Procesamiento del Lenguaje Natural, num. 29 (2002), pp.47-54 recibido 26-05-2002; aceptado 01-07-2002

Word Combinations as an Important Part of Modern Electronic Dictionaries*

Igor A. Bolshakov and **Alexander Gelbukh**
Natural Language Laboratory,
Center for Computing Research (CIC),
National Polytechnic Institute (IPN),
Av. Juan de Dios Bátiz, CP 07738, Zacatenco, DF, Mexico
{igor,gelbukh}@cic.ipn.mx, www.gelbukh.com

Abstract: It is argued that the collections of combinations of content words in the modern general-purpose electronic dictionaries are to be broadened as much as possible. Igor Mel'čuk classifies word combinations into complete phrasemes, semi-phrasemes (combinations with lexical functions), and free combinations. Our suggestion is obvious for complete phrasemes and semi-phrasemes; however, combinations with nonstandard lexical functions, poorly studied and difficult to detect, as well as terminological combinations, are also to be acquired since they are highly language-specific. In addition, we propose heuristic rules for including free word combinations into dictionaries. According to our experience with the CrossLexica word combination database, these rules do not cause any significant growth of the mean number of combinations per content word because of the mutual semantic selectivity of the words to be combined.
Keywords: computational lexicography, collocations, subcategorization, lexical attraction, terminology.

Resumen: Se dan argumentos a favor de que las colecciones de las combinaciones de palabras significativas en los diccionarios electrónicos modernos del uso general se deben extender en lo máximo posible. Igor Mel'čuk clasifica las combinaciones de palabras en frasemas completos, semifrasemas (las combinaciones con las funciones léxicas) y combinaciones libres. Nuestra propuesta es obvia para los frasemas completos y semifrasemas. Sin embargo, las combinaciones con las funciones léxicas no estándares que son poco estudiadas y difíciles de detectar, así como las combinaciones terminológicas, también se deben describir e incluir ya que son altamente dependientes de lenguaje. Adicionalmente, proponemos unas reglas heurísticas para incluir en los diccionarios las combinaciones libres de palabras. Según nuestra experiencia con la base de datos de combinaciones de palabras CrossLexica, estas reglas no causan incremento significativo alguno del promedio de las combinaciones por palabra significativa, gracias a la selectividad semántica mutua de las palabras a combinar.
Palabras clave: lexicografía computacional, colocaciones, subcategorización, atracción léxica, terminología.

1 *Introduction*

Word combinations are important lexical information, which should be given in the dictionaries since it represents knowledge specific for a given language. Neither human nor a computer cannot correctly produce and understand expressions in a given language without knowing what words can be combined, and what cannot, in this specific language.

For example, one cannot correctly translate the expressions *deliver a lecture* or *deliver a letter* into (or from) Spanish without knowing that the correct combinations are *impartir clase* and *entregar la carta*, correspondingly. It can be said that most common errors made by for-

* Work done under partial support of Mexican Government (CONACyT and SNI) and CGEPI-IPN, Mexico.

eigners – and the ones most difficult to understand for their native speaking listeners – are the errors related to word combinations.

However, this kind of information is usually not, or at least not completely and consistently, given in the dictionaries because of the technical difficulties related to its acquisition and characterization.

In this paper we present heuristic rules that simplify their selection for the dictionary. Throughout the paper we will use mostly Spanish examples, unless otherwise is specified.

Let us first agree on terminology. We call a word combination two or more syntactically connected words. By syntactic connection we mean syntactic dependency, as it is understood in the dependency-based grammar formalisms. Roughly speaking, such a combination consists of a syntactically governing word and a syntactically dependent word that immediately modifies the governor, as is indicated by the arrows in the examples below.

In this sense, one can consider as word combinations not only

(1) *prestar* → *atención* 'pay attention'
propuesta ← *concierne*
'proposal concerns',
palabra → *clave* 'key word',
muchacha → *bonita* 'nice girl'

but also

(2) *he* → *tomado* 'have taken',
está → *caminando* 'is walking',
él ← *es* 'he is', or
así → *como* → *para* 'as well as for',

though not contiguous text fragments like

(3) *de* ↷ *la* ← ... 'of the ...',
es ↷ *una* ← ... 'is a ...', or
... → bonita ↶ hablaba 'nice ... was speaking'

since there are no immediate syntactical dependencies between the components of such fragments.[1] However, we will consider only word combinations containing content (i.e., non-functional) words; i.e., our discussion does *not* apply to the examples like (2) or (3).

We are not aware of any generally accepted definition of a content word, so we will define them as nouns, non-auxiliary verbs, adjectives, and adverbs. We do not consider any prepositions content words, in spite of that they can be either content (*por la costa* 'by the coast', *para niños* 'for children', *después de la guerra* 'after the war') or purely functional (*entrar en casa* 'enter the house'). Personal pronouns are also functional since they have no semantics of their own.

For simplicity, we will consider word combinations with only two content words, i.e., we study, say, the combinations *campo de temperatura* 'temperature field' and *temperatura uniforme* 'uniform temperature' but not *campo de temperatura uniforme* 'uniform temperature field'. In particular, we do consider two-component combinations like *dar* (una) *rosa* (*a alguien*) 'give (a) rose (to smbd.)' with an omitted obligatory valence, i.e., *dar* → *rosa* is a valid word combination.

Expanding to some degree the repertoire of the objects of our study from the syntactical viewpoint, we will consider as word combinations those fragments whose dependency trees consist of two content words linked through functional words (i.e., prepositions or auxiliary verbs) and/or obligatory functional words depending on the content words (these are, e.g., articles of nouns). Thus, we consider as word combinations not only the examples like (1) but also

(4) *entrar* → *en* → (*la*) *casa*
'enter (the) house',
(*el*) *enemigo* ← *ha* → *sido* → *derrotado*
'(the) enemy has been worsted',
actuar → *a_pesar_de* → *circunstancias*
'to act in_spite_of the circumstances',

where the content words are underlined, word sequences forming indivisible entities are shown using the underscore sign, and the arrows stand for syntactic dependencies.

Mel'čuk (2001) identifies three classes of word combinations of types as in the examples (1) and (4):

- ***Complete phrasemes***[2] are stable word combinations whose meaning cannot be derived from the meaning of the components: *marchar sobre ruedas* 'advance very successfully', *ponerse como jitomate* 'become very red'.
- ***Semiprasemes*** are stable word combinations whose meaning includes the meaning

[1] Some authors include such fragments in collocations while we would call them n-grams

[2] They are also called idioms.

of one component, whereas the second component is used not in its primary meaning. Semiphrasemes are also referred to as *collocations*[3] or word combinations with *lexical functions* (LF) (Mel'čuk 2001, Mel'čuk and Zholkovsky 1988). Examples of LFs are: *prestar atención* 'pay attention', *lanzarse al ataque* 'open an attack', *té cargado* 'strong tea', *compromiso falso* 'vain promise', *propuesta concierne* 'proposal concerns'.

- ***Free combinations*** are all other word combinations, i.e., those ones whose meaning is merely composed of the meanings of the components: *ver el bosque* 'to see the forest', *vestido blanco* 'white dress', *persona extraña* 'strange person'.

In this paper we are interested in all word combinations including the free ones. Indeed, each modern linguistic application (text editing, learning foreign languages, automatic analysis or generation of texts, machine translation, etc.) requires explicit specification of all features of lexemes in the given language. Word combinations are a part of such knowledge. Both a human user (for text editing or language learning) and a program (for text analyzing, generation, or translation) need to know word combinations because:

- In text generation, to select a correct and idiomatic word combination it is important to know the restrictions the language imposes on combinatorial properties of words.
- In text analysis, only relying on syntactic context high quality word sense disambiguation can be achieved.

For these purposes, the word combinations are to be stored in the dictionary for each lexeme, along with all other data about the lexeme. The reason for their explicit representation in the dictionary is that, as one can easily see in the examples given above (e.g., *prestar atención* 'pay attention'), word combinations are highly language dependent and thus cannot be predicted or logically deduced even if the user knows the corresponding combinations for another language.

Both for a human and machine, the modern dictionary have electronic form. The type of the user (human versus computer) determines only the interface and the degree of formalization of this information, cf. also (Bolshakov *et al.* 1999).

Since technology has removed storage limitations for linguistic data, infinite possibilities have arisen to satisfy the demand on word combinations representation. Hence, types of word combinations should be revised and reinventoried in order to decide what types can enrich existing collections.

2 *All Phrasemes are Needed Unconditionally*

What word combinations are needed unconditionally was has been specified long ago in the Meaning ⇔ Text Theory (MTT) (Mel'čuk 2001, Mel'čuk and Zholkovsky 1988): these are all types of phrasemes. In the twentieth century, complete phrasemes and (partially) semiphrasemes were included in educational and academic monolingual dictionaries as well as translation dictionaries. In recent 25 years, fragments of Explanatory Combinatorial Dictionaries (ECDs) have been compiled for Russian, French, and English according to the MTT recommendations. These dictionaries describe to the depth from several hundreds to several thousands of lexemes each. Nevertheless, they do not cover any language completely, since this would require full description of more than 100,000 lexical units.

Only very specialized groups of experts can expand the fragments of already existing ECDs or create new ones. We are aware of only two such groups, for Russian and French, while the emergence of strong such groups for English is doubtful. According to the modern linguistic mainstream's viewpoint, the textual links between words are determined by their co-participation in the same constituent. This does not favor the description of word combinations. As a result, word combination acquisition was reduced to the search of co-occurrences in windows moving along the text (Sidorov 1999, Smadja 1991).

On purely practical grounds, to leave dictionaries complied by traditional methods so

[3] There is no general consensus about the term *collocation*. Many researchers consider it merely a co-occurrence of words within a short span of text (Smadja 1991). We, however, use this term only in connection with immediate syntactic dependencies between the combined words. In this sense, the informal treatment of collocations in (Calzolari and Bindi 1990, Satoshi Sekine *et al.* 1992) is just what we mean.

poor in word combinations seems inadequate. A special stress at LFs is topical now, even without their strict codification according to the MTT rules. One can consider the paper-form dictionary (Benson *et al.* 1989) as an example of a non-codified ECD for English. Small in size, this dictionary is quite popular, especially among non-native speakers.

3 Nonstandard Lexical Functions are not yet Studied

Igor Mel'čuk defines a lexical function (LF) as a relation f between lexemes L such that:

1. There exists nearly the same correspondence between meanings '$f(L)$' or '$f(L)$ + L' and 'L' for a number of different lexemes L. E.g., *té + cargado* : *cargado* ≈ *apoyo + incondicional* : *apoyo* ≈ *gran + especialista* : *especialista*, etc. 'strong tea' : 'tea' ≈ 'steady encouragement' : 'encouragement' ≈ 'great specialist' : 'specialist'.
2. For a number of different L_1 and L_2, it holds $f(L_1) \neq f(L_2)$, as in the above examples.

Among LFs, there exist so-called *standard* ones. They satisfy two additional conditions: the number of different arguments of f and the number of its different values are rather large. For *non-standard* LFs, one or both of these conditions do not hold.

Standard LFs are defined, listed, and studied rather well (Wanner 1996). As to non-standard LFs, not much more than the definition is known about them. Mel'čuk supposes there are tens of thousands of non-standard LFs in any language. However, only two examples have been discussed in literature, both for Russian:

- the LF expressing brown color for eyes, hairs, horses, and all other brown things, and
- the one expressing the meaning 'of rye flour of dark color' differently for bread and all other baked wares.

We propose a third example: the meaning 'highly situated in an official hierarchy' applicable to officials without directly indicated position. In Russian it is expressed as *vysokij* lit. 'high' for *gost'* '*guest*', *vysokopostavlennyj* lit. 'of high rank' or *krupnyj* lit. 'big' for *chinovnik* 'official', *bol'shaja* 'big' for *shishka* 'bigwig', and only *vysokopostavlennyj* lit. 'of high rank' for other relevant persons. A Spanish example of this lexical function is *alto funcionario* 'high official'; note the difference with the free word combination *funcionario alto* 'tall official'.

So rare examples of non-standard LFs discovered during MTT's 30-year history contrast to the announced abundance. The problem is not explored and seems very difficult. To distinguish nonstandard LFs from free combinations is really too hard. We have collected more than 660 modificative and attributive expressions for the English word *person*, but did not find among them any complete phrasemes (except *nether person*) or any LFs.

Since the average user of lingware systems cannot distinguish non-standard LFs from other word combinations, it is hardly necessary to distinguish them within the applied systems, too. Nevertheless, the newly acquired and not yet classified combinations proved to be selective by their semantics. For example, from the modifiers for English word *person,* the majority is applicable to *human*, *man*, and *woman*; some of them are combinable with *lad*, *guy*, *boy*, *teeny*, *lass*, *girl*, etc., but this list is rather short. For other semantic groups of English content words, the corresponding modifiers are quite different.

4 Terminological Combinations are Language-Specific

Generally, terminological word combinations heavily depend on language. As an example, consider two equivalent terms: Spanish *recubrimiento* and Russian *pokrytie* 'covering'. Both terms easily form numerous hyponymous terms through adjoining syntactically dependent adjectives or nouns: *recubrimiento anticorrosivo, bituminoso, galvánico, metálico, protector*, etc. 'anticorrosive, bituminous, galvanic, metal, protective covering'. Comparing some of these modifiers in the two languages, one can find their literal translation sometimes impossible; see Table 1.

Many similar examples can be given. This means that all modificative (and attributive) expressions combinable with terms affluent in modifiers should be represented in specialized dictionaries with the utmost possible completeness. As to general-purpose dictionaries, special terminology is to be reflected in them, too, because:

1. Terminology of various fields of science and technology constantly penetrates to the everyday life. Thus, even non-specialists

Meaning	Russian term	Literal Spanish translation	Correct Spanish term
obtained by dipping into hot substance	*goryachee*	*caliente*	*en caliente*
sound-absorbing	*zvukoizolatsionnoye*	*aislante al sonoro*	*antisonoro*
admitting resilient deformation	*uprugoe*	*elástico*	*flexible*
of medium reflecting power	*polumatovoe*	*semimate*	*semibrillante*
ultimate	*vneshnee*	*externo*	*de terminación*
resistent to fall-outs	*atmosferostoykoe*	*resistente a la atmósfera*	*resistente a la intemperie*

Table 1: Comparison of terms in the two languages.

need to know with what adjectives and verbs can be used the new or renewed words like Spanish *fichero* / *archivo* 'file'[4] *or servidor* 'server'.

2. There are numerous groups of lexical homonyms one of which has an everyday-life meaning while another is a term. Modifier sets combinable with such homonyms usually do not intersect and thus are a good tool for word sense disambiguation. For example, English $coat_1$ 'dress' is combinable with *fur, pilot, sack, storm,* etc., whereas $coat_2$ 'superimposed layer' with *finish, priming, scratch*, etc.

It is still not convenient to combine general-purpose and special dictionaries, but all those words and word combinations that have been derived from, or proved to be homonymous to, scientific and technical terms should be considered for general-purpose dictionaries.

5 *Free Word Combinations are to be Acquired Selectively*

In the 1990s, the authors of this paper have created a database of ca. 800,000 Russian word combinations of various types and classes, including free combinations (Bolshakov 1994, Bolshakov and Gelbukh 2001). For phrasemes of various types, no special rules were necessary. As to free word combinations, the following mild heuristic rules were applied for their acquiring:

1. If a given word combination is not yet represented in this dictionary but is found in some another dictionary (bilingual, educational, academic, or specialized) then this combination should be included unconditionally.
2. If a new combination is encountered in a text (in the corpus) then it is included if the linguist intuitively expects it to appear in other texts, with the exceptions described below in this section.

According to the second rule, such English combinations as *to loom large, reshuffled cabinet, economic malaise, stalemated talks* (examples found by the authors in Internet) are to be included. Other such combinations can be doubtful to a non-native speaker. In any case, the ultimate decision on each such combination found in the corpus (or in Internet) must be made by a well-read native speaker.

However, not all such combinations that should ultimately be presented to the user of the dictionary system need to be included explicitly. A vast majority of them can be generated by the system automatically in runtime using the information explicitly present in its internal dictionary, as it is done in CrossLexica system (Gelbukh *et al.* 2002), which uses, among others, the following rules:

- From the combinations with absolute synonyms, only those with the "main" synonym are included. The combinations with any other member of the synset are constructed automatically in runtime using the available combinations and synonymy links. E.g., if in the pair *dwellings* / *residences* the first word is considered the main synonym and such combination as *search* (*one's*) *dwellings* is explicitly stored in the dictionary then *search* (*one's*) *residences* is constructed automatically. Though true absolute synonyms are rare, we consider equivalent to them also the abbreviations like *EE.UU. = Estados Unidos de América* 'USA' = 'United States of America' or *Coca* = *Coca-Cola*, orthographical variants like *postgrado* = *posgrado* 'post-graduate studies' or Russian *tunnel'* = *tonnel'*

[4] The Spanish term depends on country.

'tunnel', and dialectal variants like *fichero* = *archivo* 'file', *gafas* = *anteojos* 'spectacles' or English *favor* = *favour*.

- The combinations of the kind *tomar* 'drink' *Mirinda* / *Fanta* / *7Up*, etc. or viajar por 'to travel across' *Bolivia* / *Honduras* / *Nepal*, etc. are not included explicitly. These species of drinks and countries, as well as of some fruits, berries, etc., are rare in texts and their combinations have no peculiarities, so that all combinations including their generic words *refresco* 'drink', *país* 'country', etc. are easily transformed to combinations with the species. Of course, this makes no obstacle for acquiring large and specific word combination groups for such peculiar (in what concerns word combinations) countries as USA or Afghanistan, or for some peculiar drinks.
- Combinations like English *tears of the inspector, dress of the probationer, president's sperm*, etc. that combine parts or secrets of human body or standard human reactions or clothes with nouns characterizing individuals of narrow human groups are not included explicitly. They are constructed in runtime using the genus–species hierarchies with the WordNet-type chains (Vossen 2000) such as *inspector* < *specialist* < *adult* < *man* or *woman* and more common combinations like *tears of the man*, *dress of the man*, *man's sperm*, etc.
- The two-component combinations of the type *to drink a barrel* (*of* ...), *to nail a piece* (*of* ...), *to commit a series* (*of* ...), etc. are not included explicitly. In runtime, the three-component combinations *to drink a barrel of water, to nail a piece of board, to commit a series of murders* can be generated using the combinations *to drink water, to nail a board, to commit a murder* and the semantic species–genus links: *barrel of beer* ⇐ *beer*, *piece of board* ⇐ *board, grip of a blockade* ⇐ *blockade* (these links are also lexical functions of a different, paradigmatic type, see Section 3).

As we have already explained, that all such combinations are transparently presented to the user as if they were present in the system's dictionary. However, their on-the-fly generation has the advantage of (a) reducing the manual work of the linguist necessary to compile the dictionary and (b) assuring that no such combination will be forgot. We do not even mention the disk space and memory requirements gain.

The amount of the new collocations generated by such rules depends on the desired precision vs. recall balance. Currently the Russian CrossLexica dictionary includes 1.4 million word combinations (of them, 595 thousand attributive) and 1.0 million unilateral semantic links (of them, 804 thousand semantic derivatives) used for generation; all these word combinations and semantic links were collected manually. The current implementation automatically generates about 1 million new word combinations with at least[5] 95% precision (correctness rate). The output of the generation module can be manually verified and the erroneous word combinations put in the stop list to prevent them from being presented to the user in the future.

6 Statistical Approach is Limited in its Scope

Subjectivity of the rules discussed above is obvious. Additionally, purely manual methods of word combination acquisition seemingly bring the problem out of the scope of computation. Let us imagine purely statistical methods applied to a large text corpus for this purpose.

Let the language contains N = 100,000 lexical units with the frequency of occurrence corresponding to the Zipf law

$$P(k) = \frac{1}{\ln N}\frac{1}{k},$$

where k = 1, 2, etc. is the statistical rank of a unit. We will not distinguish content and functional words.

To roughly estimate the upper bound C_{up} of the necessary corpus size, suppose that any two words with ranks $k \neq m$ form a word combination co-occurring statistically independent. Taking the words with the middle ranks $k \approx N / 2$, $m \approx N / 2$, we have

$$C_{\text{upper}} \approx \left(\frac{N \ln N}{2}\right)^2 \approx 3.3 \times 10^{11} \text{ words.}$$

To estimate the lower bound, we suppose that each word generates a limited number M of combinations of a given type, M being about 14

[5] Except for the module of generation of attributive collocations, which gave higher error rate and is now under re-implementation.

(cf. below on CrossLexica). Supposing that a given context type occurs only in half of cases and taking the Zipf law within the group, we have

$$C_{\text{lower}} \approx (N \ln N)\,(M \ln M) \approx 4.3 \times 10^6 \text{ words}.$$

A wordform with a delimiting space covers about 10 bytes in texts in most of European languages; a standard encyclopedic volume contains ca. 6 MB, while a standard CD up to 600 MB. Thus, the upper bound of the corpus size is 3300 GB, i.e. 550,000 encyclopedic volumes or 5,500 CDs, whereas the lower bound is 43 MB, i.e. only 7 volumes or 1/14 of a CD.

The scatter of the estimates is immense, but even if C_{low} is nearer to the truth, a formidable obstacle should be taken into account: the corpus should be utmost polythematic, in order to include the total vocabulary mentioned, and should be tagged by syntactic dependencies.

Possessing such facilities, we could apply the strict statistical criterion of representative co-occurrence of any two content words in the form of mutual information (Manning and Schütze 1999):

$$Q = \log \frac{P(k,m)}{P(k)P(m)},$$

where $P(k,m)$ is the empirical frequency of the co-occurrence and $P(k)$ and $P(m)$ the frequencies for the combined words taken apart. Pairs of words systematically forming word combinations result in the Q significantly exceeding zero, while for the rest, Q is about or below zero.

Though we did not have at our disposal any corpus with the properties described above, we have applied statistical methods for looking for new combinations of the most frequent words (initial few thousands in ranks). However, the common words are frequently supplied in dictionaries with numerous combinations, so that these results often proved to be redundant.

Our conclusion based on the experience with CrossLexica dictionary (Bolshakov 1994, Bolshakov and Gelbukh 2001) is as follows: one cannot get rid of manual methods of acquiring new word combinations. Even using statistical methods we are doomed to human post-editing of the results. Hence, subjective rules for including free combinations remain inevitable.

With such, at least partially manual, acquisition, we initially expected that the mean combinatorial productivity of a word for any type of combinations, i.e., the mean number of modificative-attributive expressions for a noun, the mean number of predicates for a subject noun, the mean number of various valent and circumstantial complements for a verb, etc., would grow infinitely. The practice of CrossLexica showed that this was not the case.

For content words of any part of speech participating in all explored types of word combinations, the combinatorial productivity have stabilized at 9 to 14 combinations of each type, and these mean values had no trend to increment during the last seven years of the acquirement, whereas the total number of collected combinations has increased twice during this period. With the growth of the corpus length looked through, only the number of words obtaining the full combinational characterization has been steadily increasing. This growth cannot be avoided if one wants to guarantee good (at least 65%) coverage of an arbitrary unspecialized text. This stabilization can be explained rationally. On numerous examples it was ascertained that word combinations considered as free are significantly restrained by mutual semantic selectivity of the combined words.

7 Conclusions

The collection of word combinations to be stored in general-purpose electronic dictionaries should be broadened as much as possible. Besides all complete phrasemes and combinations with lexical functions, it makes sense to include the word combinations with non-standard lexical functions, terminological combinations, and free combinations satisfying rather liberal rules.

These rules do not require classifying combinations beforehand and do not cause any significant growth of combinatorial productivity of content words.

To our experience, one cannot get rid of manual methods of acquiring of combinations or at least post-editing automatically acquired ones. On the other hand, with the heuristics discussed in this paper, a vast majority of (potentially) existing word combinations can be generated automatically in runtime using the combinations explicitly stored in the dictionary and WordNet-like relationships.

References

Benson, M., E. Benson, and R. Ilson. 1989. *The BBI Combinatory Dictionary of English.* Amsterdam / Philadelphia: John Benjamin.

Bolshakov, I. A. 1994. Multifunction thesaurus for Russian word processing. *Proc. of 4th Conference on Applied Natural language Processing*, Stuttgart, October 13-15, p. 200–202.

Bolshakov, I. A., A. Gelbukh. 2001. A Very Large Database of Collocations and Semantic Links. Mokrane Bouzeghoub et al. (eds.) *Natural Language Processing and Information Systems* (NLDB-2000), *Lecture Notes in Computer Science* 1959: 103–114, Springer.

Bolshakov, I. A., A. Gelbukh, S. N. Galicia-Haro. 1999. Electronic Dictionaries: For Both Humans and Computers. V. Matousek et al. (eds.). *Text, Speech and Dialog* (TSD'99), *Lecture Notes in Artificial Intelligence* 1692: 365–368, Springer.

Calzolari, N., R. Bindi. 1990. Acquisition of Lexical Information from a Large Textual Italian Corpus. *Proc. of COLING-90*, Helsinki.

Gelbukh, Alexander, Grigori Sidorov, and Igor Bolshakov. 2002. Coherence Maintenance in Man-Machine Dialogue with Ellipses. J. *Computación y Sistemas, revista iberoamericana de computación*.

Manning, Ch. D., H. Schütze. 1999. *Foundations of Statistical Natural language processing*. Cambridge, Massachusetts / London: The MIT Press.

Mel'čuk, Igor. 2001. Fraseología y diccionario en la lingüística moderna. In: I. Uzcanga Vivar *et al.* (eds.) *Presencia y renovación de la lingüística francesa*. Salamanca: Ediciones Universidad, p. 267–310.

Mel'čuk, I., A. Zholkovsky. 1988. The explanatory combinatorial dictionary. In: M. Evens (ed.) *Relational models of lexicon*. Cambridge, England: Cambridge University Press, p. 41-74.

Satoshi Sekine *et al.* 1992. Automatic Learning for Semantic Collocation. *Proc. of 3rd Conf. Applied Natural Language Processing*, Trento, Italy, p. 104–110.

Sidorov, G. 1999. Methods of analysis of combinability of words in Russian (in Russian). *Taal en cultur (Language and culture). Materials of the conference "Belgium–Holland–Russia,"* Maastricht, Holland, and Moscow, Russia, p. 294–302.

Smadja, F. 1991. Retreiving collocations from text: Xtract. *Computational Linguistics* 19 (1): 143–177.

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

Wanner, Leo (ed.). 1996. *Lexical Functions in Lexicography and Natural Language Processing*. Studies in Language Companion Series, ser. 31. Amsterdam / Philadelphia: John Benjamin.