ISSN 2007-9737

Inferences for Enrichment of Collocation Databases by Means of Semantic Relations

Alexander Gelbukh

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Mexico

www.gelbukh.com

Abstract. A text consists of words that are syntactically linked and semantically combinable—like "political party," "pay attention," or "stone cold." Such semantically plausible combinations of two content words, which we hereafter refer to as collocations, are important knowledge in many areas of computational linguistics. We present the structure of a lexical resource that provides such knowledge—a collocation database (CBD). Since such databases cannot be complete under any reasonable compilation procedure, we consider heuristic-based inference mechanisms that predict new plausible collocations based on the ones present in the CDB, with the help of a WordNet-like thesaurus: if an available collocation combines the entries A and B, and B is 'similar' to C, then A and C are supposed to constitute a collocation of the same category. Also, we describe the semantically induced morphological categories suiting for such inference, as well as the heuristics for filtering out wrong hypotheses. We discuss the experience in inferences obtained with CrossLexica CDB.

Keywords. Collocations, inference rules, enrichment, synonyms, hypernyms, meronyms.

1 Introduction

Texts, at least usual ones, consist of words that are syntactically linked and semantically combinable (plausible)-like *political party*, *pay attention*, or *stone cold*. We hereafter refer to such semantically plausible combinations of two content words as collocations (discussion of the term *collocation* is given in Section 2). We oppose them to senseless combinations like **green ideas* impossible in usual texts, as well as to combinations of content and non-content words like *is growing* or *she went*, quite usual in texts.

Knowledge of what word combinations are plausible collocations and what are not is important knowledge in numerous applications of natural language processing and computational linguistics, such as word sense disambiguation [12], syntactic disambiguation [14], natural language generation and machine translation [18], or sentiment analysis [10], especially concept-based sentiment analysis [1], to mention only a few.

The reasons for plausibility or implausibility of word combinations are so diverse and difficult to describe that compiling a collection of plausible word combinations proved to be quite a practical solution; we call such a collection collocation database (CDB). While recently neural network-based methods are very actively explored for extracting information on compatibility of words in a text [21, 16], lexicon-based methods are still very popular and remain their importance as a source of highly reliable and perfectly human-interpretable information about various aspects of human language [8, 25].

There is, however, one problem with such a practical approach to characterization of semantically plausible word combinations: their number is so big that any, even very large, CDB will always be incomplete. Automatic detection and generation of new word combinations is an important problem [11, 15]. As one of possible solutions to this problem, we suggest a heuristic-bases method of automatic generation of new, highly probably plausible, word combinations absent in the CDB basing on the ones present in the CDB, with the help of a WordNet-like semantic dictionary.

It is very important to clarify that our goal here is not to analyze the linguistic laws underlying collocation formation, which are very complex and are to be a topic of a much more detailed study. Instead, our goal is to describe a totally automatic heuristic procedure for generation of millions of potentially (but not necessarily) correct collocations. These suggestions can be later subject to human verification; in this case, the present paper can be considered as describing an efficient tool for a lexicographer that suggests candidates for inclusion in the collocation dictionary. Alternatively, and this is how we used these heuristics in our electronic dictionary CrossLexica, the suggestions can be directly provided to the end user, with a clear warning (color coding in the case of CrossLexica) letting the user know that these automatically generated suggestions have not yet been manually verified. Finally, these suggestions can be used internally by downstream applications such as syntactic or semantic analyzers to improve their accuracy in most cases.

Whatever be the application of the heuristic procedure described in this article, it is very important to emphasize that our goal here is not a complete and accurate linguistic characterization of collocation formation, but only balancing recall (quantity) with precision (quality) in automatic generation of suggestions. As any automatic procedure, our heuristics can sometimes generate wrong output, and our work presented in this article consisted in refining our heuristics for them to produce fewer wrong suggestions but without drastically decreasing the number of correct suggestions produced.

Specifically, in this paper we describe:

– The most important types of word combinations (collocations) worth including in CDBs;
– The semantic links relevant for the inference of new collocations;
– The semantically induced morphological categories, that can be used for such inferences;
– The restrictions imposed on the rules for decreasing the number of wrong inference results;
– Our experience with automatic enrichment of the CrossLexica CDB that Prof. I. A. Bolshakov and I have developed in the 1990s [5, 6], for the Russian language.

In spite of that we give mostly English examples (many of them were borrowed from [2, 4]), all our experience convinces us that the inference operations in lexical combinability are universal in their types and scope and are applicable to many, if not all, languages.

In Section 2, we will explain the difference between CDBs and other lexical resources involving relationships between words, such as WordNet. Then we will proceed to a formal definition of collocation and of the structure of a CDB, and finally we will list the types of collocations included in a typical CDB.

Basing on these formal definitions, in Section 3 we will give more details on the necessity of automatic enrichment of a CDB and explain our general scheme of enrichment process. This scheme will be specialized for different variants of reasoning in Sections 4 to 7.

Since our inference procedure is heuristic-based, sometimes it would generate wrong hypotheses unless special precautions were taken to filter out error-prone and doubtful cases. These precautions are described in Section 8.

In Section 9, we will describe some generalizations and special cases not considered in the previous sections for simplicity of exposition. Finally, in Section 10 we will describe a practical application of the discussed procedures in a realistic-size CDB CrossLexica, and in Section 11, we will draw the conclusions.

2 A Collocation Database

In this section, we explain the difference between CDBs and other dictionaries and lexical resources that involve relationships between words, such as WordNet. We give a formal definition of collocation and of the structure of a CDB and list the types of collocations included in a typical CDB.

2.1 WordNet-like Thesauri Versus Collocation Databases

Large dictionaries of relationships between words have a long history, probably starting with the famous Roget thesaurus. In the recent decade, very large databases (VLDBs), of various links between words have appeared. The well-known VLDB containing semantic relations between English words is WordNet [13]. Its descendant EuroWordNet [23] contains in essence the same set of semantic relations for several other European languages.

The dictionaries of WordNet type give mostly semantic links such as synonyms, hypernyms (is_a), meronyms (part), etc. However, there two types of links between words can be distinguished:

– *Paradigmatic* links describe the words that normally do not occur together in the same text but instead can be, in a way, *substituted* for each other in a text: *John bough a car*—*John bough an automobile* (synonym)—*John bough a vehicle* (hypernym)—*John bough a wheel* (meronym), etc.

– *Syntagmatic* links describe the words that can normally occur together in the same sentence and related to each other: car—buy (*to buy a car*), car—good (*a good car*), car—dealer (*car dealer*).

Thus, the primary purpose of WordNet-type dictionaries is to give the paradigmatic links between words, in the sense described above. Note that there was certain effort made devoted to inclusion of some kinds of syntagmatic information—such as subcategorization frames—in WordNet. Still, this information is supplementary, while the primary goal of WordNet in its current state is to provide the paradigmatic semantic relationships.

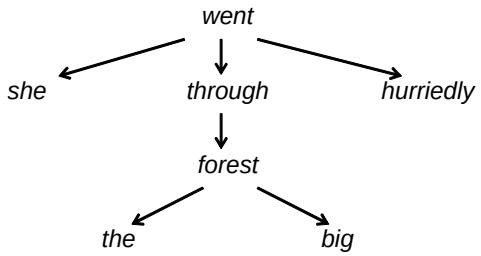
In contrast, in this paper we are interested in the syntagmatic relationships words, such as *pay attention*, *buy a car*, *thick soup*, etc. We call a lexical resource that provides this kind of relationships between words a collocation database.

The collocation databases (CDBs) have already found their own niches of applications, both interactive (word processing, foreign language learning) and non-interactive ones (homonymy resolution, disambiguation of parsing results, segmentation of texts). In any application, better results are directly implied by the completeness of the available collocation collection.

2.2 Collocations

The easiest way to explain formally our definition of a collocation is with the use of dependency grammar formalism [19]. The term *collocation* is used in different meanings in different branches of linguistics. For example, in statistical text processing and corpus linguistics, it is used for any pair of words with high probability of co-occurrence. In another tradition, it is used for certain syntactic relations between words. To avoid any misunderstanding, we give our own definition of this term for the purposes of this paper.

Dependency grammars arrange words of a sentence in a dependency tree, in which collocations cover some subtrees (usually, chains). For example, the sentence *She hurriedly went through the big forest* has the following dependency tree:



Here, by the arrow → we denote an immediate dependency between the two words. In a constituency-based formalism, an immediate syntactic dependency between two words can be roughly thought of as a relation between the head of a constituent and the head of a daughter constituent.

In this tree, one of such chains is ***went*** → *through* → ***forest***, with the highlighted content words at the end nodes and an auxiliary word at the middle node; others are ***hurriedly*** ← ***went***, ***big*** ← ***forest***, etc.

We refer to as a *collocation* such a pair of content words that form a chain a syntactic

dependency tree, with possible auxiliary words between them in the chain. These linking auxiliary words, together with the categories of the content words in question and the type of the syntactic link, serve for categorization of collocations. Thus, the pair (*go*, *forest*) forms a collocation whose type is characterized by the preposition *through*. We will say that there exist a collocational link (of the type characterized by the preposition *through*), between the words *go* and *forest*.

The examples above represent collocations of a ruling verb and its (prepositional), complement, of a verb and its adverbial modifier, and of a noun and its adjective modifier. There exist many other categories of collocations.

It is important to emphasize that the links within collocations, being superficially syntactic, relates semantically combinable content words. For instance, one word in the collocation can fill a syntactic (and simultaneously semantic), valence of the other word. Thus, collocational links are semantic in nature, and these links are immanent for semantic representation of a sentence.

Note again that these semantic correlations have nothing to do with those of WordNet: the words semantically linked in WordNet are not syntactically connected in texts; in fact, they even rarely co-occur in a sentence.

2.3 Collocation Database

A collocation database (CDB), is a lexical resource providing the information on whether two given words can typically form a collocation, and of what type. Note that only semantically plausible collocations are included in the CDB—that is, the ones that can occur in quite natural contexts (here we do not deep into the issue of what contexts are natural; the reader can roughly think of typical collocations as of frequent ones).

The database is realized as three-column table: first word, second word, and the type of their relationship (which includes the auxiliary words such as prepositions, as well as any other necessary grammatical information).

The grammatical information and the types of collocations stored in the CDB are quite rich. Syntactic dependencies connect words of different parts of speech (POS), as shown in Fig. 1. We consider only four main POSs: nouns ***N***, verbs ***V***,

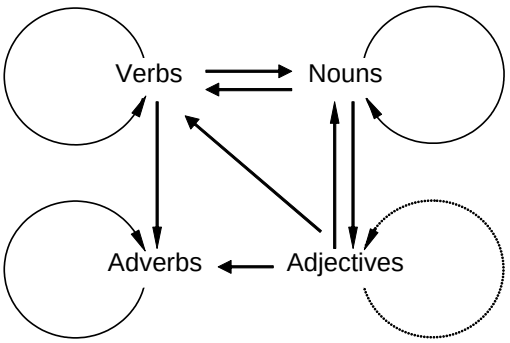


Fig. 1. Various links within collocations

adjectives ***Adj***, and adverbs ***Adv*** in their usual syntactic roles. The role of a noun can be also performed by a noun group (*mass media*), the role of a verb can be performed by a verb group (*is close*), while a prepositional group can play the role of an adjective (*man* → (*from the South*)) or an adverb (*speak* → (*at random*)).

Each arrow in Fig. 1 represents an oriented dependency link (maybe not direct), and nodes linked by arrows are components of a collocation. A CDB can retrieve collocations starting from any component.

A specific syntactic link between two content words can be realized in natural languages through: (1) a preposition or other auxiliary word in between, (2) a specific word order of the linked words, (3) a specific finite form of a verb, (4) grammatical agreement between words, (5) a special value of grammatical case of noun or adjective (e.g., in Slavic languages), or (6) a combination of these ways. All features categorizing a given dependency link should be stored in a CDB in a way sufficient for their correct textual representation. Since different numbers of nouns usually correspond to different sets of collocations, these forms are considered independent entries of a CDB. By the same reason, verb forms of different aspects in Slavic languages are considered separately too.

All types of collocations are registered in CDBs: (1) free combinations (*white dress*, *see* (*a*) *book*, etc.); (2) lexically restricted combinations (*heavy rain*, *give attention*, *do favor,* etc.), *cf*. the notion of lexical functions in [24]; (3) phraseologically fixed (mainly figurative), combinations (*kick the bucket*).

The restricted and phraseological combinations are highly idiomatic and are inserted to CDB just for this reason. Meantime, the criterion of enlisting free combinations is merely their 'commonness', which seems rather diffuse. Nevertheless, the semantics of collocation components restrict their combinability (see below).

Here we consider only two-component (binary) collocations, thus ignoring multi-valence combinations with mutually dependent components. Clearly, a more developed representation can be introduced in the future; for the moment, within binary collocation sets it is possible to represent ternary collocations in a rather primitive manner, using dots for the omitted obligatory valence in each binary projection: *give a book...* and *give... to the boy*. Also, while most lexical entries are single content words, there is also a possibility to use the "words with a space"—indissoluble content word combinations such as *mass media*, *TV set* or *hot dog*, which partially compensates for the current binary structure of the CDB. Note that, if necessary, such a word pair can also, independently, be included in the CDB as a collocation *hot* ← *dog*.

2.4 Specific Types of Collocations

Principally, collocation types are language dependent. However, the most numerous of them proved to be universal, at least for major European languages: Romance, Germanic, and Slavic. Following are the specific types illustrated by English examples. The syntactic link between collocation components is called hereafter relation. Each specific link considered from both sides thus giving two different relations.

– ***HasModifier*** is a relation, in which a given word (noun, adjective or verb), is modifierd by some other word: an adjective or an adverb. This gives collocations: (Noun ***HasModifier*** Adj), (Verb ***HasModifier*** Adv), (Adj ***HasModifier*** Adv), and (Adv ***HasModifier*** Adv), e.g., (*act* ***HasModifier*** *criminal*), (*prepare* ***HasModifier*** *readily*).

– Note that in English, a modifier of a noun can be expressed by putting adjective or another noun in preposition to the modified noun. In Spanish the modifier is usually adjective in postposition agreeing with noun in number and gender; in Russian the adjective is usually in preposition and agrees in number, gender, and case. For a given language, ***HasModifier*** relation implies all specific constraints. In WordNet, adjective modifiers are considered as a separate important class of semantic entities, but it left unknown what nouns can combine with these adjectives.

– ***IsModifierOf*** is relation inverse to the previous one. It determines collocations: (Adj ***IsModifierOf*** Noun), (Adv ***IsModifierOf*** Verb), (Adv ***IsModifierOf*** Adj), and (Adv ***IsModifierOf*** Adv). Examples: *national* ***IsModifierOf*** *economy*, *very* ***IsModifierOf*** *quickly*, *rather* ***IsModifierOf*** *well*.

– ***IsSubjectOf*** determines (Noun ***IsSubjectOf*** Verb) collocations, where Noun is grammatical subject and Verb is its grammatical predicate. E.g., (*heart* ***IsSubjectOf*** *sink*), reflects the collocation *heart sinks*. A predicate agrees with a subject in person and number—in many languages, and additionally in gender—in Russian past tense.

– ***IsNounObjOf*** determines (Noun ***IsNounObjOf*** Verb) collocations, where Noun is object of Verb, direct, indirect or prepositional one: *shake hands*, *arrange* (*with*) *enemy*, etc.

– ***IsVerbObjOf*** determines (Verb ***IsVerbObjOf*** Verb) collocations, in which one verb in infinitive is subordinated to another verb: *prepare* (*to*) *sleep.*

– ***IsNounComplOf*** determines (Noun ***IsNounComplOf*** Noun) collocations, where one noun is subordinated to another one: *adjustment* (*of a*) *clock*.

– ***IsVerbComplOf*** determines (Verb ***IsVerbCompl*** Noun) collocations, where Noun rules Verb in infinitive: *readiness* (*to*) *use*.

– ***GovPattern*** represent government patterns, according to which a given word rules other words (usually nouns), as its own valencies. They contain also the lists of specific collocations for each pattern. In the case of verbs, these are approximately their subcategorization frames. For example, the verb *give* has the pattern *who/what gives?* with

examples of dependents *boy, father, government...;* the pattern *what is given?* with examples *hand, money, book...reach*; and the pattern *to whom is given?* with corresponding examples. ***GovPattern*** is an inverse set of relations to ***IsSubjectOf, IsNounObjOf, IsVerbObjOf, IsNounComplOf, IsVerbComplOf*** and analogical relations for adjectives and adverbs.

The comparison of the relations listed above with dependency relations defined in the Meaning ⇔ Text Theory (MTT) [19, 20], shows that the former cover the latter, except for auxiliary relations. To be more accurate, some CDB, relations amalgamate several more fine-grained relations of the MTT.

Comparison of collocations mentioned above with lexical functions (LF), by Mel'čuk [24], shows that lexically restricted of them just correspond to LFs or compositions of LFs. However, LFs represent only a part of a total collocation collection.

3 General Inference Scheme

In this section, we will give more details on the necessity of automatic enrichment of a CDB and explain our general scheme of enrichment process.

3.1 The Problem of Enrichment of a CDB

High completeness of collocation collections (say, 99% coverage of any text) seems unreachable, just as for dictionaries of separate words. What is more, the efforts for collecting word co-occurrences through a text corpus significantly exceed those for separate words. Indeed, if one word of a collocation has the statistical rank N_1 in the large corpus, and the other word has the rank N_2, then in supposition that the both occur nearly independently and are subject to Zipf law, the estimate of the co-occurrence probability is $O(1/N_1N_2)$, as compared with $O(1/N)$ for a separate word of the rank N. Meantime, the less probability of an event, the longer and more diversified corpus is needed—to guarantee statistically significant results.

Hence, for compiling a highly complete CDB, it would be necessary to automatically scan through—with further manual control and post-editing—a huge and highly poly-thematic corpus at expense of a tremendous labor. What is more, natural language is not static, so that new candidates for stable collocations appear in texts continuously.

With such aggravations, the compilation of a large collection of collocations seems to be a problem not only of a sophisticated statistical processing but also of experimentation with automatic enrichment of CBDs, i.e., of automatic generation (inference) of new collocations based on their already available amount. Even if some inferences would give a rather high percentage of wrong collocations, the correct ones, after checks by native speakers, might be incorporated into CDBs, thus directly increasing their size. At the same time, the errors of automatic inference are usually very instructive for further research.

This paper considers enrichment of collocation collections by means of automatic generation of new plausible collocations in runtime.

Already available entries of Collocation Databases are taken as components of collocations to be generated.

3.2 The General Inference Scheme

WordNet-type relations are considered a tool for the generation. So we suppose that a CDB is supplied beforehand with semantic relations relevant for the generation. All semantic relations impart universality to CDB, whereas for generation there proved to be relevant synonymy, hypernym / hyponym (genus-species relations), and meronym / holonym (parts-whole relations).

The inference rules are of production type. Let a collocational link **D** of a specific dependency category (e.g., ***HasModifier***) combine the entries *A* and *B*, and *B* has semantic 'similarity' of a class **S** with an entry *C*. Then our hypothesis consists in that *A* and *C* constitute a collocation of the same category **D**:

$$(A\ \mathbf{D}\ B)\ \&\ (B\ \mathbf{S}\ C) \Rightarrow (A\ \mathbf{D}\ C). \qquad (1)$$

The dependency link **D** can be of any direction—either ***IsDependentOn*** or its converse ***HasDependent***—as we will show later (note that

in both the left and the right part of the formula, the link has the same direction).

We use the term ***inference*** for the generation of new collocations basing on the formula (1). Note that the rule (1), is only a heuristic, and its result is a hypothesis that, strictly speaking, not always is true. In most cases, however, it is, and the rest of this paper will be devoted to the descriptions of special cases and precautions we consider to make this heuristic as reliable as possible.

One can also see that we study phenomena of lexical semantics and lexical combinability, not of immediate discourse semantics. However, we believe that without study of collocations the inferences within semantic representation of discourse are impossible.

4 Synonymy-based Inference

Consider first an example of the inference based on synonymy. Suppose that the noun *coating* has no collocations in CDB, but it belongs to the synonymy group (synset, in terminology of WordNet), with the *layer* member supplied by collocations. It is natural to conclude that the information connected with *layer* can be assigned to all other synset members lacking the complete characterization. Thus, starting from the collocation *to cover with a layer*, the collocation *to cover with a coating* can be inferred.

4.1 Various Definitions of Synonymy Relation

The first way to reveal synonymy between two given words, W_1 and W_2, is to study the results of substitution of W_1 for W_2 and vice versa in some contexts. If this is always possible, W_1 and W_2 are absolute synonyms. Some researchers acknowledge synonymy between two words interchangeable in a unique context [23]. Though inclusion and gradable intersection of two context sets are well known, the threshold of synonymy admission was never defined strictly.

Another way to describe members of a synset is to decompose their meanings to more ‘simple’ parts, in order to find common and diverse elements. This is a hard lexicographic problem [3]. Just now we abstract from both testing contexts and the decomposition, in order to study relations between synonyms in the set theoretical approach.

The set theory can define synonymy in various ways and with a various degree of strictness. The simplest way is to use the notion of equivalence. The whole set of words under consideration is divided into subsets of equivalence. As applied to synonymy, these subsets are synsets [13]. The elements of each synset are subject to the following conditions of equivalence relation **E**: reflexivity x **E** x, symmetry x **E** $y \Rightarrow y$ **E** x, and transitivity x **E** y & y **E** z $\Rightarrow x$ **E** z. These imply that synsets do not intersect, and each element of the whole set belongs to one synset only.

The equivalence definition of synonymy strictly corresponds only to absolute synonymy. In a common case, if x is ‘similar’ to y, and y is ‘similar’ to z, the ‘similarity’ of x to z is not yet guaranteed. The intersection of context sets for x, y, and z can be empty.

If the transitivity is discarded, the so-called tolerance relation is valid, with the reflexivity and symmetry conditions only.

Tolerance proved to be too loose for synonymy definition. Indeed, word chains with adjacent words complying with tolerance can be arbitrarily long, whereas their two utmost words can have no common contexts. We do not know set theory relations that are strictly between equivalence and tolerance and well suited to synonymy formalization.

Meantime, many lexicographers have adopted, rather spontaneously, an idea of title-forming member of a synset, which reflects the meaning of a synset most generally and neutrally. Let us call the title-forming member the dominant of a synset.

The introduction of dominant eliminates strict equivalence between synset members; they remain just ‘similar’ to each other. In lexicographic practice, it is necessary to describe some similarities between different synsets as well. For this purpose, lexicographic practitioners had introduced into printed dictionaries the *see also* labels referring to external dominants. In electronic dictionaries, the references outside are usually included into corresponding synsets as their ordinary members—without any label. The elegance of equivalence relation disappeared, but the word distribution among synsets became more

Table 1. Characterization of synonyms by feature values

Synonym	Cause	Character	Manner	State of mind	Soc. estimate
acknowledge	*facts / circumstances*	*public*	*irrelevant*	*indifferent*	*indifferent*
admit	*facts / arguments*	*personal*	*irrelevant*	*disinclined*	*negative*
own	*circumstances*	*personal*	*verbal*	*ready*	*indifferent*
avow	*circumstances / conscience*	*public*	*verbal*	*ready*	*negative*
confess	*conscience*	*personal*	*verbal*	*ready*	*indifferent*

flexible and corresponds better to lexicographers' intuition.

Since the approaches of strict equivalence and diffusely comprehended dominance coexist in computational linguistics [3, 17, 23], we should consider them both.

4.2 Synonymy as Equivalence

If the given entry μ is a member of the synset $\{s_1,...,s_N\}$ and Q other members of the synset have collocations of the given type **D**, the collocations of the same type with the word μ can be inferred as intersection of those Q collocation sets for any x:

$$\forall_{q=1}^{Q}\left(\left(\mu\ \textbf{\textit{HasSyn}}\ s_q\right)\&\left(s_q\ \mathbf{D}\ x\right)\right)\Rightarrow\left(\mu\ \mathbf{D}\ x\right). \quad (2)$$

where HasSyn is a kind of relation S introduced in (1).

4.3 Synonymy with Dominants

In order to illustrate how a dominant can be selected within a synset, let us borrow the synset {*acknowledge, admit, own, avow, confess*} from [2]. The main semantic dissimilarities between the synset members are: (1) Cause of doing, i.e. the pressure of *facts, arguments, circumstances, conscience* or some their disjunctive combination; (2) Character of doing: *public* or *personal*; (3) Manner of doing: *verbal, indirect* or *irrelevant*; (4) Subject's state of mind while doing: *ready, disinclined* or *indifferent*; (5) Social estimate of the target of doing: *indifferent* or *negative*.

In a rough approximation, [2] characterizes these synonyms by the feature values shown in Table 1. The synset has no indisputably 'neutral' member, but *acknowledge* has a maximum of irrelevant or indifferent values, so that it could be admitted dominant.

Returning to a general case, if a queried entry μ is dominant, it should be maintained in a CDB with usual collocations, so that no inference is needed. For non-dominants, the inference rules could to be taken as follows:

– If the entry μ belongs to only one synset $\{D, s_1,... \mu,...s_N\}$ with the dominant D, any collocation valid for D is supposed valid for μ: for any x:

$$\left(\mu\ \textbf{\textit{HasDom}}\ D\right)\&\left(D\ \mathbf{D}\ x\right)\Rightarrow\left(\mu\ \mathbf{D}\ x\right), \quad (3)$$

where ***HasDom*** is another type of the **S** relation.

– If the entry μ belongs to several synsets with their own dominants D_q:

$$\{D_1, s_{11},..., \mu,..., s_{1N_1}\},\ \{D_2, s_{21},..., \mu,...s_{2N_2}\},...,$$
$$\{D_k, s_{k1},..., \mu,..., s_{kN_k}\}$$

and those collocations are supposed valid whose analogues are registered in CDB for all dominants: for any x:

$$\forall_{q=1}^{k}\left(\left(\mu\ \textbf{\textit{HasDom}}\ D_q\right)\&\left(D_q\ \mathbf{D}\ x\right)\right)\Rightarrow\left(\mu\ \mathbf{D}\ x\right). \quad (4)$$

5 Hypernym-Based Inference

The hypernym-based inference can be explained by the following example. Let the term *refreshing drink* have the complete collocation set in CDB, with the verbs constituting collocations *bottle, have, pour*, etc. The same information on *Coca Cola* may be absent in the DB, it is only known that

they are hyponyms of *refreshing drink*. The inference attaches the information connected with the hypernym to all its hyponyms lacking same type of collocations. Thus, it is inferred that the mentioned verbs are applicable to *Coca Cola* too.

5.1 Various Kinds of Hierarchies

A set of hypernym / hyponym relations can be described in two possible ways:

– All relevant terms are included into a united classification hierarchy (i.e., a tree), so that a unique hypernym corresponds to each hyponym in it, except for the uppermost node (i.e. the root of the tree). We call such case ***monohierarchy***.

– There are several hierarchies, and CDB entries are distributed among them, so that an entry can participate in several hypernym-hyponym relations based on different principles of classification. We call such case ***crosshierarchy***. The whole structure is a directed acyclic graph, and each its node, except for a few uppermost nodes of partial hierarchies, can have arbitrary numbers of hypernyms. This means that several ways for enrichment through hypernym could exist.

5.2 Mono-Hierarchy

Suppose that the relation ***IsA***^{1} (here *1* stands for "one-step"), gives the immediate (nearest), hypernym h_1 for the entry μ. The h_1 is unique within a mono-hierarchy (or does not exist—for the root of the tree). The emptiness of the collocation set for the h_1 necessitates attaining hypernyms of higher levels. The transition from μ to its unique k-th hypernym can be represented by the formula:

$$(\mu\ \textbf{\textit{IsA}}^k\ h_k) = (\mu\ \textbf{\textit{IsA}}^1\ h_1)\ \&\ (h_1\ \textbf{\textit{IsA}}^1\ h_2)\ \&\ \ldots\ \&\ (h_{k-1}\ \textbf{\textit{IsA}}^1\ h_k).$$

The inference by means of hypernyms seems evident: attaining the first met hypernym h_k with a non-empty collocation set of the queried type and assigning these collocations to μ: for any x:

$$\left(\mu\ \textbf{\textit{HasUp}}^k\ h_k\right)\&\left(h_k\ \mathbf{D}\ x\right)\Rightarrow\left(\mu\ \mathbf{D}\ x\right). \qquad (5)$$

5.3 Cross-Hierarchy

In a cross-hierarchy, more than one path comes up from a CDB entry. The following inference procedure can be thus proposed. All k-th hyponyms of the entry μ, k = 1, 2,..., are searched widthwise, until at least one of them has a non-empty collocation set. If only one set is non-empty at k-th layer, the formula (5), remains valid, elsewhere the intersection of all non-empty sets is taken. To represent this mathematically, let us enumerate different non-empty homonyms of k-th layer as q = 1, 2, ..., Q. Then the following rule is valid: for any x,

$$\forall_{q=1}^{Q}\left(\left(\mu\ \textbf{\textit{HasUp}}^k\ h_{k_q}\right)\&\left(h_{k_q}\ \mathbf{D}\ x\right)\right)\Rightarrow\left(\mu\ \mathbf{D}\ x\right). \qquad (6)$$

The widthwise search excludes situations when a collocation set of k-th hypernym is taken, whereas m-th non-empty hypernym's set exists for the same μ with $m < k$.

6 Inference Based on Meronymy and Holohymy

The meronymy relation (*x* ***HasMero*** *y*), states that *x* has *y* as a part, whereas holonymy (y ***HasHolo*** x) is inverse relation: *y* is a part of *x*. In some simple cases both *x* and *y* are single words in a given language, like (*clientele* ***HasMero*** *client*) or (*tree* ***HasMero*** *trunk*) in English.

In contradistinction with synonyms and hypernyms, one can imagine the moving of collocations in both directions. E.g., the collocations (*to*) *serve / satisfy / draw in/ lose... a client* are equally applicable to *clientele* and, vice versa, nearly all collocations valid for *clientele* are valid to *client* too. That is the inference rules are: for any *x*:

$$(\mu\ \textbf{\textit{HasMero}}\ y)\ \&\ (y\ \mathbf{D}\ x)\Rightarrow(\mu\ \mathbf{D}\ x), \qquad (7)$$

$$(\mu\ \textbf{\textit{HasHolo}}\ y)\ \&\ (y\ \mathbf{D}\ x)\Rightarrow(\mu\ \mathbf{D}\ x). \qquad (8)$$

In fact, not all *x* in the formulas (7) and (8) can be taken, and there exist other complications in the case of meronymy / holonymy.

It is known [23] that meronymy / holonymy can be of at least five different types: (1) a part proper, like *finger* of *hand*, (2) a portion, like *drop* of *liquid*;

(3) a narrower location, like *center* of *city*; (4) a member, like *player* of *team*; (5) a substance the whole is made of, like *stick* of *wood*. The existence of various types of meronyms and of combinatorial representations of meronyms and holonyms makes the problem of inference highly complicated.

In particular, in some pairs of words related by the meronymy / holonymy relation, one word names a standard, or typical, portion of what is denoted by the other word. We call such a portion a quantum, for example:

(*to drink beer*) & (*pint of beer* = Quant (*beer*)) $\Rightarrow$ (*to drink a pint of beer*)

and similarly for *to add a glass of water*, *to drink a cup of wine*, *to eat a loaf of bread*, etc.

7 Morphology-Based Inference

Some morphological categories semantically motivated, i.e., are explicitly expressed at the semantic level of representation of the text. These categories can be used for inferences. In this section, we will describe such semantically induced morphological categories that can also be used for the inferences. Such categories are grammatical number of nouns—for any European language, and aspect of verbs—for Slavic languages. The separate entries of a CDB might be complementary words, e.g., singular *vs*. plural forms of a noun, and the words differing in a semantically induced category can be characterized to various extent, so that the better attended word could help to characterize the poorer attended one.

7.1 Inference Based on Grammatical Number

In all European languages, nouns have grammatical category of number, usually with two values: singular and plural (some languages have also the dual number). Since number values frequently imply different collocation sets, they should be included into a CDB as separate entries.

Any CDB can contain a collocation set (of a given type) of only one most 'habitual' value, singular or plural. Then one has no choice but to take the same set, maybe after certain filtering out, for the complementary value, plural or singular: for any *x*:

$$(\mu\, \textbf{\textit{HasComplNum}}\ y)\ \&\ (y\ \mathbf{D}\ x) \Rightarrow (\mu\, \mathbf{D}\ x). \quad (9)$$

As to the filtering, it should be recollected that at semantic level singular of a noun N is usually opposed to plural through the predicates ***Single***(*N*) *vs.* ***MoreThanOne***(*N*). Hence, we can introduce restrictive lists of words with a semantic element of singularity/uniqueness and thus well combinable only with singular, and, vice versa, containing an element of collectiveness / multiplicity and thus well combinable only with plural.

For modificatory collocations, following are several English plural-oriented modifiers: *many, multiple, numerous, various, different, diverse, equal, unequal, of all kinds*, etc. Singular-oriented modifiers are fewer: *unique, single, solitary, lonely, individual*, etc. Of cause, some of them can be used in complementary number context too, but their transfer to the complementary number is in general case incautious.

7.2 Inference Based on Verb Aspect

Verbs in Slavic languages contain grammatical category of aspect, whose alternative values *perfect* and *imperfect* have no analogies in Germanic or Romance languages. Aspect is considered separately from tense and reflects completeness *vs*. incompleteness of the verb action, i.e. whether did the action come to its logical end or not. Note, that the *perfect* value of aspect differs in meaning from *present perfect* tense in English or the *pretérito perfecto* tense in Spanish, since these two convey the idea of an action with a result lasting up to moment of speech, irrespectively of action completeness.

In Russian, verbs in imperfect are frequently modified with the following adverbs or adverbial combinations impossible in perfect: *vsjacheski* 'in every way possible,' *mnogokratno* 'repeatedly,' *postojanno* 'continually,' *dolgo* / *dolgoe vremja* 'long / a long time,' etc. Meantime, Russian perfect-oriented adverbs and adverbials are different: *vdrug* 'suddenly', *okonchatel'no* 'completely / definitively,' *bespovorotno* 'irrevocably,' *davno* 'long ago', *nedavno* 'recently,' *neskol'ko* <*let* / *mesjacev* / *nedel'* / *dnej* / *chasov* / *minut*> *nazad*

'several <years / months / weeks / days / hours / minutes> ago,' etc. Hence, for any *x*, the inference towards complementary aspect value:

$$(\mu\ \textbf{\textit{HasComplAspect}}\ y)\ \&\ (y\ \mathbf{D}\ x) \Rightarrow (\mu\ \mathbf{D}\ x), \quad (10)$$

seems possible when used with the restrictive lists of mentioned types.

8 Inference Precautions

Several precautions to be taken for avoiding the most numerous inference errors were already described; here we will describe several others restrictions. Some of these restrictions are prohibitive lists of entries, while the other point out a whole subclass of entries prohibited for specific inferences.

8.1 Not Considering Some Syntactic Relations

Some syntactic relations do not permit enrichment inferences. The most error-prone of them is ***GovPattern*** for verbs. To illustrate this statement, let us take the English synset {*choose, select, pick, cull, elect, opt, single out*}. Each of them, except *opt*, forms the target of selection in the shape of direct complement, while *opt* uses prepositional complement *opt for something* for this purpose. Each of them, except *elect* and *opt*, can introduce a prepositional complement with *among* or *from*— for options of the selections. Thus, ***GovPattern***, say, of the verb *opt* cannon be inferred correctly based on data of the other synset members.

Meantime, the relations inverse to ***GovPattern*** can be freely used for the inferences. Indeed, if ***IsNounObjOf*** gives for *country* collocational counterparts (*to*) *cross / visit / ruin*..., then all these verbs can form collocations with any specific country name.

8.2 Ignoring Classifying Modifiers

Some inferences for modificatory collocations also give wrong results. For example, *berries* can have nearly any color, smell and taste, but its hyponym *blueberries* are scarcely *yellow*. Among modifiers there exist rather broad class that is the most error-prone. Let us see the following wrong inference:

(*Argentina* ***IsA*** *country*) & (*European* ***IsModifierOf*** *country*) ⇒
*(*European* ***IsModifierOf*** *Argentina*).

To exclude such cases, we should not use modifiers, which can be referred to as classifying. They convert a specific entry to its hyponym, e.g., *country to <European / American / African> country*. As to other modifiers for *country*, e.g., *agrarian*, *beautiful, great*, *industrial*, *small*, they compose correct collocations rather frequently: *agrarian* / *beautiful* / *great... Argentina*. However, the modifiers like *southern* or *northern* seeming good for inferences with *country*, change its meaning while inference, whereas modifiers like *woody*, *hot* or *densely populated* can be inapplicable by other reasons.

8.3 Ignoring Labeled Words and Collocations

Each printed dictionary uses the set of special labels explaining word usage. The number of labels can reach 30 or more, and they are usually introduced without strict classification principles. In electronic editions destined for a human user the diffuse situation is usually the same.

Let us try to introduce a simple set of usage marks for CDBs seeming minimally sufficient for a common user. It should contains at least two coordinates:

– ***Scope of use*** with five grades: (1) neutral: no limitations on the use and no label needed; (2) special, bookish or obsolete: written use is recommended when meaning is known to the user; (3) colloquial: use in writing is not recommended; (4) vulgar: neither written nor oral use is recommended; (5) incorrect: is not recommended as contradicting language laws.

– ***Idiomacity*** reflects literal vs. figurative (metaphoric) use of words and collocations, with three grades: (1) direct use (no label needed); (2) both figurative and direct interpretations possible (*kick the bucket*), and (3) figurative use only (*hot dog*).

The labels of idiomacity and scope at a given word are transferred to its collocations in CDB. The idiomacity labels can mark separate collocations in CDB.

While inference, the majority of labeled collocations give wrong results, as for:

(*poodle* ***IsA*** *dog*) & (*hot* ***IsModifierOf*** *dog*)$_{\text{figurative}} \Rightarrow$ *(*hot* ***IsModifierOf*** *poodle*).

Hence the most cautious way is to avoid all labeled collocations in inferences.

9 Case of Ternary Collocations

Synonyms, holonyms, and especially meronyms are frequently expressed through a specific collocation containing the counterpart of the relation as its syntactic component. E.g., *gladness* has a synonym *feeling of gladness*; *cow* has the holonym *herd of cows*, and *Parliament* has the meronym *Member of Parliament*. The omission of the syntactically subordinated part of such collocations (...*of cows*, ...*of Parliament*) is possible only in the fully definite situation, compare *He came to the center* vs*. He came to the center of the city*.

While in our current CDB structure collocations are binary relations, the collocations described above need ternary (or more) representation. Here we will not describe formally the solution to this problem, which would consist in re-definition of a collocation that would permit more than binary dependency trees. Instead, we will describe the corresponding issues quite informally.

While trying to assign collocations valid for a single word to its combined counterpart, one run against the binary nature of links within CBDs under research. At the same time, it seems unnatural to store in the CDB collocations like *appeal to Member* without mentioning *Parliament* or (*to*) *drink* (*a*) *drop / cup / glass / bottle / barrel*... without mentioning *water*, *wine* or *beer.* Indeed, the semantic link between content words like *drink* and *glass* is indirect: *glass* is only container of *water, wine* or *beer*.

In order to neutralize the drawbacks of binary CDBs through runtime inference of ternary collocations, we can use the same formulas (7) and (8), where μ is taken in the shape of the 'genitive-type' collocation (δ ***Of*** *z*). Such collocations exist in many languages, *z* coinciding with *y* in (7), (8) or differing only in number. Such collocations should be marked in CDB as meronym or holonym of its component z. Thus, (μ **D** *x*) relation links the ruling node δ of genitive-type collocation μ with *x*. E.g., the rule (7) infers *to appeal to the Member of Parliament* based on collocations *to appeal to Parliament* and *Member of Parliament,* whereas the rule (8) infers *to drink a glass of vine* based on *to drink vine* and *glass of vine.*

Various languages have other types of collocations expressing holonyms and meronyms. For example, in English the noun *user* has the collocational holonym *user community*, with *user* modifying *community*. Since *community* is a component of the collocation related to *user*, it is easy to mark the whole collocation as holonym of *user*, just as in the previous case. However, to mark *French province* as meronym of *France* is much more difficult: neither *French* nor *province* coincides with *France*.

All the types discussed above are mainly applicable to physical objects (things, substances, living creatures). Meantime, there exist numerous collocations reflecting entities with properties similar to meronyms but applicable to abstract nouns. Following are English examples: *particle of truth, shadow of doubt, pangs of conscience, fit of temper, flame of wrath, summit of glory*.

The verbs admitting *truth, doubt, conscience, temper, wrath, glory,* etc. as their valence fillers equally admit the mentioned collocations—for the same syntactic roles. For example, *to feel a doubt* implies admissibility of *to feel a shadow of doubt*, while *to achieve glory* implies admissibility of *to achieve the summit of glory*.

In such situations, the introduction of an additional sort of 'meronyms' can be recommended. For such abstract nouns (and not only for them, compare *heart of the desert, patch of fog, bout of coughing*, etc.), the meronymous collocations can be marked in a CDB in the same manner, to facilitate inference of ternary collocations instantiated above.

10 Experience of the CrossLexica Electronic Dictionary

The CrossLexica collocation database was mainly developed in the 1990s [5, 6] with Russian as the basic language and English only for queries, and is

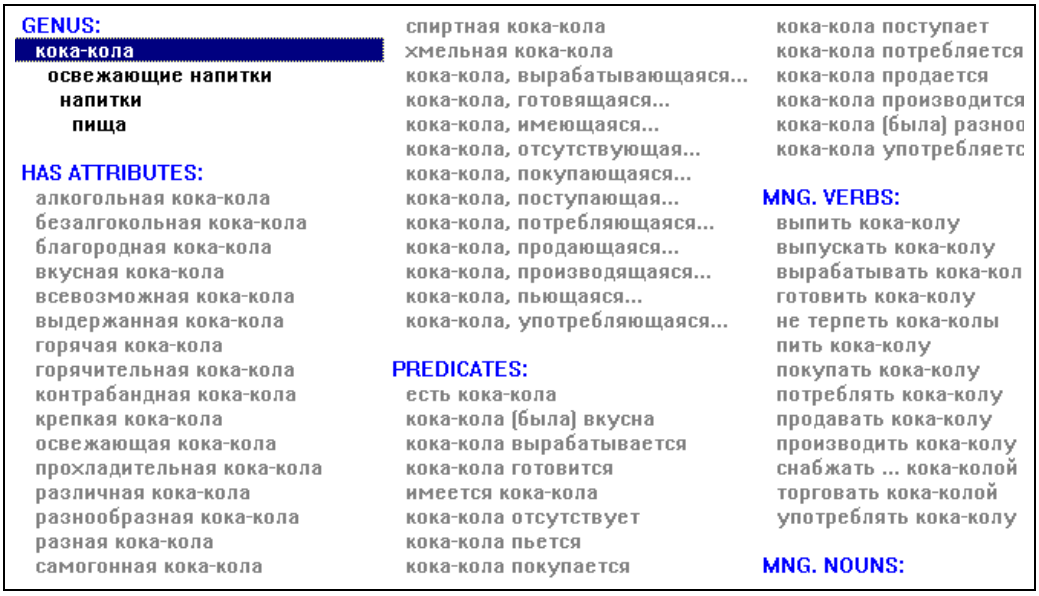


Fig. 2. An example of enrichment with wrong variants generated by the heuristics with some of the precautions disabled: the entry *Coca Cola* in Russian CrossLexica. In low-contrast font color, automatically generated collocations are shown, such as *hot Coca-Cola* or *Coca-Cola is sold*. The font color alerts the user of that the automatically generated collocations have not been manually verified.

constantly growing since that. Its proportions can be currently characterized by the following statistics of collocations, measured in unilateral links, in its core subset (though these figures are constantly growing):

Modificatory collocations	615,600
Verbs *vs*. their noun complements	348,400
Nouns *vs*. their predicates (verbs or short-form adjectives)	235,400
Nouns *vs*. their noun objects	216,800
Verbs *vs*. their infinitive objects	21,500
Nouns *vs*. their infinitive complements	10,800
Total	1,448,500

It is interesting to mention that the mean collocational fertility proved to be a rather constant value. For example, a noun can be object of approx. 24 verbs in Russian, and this value does not change during five recent years of version renewals. This shows that the so-called free collocations are nevertheless constrained semantically.

The semantic relations relevant for us now are: synonyms 193,900; holonyms / meronyms 17,300; hyponyms / hypernyms 8,500; totally 219,700. Among synonyms 39% are nouns, 28% are verbs, 22% are adjectives, 11% are adverbs, and the number of unilateral links is counted as $\Sigma_i n_i(n_i–1)$, where n_i counts *i*-th synset. Synset are considered with dominants. Hyponyms and hypernyms are taken only nouns and form cross-hierarchy.

A screenshot of the Russian CrossLexica electronic dictionary for the entry кока-кола (*koka-kola* 'Coca Cola') with examples of the enrichment without implementing the precautions described above is given in Fig. 2. The database contains only its hypernyms: (*Coca Cola* ***IsA***[1] *soft drink*), (*soft drink* ***IsA***[1] *drink*)..., so that all collocations are inferred based on *drinks* (*refreshing drinks* has no collocations) and are given in low contrast. The statistics of correct inference are as follows: ***HasModifier*** (*Has Attributes* group in Fig. 2) 10%, ***IsSubjectOf*** (*Predicates*) 93%, ***IsNounObj*** (*Mng. Verbs*) 100%, and ***IsNounCompl*** (*Mng. Nouns*) 94%. Here, very poor results for modificatory collocations are explained by disabling some of the abovementioned ideas on inferences. For

example, the modifiers алкогольная *alkogol'naja* and спиртная *spirtnaja* (both 'alcoholic') are classificational, and they are moved to other place after a revision of classification; the modifiers различная *razlichnaja* and разнообразная *raznoobraznaja* (both 'various / different') are plural-oriented and are filtered out based on other reasons, etc. Evaluation of this process (with disabled precautions) has shown that the portion of generated collocation was less than 8% of the total CDB, and more than 3% gave so high percentage of wrong collocations that the generated sub-collections were fully revised by hand and then inserted to the CDB as its immanent part.

Our heuristics showed varying precision for different semantic fields. For instance, the transfer of collocations from the name of genus to the name of species worked almost ideally for dogs (such as transferring collocations of *dog* to *riesenschnauzer*) and with acceptable quality for flowers, though generating some number of wrong suggestions. However, they showed practically unusable results for species of berries, most of which finally required manual specification of all their collocations instead of automatic transfer of collocations from the word *berry*.

We keep refining our heuristics in parallel with detailed characterization of words occurred in texts more and more rarely, and some potential candidates for inference disappear with them. In this way, the total portion of inferred collocations diminishes. However, expansion and perfection of the synonyms and the cross-hierarchy act in the opposite direction. That one of these two opposite tendencies could prevail is unlikely. For example, there are *callas* among flowers, the specie nearly unknown (totally imported) in Northern countries. We cannot even imagine how large a Russian text corpus should be for gathering all evidently possible collocations with callas: (to) *throw, lay on, choose, grow, present, buy, gather*, etc. As to humans, they immediately use such collocations just after knowing that callas are flowers. The computer systems should act analogously.

11 Conclusions

The method is developed of generating new collocations based on an available collocation database and several semantic relations touching one component of source collocations. In the target collocations, the related component is changed to semantically similar one. Semantic similarity is determined by synonyms, hypernyms, holonyms, and semantically induced morphological categories.

The enrichment is performed by means of production-type inference rules, looking like deduction formulas of mathematical logic. However, with any semantic similarity including generic terms, the inference rules remain mere heuristics, and the 100-percent correctness of results can be never reached. Even after taking precautionary heuristics (i.e., prohibitive subclasses or word lists), the results frequently leave much to be desired. Thus, any generated collocation should be given to a user with marks of its tentative nature. On the contrary, the inferences proved to be quite opportune for semi-automatic characterization of rather rare words not yet fully described in the immanent part of a CDB.

Our rather pessimistic viewpoint does not exclude a further progress based on a deeper semantic research and experiments with systems similar to the exposed Russian CrossLexica. In fact, computational linguistics currently has no other solution to grow up collocation sets. However, the inferred part of collocation sets will be always rather marginal for the user since the most frequent collocations are included in the CDB directly.

Acknowledgements

This work has been done with a partial support of the Instituto Politécnico Nacional grants SIP 20172008 and SIP 20181792. The CrossLexica system is being developed in close collaboration with Professor I. A. Bolshakov.

References

1. **Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., & Hussain, A. (2015).** Concept-Level Sentiment Analysis with Dependency-Based Semantic Parsing: A Novel Approach. *Cognitive Computation*, Vol. 7, No. 4, 2015, pp. 487–499. DOI: 10.1007/s 12559-014-9316-6.

2. **Apresyan, Y.D. (1979).** *English-Russian dictionary of synonyms in Russian*. Moscow, Russky Yazyk Publ.
3. **Apresjan, Ju. (2000).** *Systematic lexicography*. Oxford Univ. Press.
4. **Benson, M., et al. (1989).** *The BBI Combinatory Dictionary of English*. John Benjamin Publ., Amsterdam, Philadelphia.
5. **Bolshakov, I.A. (1994).** Multifunction thesaurus for Russian word processing. *Proceedings of 4th Conference on Applied Natural Language Processing,* Stuttgart, Vol. 13–15, pp. 200–202.
6. **Bol'shakov, I.A. (1994).** Multifunctional thesaurus for computerized preparation of Russian texts. *Automatical Documentation and Mathematical Linguistics,* Allerton Press Inc, Vol. 28, No. 1, pp. 13–28.
7. **Bolshakov, I.A. & Gelbukh, A. (2000).** *Classification of Collocations in a Lexical Database by Meaning if the Combined Words.* In: A. Guzman Arenas, F. R. Menchaca García (eds.) *Selected papers CIC,* Instituto Politécnico Nacional, pp. 5–15.
8. **Chen, Y. (2917).** Dictionary Use for Collocation Production and Retention: a CALL-based Study. *International Journal of Lexicography,* Vol. 30, No. 2, 2017, pp. 225–251. DOI: 10.1093/ijl/ecw005.
9. **Calzolari, N. & Bindi, R. (1990).** Acquisition of Lexical Information from a Large Textual Italian Corpus. *Proc. of COLING-90*.
10. **Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017).** Sentiment Analysis Is a Big Suitcase. *IEEE Intelligent Systems*, Vol. 32, No. 6, pp. 74–80. DOI: 10.1109/MIS.2017.4531228.
11. **Castro-Sánchez, N.A., Cruz-Domínguez, I., Sidorov, G., & Martínez-Rebollar, A. (2015).** Toward an Automatic Extraction of Collocations in Verb Definitions from a Spanish Explanatory Dictionary. *Revista Signos. Estudios de Lingüística*, Vol. 48, No. 88, pp. 174-196. DOI: 10.4067/S0718-09342015000200002.
12. **Corra, E.A., Lopes, A.A., & Amancio, D.A. (2018).** *Word sense disambiguation*. Information Sciences: An International Journal, Vol. 442, No. C, 2018, pp. 103–113.
13. **Fellbaum, Ch. (1998).** *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, London.
14. **Gelbukh, A. & Calvo, H. (2018).** *Automatic Syntactic Analysis Based on Selectional Preferences*. Springer, pp. 165.
15. **Gelbukh, A., Sidorov, G., Han, S.Y., & Hernández-Rubio, E. (2004).** Automatic enrichment of very large dictionary of word combinations on the basis of dependency formalism. *MICAI, Mexican International Conference on Artificial Intelligence*, Springer, pp. 430–437.
16. **Gómez-Adorno, H., Posadas-Durán, J.-P., Sidorov, G., & Pinto, D. (2018).** Document embeddings learned on various types of n-grams for cross-topic authorship attribution. *Computing*. DOI: 10.1007/s00607-018-0587-8.
17. **Hirst, G. (1995).** Near-synonymy and the Structure of Lexical Knowledge. *Working Notes of AAAI Symposium on Representation and Acquisition of Lexical Knowledge Stanford University*, pp. 51-56.
18. **Liu, X. & Huang, D. (2017).** Translation Oriented Sentence Level Collocation Identification and Extraction. In: Machine Translation. *CWMT 2017. Communications in Computer and Information Science*, Vol. 787. Springer.
19. **Mel'čuk, I. (1988).** *Dependency Syntax: Theory and Practice*. SONY Press.
20. **Mel'čuk, I. & Pertsov, N. (1987).** *Surface Syntax of English: A Formal Model within the Meaning-Text Framework*. Benjamins, Amsterdam.
21. **Poria, S., Peng, H., Hussain, A., Howard, N., & Cambria, E. (2017).** Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing,* Vol. 261, pp. 217–230.
22. ***The Spanish WordNet*. (1999).** Version 1.0, EuroWordNet, LE2-4003 & LE4-8328. CD ROM.
23. **Vossen, P. (2000).** *EuroWordNet General Document*. Vers. 3 final. www.hum.uva.nl/~ewn.
24. **Wanner, L. (1996).** Lexical Functions in Lexicography and Natural Language Processing. *Studies in Language Companion Series,* No. 31.
25. **Zhou, J., Chen, B., & Lin, Y. (2017).** *An Approach to Constructing Sentiment Collocation Dictionary for Chinese Short Text Based on Word2Vec*. *Emerging Technologies for Education,* SETE´17. Lecture Notes in Computer Science, Vol. 10676, Springer.

Article received on 16/03/2017; accepted 07/12/2017.
Corresponding author is Alexander Gelbukh.