English so far, and though collocation testing could be more easily and reliably done with collocation databases, we have to look for other resources. Just this resource is Internet.

4 Evaluations of Collocation Statistics via Internet

Hence, our goal is to elaborate a mechanism for assessing whether a word can be replaced with its synonym while keeping collocational cohesion of the text, i.e., a means for deciding which synonyms of a given word can form good collocations with a word in the context.

Consider an example. Suppose the modifying adjective *large-scale* and the noun *project* somewhere to the right of it are found in the same sentence. According to the synonymic dictionary, *large-scale* enters into the synset {*colossal, gigantic, grandiose, great, huge, large-scale, tremendous, very large*}. It is necessary to collect statistics in Google on potential collocations that each synonym of the synset could form with the noun *project*.

Google permits collecting statistics only on the number of pages where the two words (or multiwords) co-occur. Only two options of their mutual disposition are measurable: juxtaposition (can be obtained by querying the tested pair in juxtaposition within quotation marks) and arbitrary co-occurrences within a page (queried without quotation marks). The corresponding statistics are given in Table 1.

Table 1. Google statistics of collocations with *project*

Collocation	In quot.	Portion	W/o quot.	Portion	MGV	Portion
colossal project	793	0.5%	123,000	0.5%	9,876	0.5%
gigantic project	2,670	1.7%	255,000	1.0%	26,093	1.3%
grandiose project	1,540	1.0%	83,200	0.3%	11,319	0.5%
great project	80,300	51.6%	9,710,000	38.9%	883,013	44.8%
huge project	34,400	22.1%	4,100,000	16.4%	375,552	19.0%
large-scale project	28,700	18.4%	2,660,000	10.7%	276,300	14.0%
tremendous project	1,620	1.0%	1,340,000	5.4%	46,591	2.3%
very large project	5,570	3.6%	6,690,000	26.8%	193,037	9.8%
Total:	155,593	100.0%	24,961,200	100.0%		

MGV (Mean Geometric Value) in Table 1 is the square radix of the product of numbers obtained in quotation marks and without them. The portion distribution for the collocation set is calculated just for MGVs for the whole set.

As one can see, the ratio between the non-quoted and quoted (sequential co-occurrence, a probable collocation) evaluations is rather big and varies in a broad range. The large ratio values are natural since the non-quoted evaluations count all co-occurrences even at far distance within the page, so that the majority of them do not correspond to collocations of the two components at hand. On the contrary, quoted evaluation corresponds to sequential co-occurrences which probably correspond to collocations. However, not all collocations are counted in this way, since pages with distanced collocations like *great industrial* (*commercial*, *political*, *web*, ...) *project* are not taken into account. Thus the correct number of pages with a given collocation is between the two figures and cannot be measured exactly in this way.

Since only comparative estimations are necessary for our purposes, we evaluate the usage proportions of synonyms within the synset (summing up to 100%) separately for quoted and non-quoted measurements and then take the mean value in each pair of such evaluations. These values are given in the right-hand column of Table 1.

The synonyms with cumulative portion less than a certain threshold μ of marginality are considered unusual in the given collocation and thus not recommended for use in SP. If we take the threshold $\mu = 3\%$, the recommended subset of synonyms in context of *project* is {*gigantic, great, huge, large-scale, very large*}. Just these are the words that we will consider further for various types of the paraphrasing.

Consider now an example where statistics is gathered for a synset's members participating each one in two different collocations: the phrase *heads of various departments*, the synset to be tested being {*departments, offices, services*}; see Table 2.

Table 2. Google statistics of collocations with synonyms of *departments*

Collocation	In quot.	W/o quot.	MGV	Portion
heads of departments	72,700	989,000	268,142	50%
heads of offices	2,320	1,060,000	49,590	12%
heads of services	2,130	5,030,000	103,508	38%
various departments	287,000	5,440,000	1,249,512	34%
various offices	59,000	5,150,000	551,226	17%
various services	297,000	11,200,000	1,823,842	49%

The portion distributions for various collocation sets (in our example, the first three *vs*. the last three rows in Table 2), are again calculated through MGVs. To combine the data of various sets, we use the mean arithmetic values of the corresponding portions in the different sets. This gives the following distribution:

departments	42%
offices	15%
services	43%

This shows a low portion of *offices*, so this synonym is much less recommendable in this context than the two others (cf. the data of separate collocations). By this composition of tests considered independently, the portion of some synonyms can fall below the marginality threshold.

5 Various Types of Paraphrasing

Paraphrasing can have various objectives. Having in mind the fist example in the previous section as illustration, we can classify its types as follows.

Text compression For this, the shortest synonym is taken in each synset (either independently of any statistical evaluations or selecting from the words that passed the marginality threshold). In our example, this is *huge*. This gives a significant gain in space only when there are abbreviation(s) among absolute synonyms.

Text canonization For this, the most frequently used synonym is taken. Of course, it may prove to be the same one as in the source text. In our example, this is *great*. The text becomes more canonic—or banal, without variations. It is useful from the viewpoint, say, of legislative bodies, since in the legislation even common words can be considered strict terms. It is also useful for persons with limited language knowledge, i.e. for foreigners or children, since this renders texts in a more intelligible way.

Text simplification Any text will be more intelligible for language-impaired person if we select among synonyms a "simpler," colloquial variant [5]. It is not always the most frequently used synonym, though in our example this is probably also *great*. We consider language-impaired persons as native adults with rather low educational level whose language abilities scarcely could be improved.

The algorithm of synonymous paraphrasing for the simplification is roughly as follows. If for a given word there are any synonyms marked in the dictionary as *colloquial*, we select the most frequent one of them. Otherwise, if there are any neutral synonyms (without any stylistic mark), we select the shortest one of them, assuming that the shortest is the simplest for average language-impaired person's mentality. In particular, in this way the scientific, bookish or obsolete words will be substituted with colloquial or neutral synonyms.

Conformistic variations For this, the synonym is taken randomly with the distribution corresponding to the frequencies obtained though Internet evaluations. Such a choice fully corresponds to the present usage of the given synset's members.

Individualistic variations We may imagine individualistic (counter-conformistic) variation as selection of the most rarely used option among those exceeding the marginality threshold. Since the value of the threshold is taken on rather subjective grounds, this tactics may be considered risky and sometimes give erroneous results.

6 Basic Algorithm of Interactive Paraphrasing

Below we outline—with significant simplifications, especially as to the conformistic style mode—the interactive SP procedure.

```
1.  Ask $mode \in \{compression, canonization, simplification, conformistic, individualistic\}$
2.  Ask marginality threshold $\mu \in (0,1)$ and sensitivity threshold $\lambda \in (0,1)$
3.  For each non-functional word (or multiword) $w$ which is a member of a synset
4.      Let $S$ = union of all relevant synsets $s$ for $w$
5.      For each word $v$ in $S$
6.          If its appropriateness $a(v) < \mu$ then set $score(v) = 0$
7.          Else
8.              If $mode = compression$      then set $score(v) = 1 / length(v)$
9.              If $mode = canonization$     then set $score(v) = a(v)$
10.             If $mode = simplification$   then set $score(v)$ as described in Section 5
11.             If $mode = conformistic$     then set $score(v) = random$ from 0 to $a(v)$
12.             If $mode = individualistic$  then set $score(v) = 1 / a(v)$
13.     If $score(w) / \max_S score(v) < \lambda$ then
14.         Suggest to the user all variants $v$ in $S$, $score(v) \neq 0$, in the ordered of $score(v)$
```

By relevant synsets in line 4 we refer to a word sense disambiguation procedure if it is available; otherwise all senses are considered relevant. Since we cannot distinguish between (relevant) senses, we have to consider all members of all such synsets to be equally possible synonyms of the given word, hence the union; however, the synonyms of wrong meanings are unlikely to pass the marginality threshold in line 6.

Appropriateness is determined as described in Section 4. If syntactic heuristics selecting possible dependency links for a given word are available, the context words to try the collocations with are selected accordingly. Otherwise, all non-function words in the same sentence within certain linear distance from the given word are used.

The condition in line 13 is needed to force the algorithm to suggest corrections only where they are really necessary and not at every word.

After the work has been finished, the user can assess the result as described in Section 8 and compare the obtained score with that of the optimal transformation consisting in automatically accepting the variant with the highest score for each word.

7 An Experiment on Text Fragment Paraphrasing

For a real experiment with SP, an English fragment from a Russian newswire site Gazeta.ru was taken. Our initial assumption was that the translators from Russian to English from the Russian news agencies are not as skilled in the language as their native English-speaking colleagues, so that the results of paraphrasing might be of practical interest. The fragment taken was as follows:

> *The Georgian foreign_minister (foreign_office_head) is scheduled (planned, designed, mapped out, projected, laid on, schemed) to meet (have a meeting, rendezvous) with the heads (chiefs, top_executives) of various (different, diverse) Russian departments (offices, services) and with a deputy of Russian foreign_minister (foreign_office_head). "Issues (problems, questions, items) concerning (pertaining, touching, regarding) the future (coming, prospective) contacts at the higher (high-rank) level will be discussed (considered, debated, parleyed, ventilated, reasoned, negotiated, talked about) in the course of the meeting (receptions, buzz sessions, interviews)," said Georgian ambassador to Russia Zurab Abashidze. The Georgian foreign_minister (foreign_office_head) will be_in (visit) Moscow on a private (privy) visit (trip), the Russian Foreign Ministry reported (communicated, informed, conveyed, announced).*

Let us discussed the transformations listed in Section 5 as applied to this text.

Text compression The potential improvements are: *scheduled* → *planned*, *departments* → *offices*, *issues* → *items*, *concerning* → *touching*, *discussed* → *debated*, *private* → *privy*, *visit* → *trip*. Not all of them would pass the statistical test. For example, a combination *foreign minister is scheduled* is 60 times more frequent than *foreign minister is planned*.

Text canonization Our tests showed that a few words can be changed in the text: (*will*) *be_in* → *visit* (*Moscow*), 3 times more frequent; (*Ministry*) *reported* → *announced*, 1% more frequent. On the other hand, in most cases the translator has chosen the correct synonym. For example, *issues* is 3 times more frequent with *concerning* than *problems*; *future* 20 times more frequent than *prospective* with *contacts*; *visit* 13 times more frequent than *trip* with *visit*. Thus, the overall quality of translation in the text under consideration can be assessed as quite good.

Text simplification Here, the first candidates to substitution are the words having colloquial variants: *discussed* → *talked about* and *meetings* → *buzz sessions*. The other substitutions are the same as for text compression.

Conformistic variations Here is a possible variant of such SP:

> *The Georgian foreign_office_head is planned to have a meeting with the heads of diverse Russian offices and with a deputy of Russian foreign_office_head. "Questions touching the future contacts at the high-rank level will be debated in the course of the interviews," said Georgian ambassador to Russia Zurab Abashidze. The Georgian foreign_minister will visit Moscow on a private trip, the Russian Foreign Ministry informed.*

Individualistic variations Here is a possible variant of this type of SP:

> *The Georgian foreign_office_head is projected to rendezvous with the top_executives of diverse Russian departments and with a deputy of Russian foreign_office_head. "Issues*

regarding the prospective contacts at the high-rank level will be parleyed in the course of the receptions," said Georgian ambassador to Russia Zurab Abashidze. The Georgian foreign_office_head will visit Moscow on a privy visit, the Russian Foreign Ministry conveyed.

8 Another Application: Style Evaluation

The most usual way to evaluate the style of the text is currently through easily gathered statistics of word length in letters, sentence length in words, and paragraph length in sentences. This is too formalistic to give good results.

Meantime, the use of synonyms can evidently estimate an important aspect of the literary style. For example, repeated use of the same synonym for the given notion makes the text banal and dull, though maybe good technical writing. Diverse but conformistic use of synonyms considered by many a good literary style, but some journalists prefer counter-conformism (cf. Section 6).

So we suppose that a user of an automatic style checker wants to obtain the evaluation in points that assess four characteristics: compressibility, variation, conformism, and individualism.

The algorithm for assessing compressibility works (with some simplifications) as follows.

```
Set $\mathit{Compressibility}$ to 0
For each non-functional word $w$ in the text
    Set $S$ = union of all relevant synsets containing $w$
    Remove from $S$ the members $v$ with appropriateness $a(v) < \mu$
    Let $v_0$ be the shortest word in $S$
    Increase $\mathit{Compressibility}$ in $\mathit{length}(w) - \mathit{length}(v_0)$
```

Again, by relevant synsets we refer to a word sense disambiguation procedure if it is available; otherwise, all synsets are considered relevant. By appropriateness we refer to the procedure discussed in Section 4, where μ is the marginality threshold.

For measuring variation, conformism, and individualism, we need to compare the usage statistics g in Internet and f in the given text, for each word used in the text.

```
Consider only synsets relevant for at least one non-functional word in the text
For all words $w$ in all synsets $s$
    Set $g_s(w) = 0$ (Google statistics)
    Set $f_s(w)$ = the number of occurrences of $w$ for which $s$ is relevant
        or $f_s(w) = 1$ if no occurrences (for smoothing)
    Set $\varphi_s(w) = 1 / f_s(w)$ (inverse frequency, for individualism)
For each occurrence of a word $w$
    For each synset $s$ relevant or it
        For each member of $s$
            Increase $g_s(w)$ in the frequency obtained from Internet
For each synset s
    Normalize $g_s, f_s, \varphi_s$ within $s$ so that $\sum_w g_s(w) = \sum_w f_s(w) = \sum_w \varphi_s(w) = 1$
Set $\mathit{variation}$ to average dispersion of $f_s(w)$ within synsets
Set $\mathit{conformism}$ to average cosine similarity between $g_s$ and $f_s$
Set $\mathit{individualism}$ to average cosine similarity between $g_s$ and $\varphi_s$
```

By Internet statistics in line 10 we again mean the procedure described in Section 4, which depends on the context of each specific occurrence of a word, which we implicitly average across all occurrences.

The above procedure generates the absolute value of the corresponding characteristic. What is more interesting for the user, however, is the relative value: is the text optimal or can be significantly improved? This can be assessed as the ratio between the absolute score obtained for the given text and that of a text optimized as described in Section 6 by automatically choosing the best variant at each text position.

9 Yet Another Application: Linguistic Steganography

Linguistic steganography [6, 7] is a set of methods and techniques permitting to hide within a text any binary information, based on some purely linguistic knowledge. For hiding the very fact of enciphering, the resulting text should not only remain innocuous but also conserve grammatical correctness and semantic cohesion. For hiding information, some words in the source text are replaced by other words in the way controlled by the bit sequence to be hidden and detectable at the receiver's side. In the best case, the resulting text conserves the meaning of the source one.

Chapman *et al.* [7] proposed to take beforehand a synonymy dictionary and enumerate the words within each its group. When their steganographic algorithm encounters in the text a word from one of these groups, it replaces the word by its synonym having intra-group number equal to the binary content of a small current portion of the information to be hidden. It is clear that while scanning the resulting text, the reverse algorithm will find the representatives of just the same synonymy groups and restore the hidden information basing on their numbers within the groups.

The described idea does work, but context-independent synonymous changes usually do not preserve the meaning. Additionally, the resulting texts become semantically non-cohesive (incomprehensible) and thus can loose their initial innocuity.

We propose to divide the synonyms into two large classes. Synsets of absolute synonyms are used in the same context independent manner. However, the synsets of non-absolute synonyms are previously tested for conforming to collocations in the text the source synonym is in. Only those non-absolute synonyms that pass all collocational tests form the relevant groups are used. The inner numbers within these (usually reduced) groups are taken for steganography.

The proposed steganographic algorithm has two inputs:

- The source natural language text of the minimal length of approximately 200 per bit of the information to be hidden. The text format can be arbitrary, but it should be orthographically correct, to avoid later corrections by someone else. The text should also be semantically "common," since the presence of lists of names, sequences of numbers, and the like increase the text length required for hiding.
- The information to hide, merely as a bit sequence.

The steps of the algorithm are as follows:

Search of synonyms From left to right, (multi)words that are entries of the synonymy dictionary are extracted (in case of ambiguity, the longest multiword is taken).

Formation of synonymy groups The synsets are analyzed one by one. If the synset consists of absolute synonyms, only they are immediately taken and ordered in a predetermined manner (e.g., alphabetically). If this is a synset of non-absolute synonyms, then it is checked whether the textual synonym is homonymous and its homonyms are the members of other synsets. All newly found homonyms are grouped for further collocation proving.

Collocation proving of synonyms All members of non-absolute synsets are checked against their context, group by group. The context words in the text that could form a collocation with members of the tested synset are sent as a query to Google. Each query is sent in two forms, in quotation marks and without them. The statistics obtained is normalized in the manner described in Section 4. Each pair {synonym, context word} is statistically evaluated against Internet as a pair of components of a collocation. If the synonym has several senses, all of them are tested. If the context word is absolutely synonymous or not synonymous, the tests are carried out only with it. Otherwise (if the context word belongs to a group of non-absolute synonyms), the tests are done with all of them. At each step, the synonym under test is excluded from its group if a certain threshold μ is not reached. The synonyms that pass this test are ordered within the reduced synsets in the predetermined manner. All non-functional context words, both to the left and to the right from the current word, are taken within the current sentence.

Enciphering The sequence of the obtained synonymy groups is scanned from the left to the right. The current group is cut in length to the nearest power p of 2. The next piece of length p is picked up from the bit sequence to be hidden, and the synonym is taken with the number equal to this piece. It replaces the synonym in the source text. If grammatical features of the newcomer (number or person, depending on the part of the speech) differ from the source word, special operations of agreement are performed.

This process continues until one of the inputs is exhausted. In normal situation, the hided information sequence ends earlier and hereafter the source text does not change.

10 Conclusions and Future Work

We have proposed a method for synonymic paraphrasing of natural language texts by contextually controlled synonymic variation of individual words. We quantitatively measure the naturalness (appropriateness) of a word or its synonyms in the given context as the frequency of collocations of the given word with the words from the context. The corresponding frequency in the general texts is measured as their relative frequency in Internet, using an Internet search engine. As a synonymy dictionary we use WordNet.

We have pointed out at least two practical applications of our method. The first one is style checking and correction. For each word in the text, we generate its possible synonymic variants appropriate in the given context; if there are much better variants, we suggest them to the user as possible improvements. What is more, comparing the average appropriateness of words in the given text with that of the best word choice at each position generated automatically, we can assess the stylistic quality of the given text as optimal or allowing significant improvement.

The second application is linguistic steganography: hiding arbitrary information within a text without changing its grammaticality, cohesion, and even meaning. Recently this has been an active research area having in its turn a number of important applications. One is a way of secret communication in the situation what the very fact of communication is to be kept secret (and thus, usual cryptographic methods prove insufficient). Another one is watermarking: digitally "signing" the text in such a way that the signature can be restored even after significant and probably intentional transformations of the text or its format (e.g., plagiarism).

In the future, we plan to consider other applications of the suggested ideas, e.g., word choice in automatic translation. Also, the measure of the linguistic appropriateness is to be extended to better take into account various linguistic phenomena.

References

1. Apresian, Ju. D., *et al. ETAP-3 Linguistic Processor: a Full-Fledged NPL Implementation of the Meaning–Text Theory*. Proc. First Intern. Conf. Meaning–Text Theory, MTT 2003, Paris, Ecole Normale Supérieure, June 2003, p. 279-288.
2. Bogatz, H. *The Advanced Reader's Collocation Searcher* (*ARCS*). ISBN 09709341-4-9, www.asksam.com/web/bogatz, 1997.
3. Bentivogli, L., E. Pianta. *Detecting Hidden Multiwords in Bilingual Dictionaries*. Proc. 10th EURALEX Intern. Congress, Copenhagen, Denmark, August 2002, p. 14–17.
4. Bolshakov, I. A., *Getting One's First Million... Collocations*. In: A. Gelbukh (Ed.) *Computational Linguistics and Intelligent Text Processing*. Proc. 5th Intern. Conf. on Computational Linguistics CICLing-2004, Seoul, Korea, February 2004. Lecture Notes in Computer Science No. 2945, Springer, 2004, p. 229-242.
5. Carrol, J., G. Minnen, D. Pearse, Y. Canning, S. Delvin, J. Tait. *Simplifying text for language-impaired readers*. Proc. 9th Conference of the European Chapter of the ACL EACL′99, Bergen; Norway, June 1999.
6. Chapman, M., G. Davida. *Hiding the hidden*: *A software system for concealing ciphertext as innocuous text*. In: Yongfei Han, Tatsuaki Okamoto, Sihan Qing (Eds.) Proc. 1st Intern. Conf. on Information and Communication Security ICICS 97. Lecture Notes in Computer Science 1334, Springer, 1997, p. 335-345.
7. Chapman, M., G. I. Davida, M. Rennhard. *A Practical and Effective Approach to Large-Scale Automated Linguistic Steganography*. In: G. I. Davida, Y. Frankel (Eds.) *Information security*. Proc. of Intern. on Conf. Information and Communication Security ICS 2001, Lecture Notes in Computer Science 2200, Springer, 2001, p. 156-165.
8. Fellbaum, Ch. (Ed.) *WordNet*: *An Electronic Lexical Database*. MIT Press, 1998.
9. Mel'čuk, I. *Dependency Syntax*: *Theory and Practice*. SONY Press, NY, 1988.
10. *Oxford Collocations Dictionary for Students of English*. Oxford University Press. 2003.
11. Smadja, F. *Retreiving Collocations from text: Xtract*. Computational Linguistics. Vol. 19, No. 1, 1990, p. 143-177.
12. Vossen, P. (Ed.). *EuroWordNet General Document*. Vers. 3 final. www.hum.uva.nl/~ewn.
13. Biemann, C., S. Bordag, G. Heyer, U. Quasthoff, C. Wolff. *Language-independent Methods for Compiling Monolingual Lexical Data*. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (CICLing-2004). Lecture Notes in Computer Science, N 2945, Springer, 2004, pp. 214–225.