Word Combinations as an Important Part of Modern Electronic Dictionaries*

Igor A. Bolshakov and Alexander Gelbukh

Center for Computing Research (CIC), National Polytechnic Institute (IPN),
Mexico City, Mexico
{igor,gelbukh}@cic.ipn.mx

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. I. 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.

1 Introduction

A word combination is usually supposed to be two or more syntactically connected words. In this sense, we may consider as word combinations not only

pay attention, proposal concerns, collocation set, beautiful flower, stone wall... (1)

but also *have been, is growing* or *she went*, but not contiguous text fragments like *of the*, *be a* or *and to*, since there are no immediate syntactical dependencies between the components of these fragments.

We will consider only word combinations containing content (non-functional) words. We do not know any strict definition of a content word, so we will define them as nouns, non-auxiliary verbs, adjectives, and adverbs. Prepositions, which can be both content (*<u>by</u> the riverside*, *<u>for</u> children*, *<u>after</u> the war*) and purely functional (*enter <u>into</u> the room*), are always considered functional. 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 *differential equations* and *partial derivative* but not *differential equations in partial derivatives*. However, we do consider two-component combinations like *give* (*a*) *rose* (*to somebody*) with an omitted third component in the role of an obligatory valence.

Expanding to some degree the repertoire of the objects of our study from the syntactical viewpoint, we will consider as word combinations those linear fragments

---

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

whose dependency trees consist of two content words linked through functional words (these are mainly prepositions) 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 (1) but also

$$\begin{array}{c} \underline{\textit{come}} \rightarrow \textit{into} \rightarrow (\textit{the})\ \textit{room}, \\ (\textit{the})\ \underline{\textit{enemy}} \leftarrow \textit{will} \rightarrow \textit{be} \rightarrow \underline{\textit{defeated}} \\ \underline{\textit{act}} \rightarrow \textit{in\_spite\_of} \rightarrow \underline{\textit{circumstances}}, \end{array} \qquad (2)$$

where the content words are underlined and word sequences forming indivisible entities are shown using the underscore sign.

Mel'čuk [7] states three classes for word combinations of types (1) and (2):

- ***Complete phrasemes*** are stable word combinations whose meaning cannot be derived from the meaning of the components: *to kick the bucket, red herring, to be for the birds...*
- ***Semiprasemes*** are stable word combinations whose meaning includes the meaning of one component, whereas the second component is used not in its primary meaning. Semiphrasemes are also referred to as ***collocations*** [1] or word combinations with ***lexical functions*** (LF) [7, 8]. Here are several LFs: *to pay attention, to launch an attack, strong tea, false promise, proposal concerns...*
- ***Free combinations*** are all other word combinations, i.e., those ones whose meaning is merely composed from the meanings of the components: *to see the forest, white dress, strange person...*

In this work we are interested in all such word combinations, including free ones. Indeed, each modern linguistic application (text editing, learning foreign languages, automatic analysis or generation of texts, machine translation, etc.) requires storing all specific features of lexemes in language. Word combinations constitute a part of such knowledge.

Both a human user (for editing text or learning the language) and a program (for analyzing, generating, or translating text) need word combinations because:

- To select a correct and idiomatic word combination, it is important to know what restrictions language imposes on combinatorial properties of words.
- Only relying on syntactic context, high quality word sense disambiguation can be achieved.

For these purposes, the word combinations should be stored for each lexeme, along with all other data about it. Both for a human and machine, the modern dictionary has electronic form. The type of user (human *versus* computer) determines only the interface and the degree of formalization of this information, cf. also [4].

Since technology has removed storage limitations for linguistic data, infinite possibilities have arisen for satisfaction of the demand on word combinations. Hence, types of word combinations should be revised and re-inventoried in order to decide what types are to be included to enrich existing collections.

---

[1] There is no general consensus about the term *collocation*. Many researchers consider it merely as a co-occurrence of words within a short span of text [11]. We, however, use this term only in connection with immediate syntactic dependencies between the combined words. In this sense, the non-formal treatment of collocations in [5, 9] is just what we mean.

2 All Phrasemes are Needed Unconditionally

The question of which word combinations are needed unconditionally was answered long ago in the Meaning ⇔ Text Theory (MTT) by I. Mel'čuk [7, 8]: these are all types of phrasemes. It was shown that without knowing phrasemes idiomatic speech is impossible, as well as analysis and generation of correct texts by a computer that models linguistic activity.

In the 20-th century, complete phrasemes and (partially) semiprasemes were included in educational and academic monolingual dictionaries as well as translation dictionaries. In recent 25 years, fragments of Explanatory Combinatorial Dictionaries (ECDs) were compiled for Russian, French, and English according to the MTT recommendations. They 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 high-level and specialized groups of experts can expand the fragments of already existing ECDs or create new ones. We are aware of only two such groups, working on Russian and French, while the emergence of 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 constituents. This does not favor the description of word combinations. As a result, the word combination acquisition was reduced to the unprincipled search of co-occurrences in windows along the text [10, 11].

On purely practical grounds, to leave dictionaries complied by traditional methods so poor in word combinations seems inadequate. Since lexicographers never neglected complete phrasemes, a special stress at LFs is topical now, even if these combinations are left not codified according to the MTT (without the titles proposed by Mel'čuk and with a poor structure of a dictionary entry) . We can consider the paper-form dictionary [1] 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

I. 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 '$f(L)$' и 'L' for a number of different lexemes L.
2. For a number of different L_1 and L_2, it holds $f(L_1) \neq f(L_2)$.

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

Standard LFs are defined, listed, and studied rather well [13]. As to non-standard LFs, not much more than the definition is known about them. I. Mel'čuk supposes there are tens of thousand 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* for other relevant persons.

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 found among them any complete phrasemes (except *nether person)* or standard LFs. We failed to find any non-standard LFs either, though they surely exist within the group.

Since average language speaker (the user of applied linguistic systems, including computer-based dictionaries) cannot distinguish non-standard LFs among other word combinations, it is hardly necessary to distinguish them within the applied systems, too. The newly acquired word combinations (left unclassified up to unforeseeable future) are usually selective. By selectivity we mean here that words that seem to be values of some LF are semantically combinable with very limited number of supposed arguments. For example, from 660 modifiers for *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 another group of 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: English *coating* and Russian *pokrytie*. Both terms easily form numerous hyponymous terms through adjoining syntactically dependent adjectives or nouns: *anodic, corrosion-resistant, metal, oxide, protective... coating*. Comparing some of these modifiers in the two languages, one can find their literal translation sometimes impossible:

Meaning	**Russian term**	**Literal translation**	**Correct English term**
‘got through dipping into hot substance’	*goryachee*	*hot*	*hot-dip*
‘superimposed twice’	*dvusloynoe*	*double-layered*	*double-course*
‘admitting resilient deformation’	*uprugoe*	*elastic*	*flexible*
‘of medium reflecting power’	*polumatovoe*	*semimatted*	*semibrigth*
‘emitting fluorescent light’	*svetjascheesya*	*gleaming*	*fluorescent*
‘of paint or varnish’	*lakokrasochnoe*	*of varnish or paint*	*paint*
‘difficult to melt down’	*tugoplavkoe*	*difficultly melting*	*high-melting, refractory*

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 need to know with what adjectives and verbs the words *uranium, vibrio, file, site, server*... can be used.
2. There exist a great number of vocables (sets of lexical homonyms) with one homonym having an everyday-life meaning and another one a terminological meaning. Modifier sets combinable with such homonyms usually do not intersect and thus are considered a good tool for word sense disambiguation. For example, $coat_1$ 'dress' is combinable with *fur, polo, pilot, sack, storm, tent*, etc., whereas $coat_2$ 'superimposed layer' with *finish, priming, of paint, scratch*, etc.

It is still unreasonable to combine general-purpose and special dictionaries, but all those words and word combinations that have been acquired from, or turned to be homonymous to, scientific and technological terms should be considered for inclusion in general-purpose dictionaries.

5 Free Word Combinations are to be Acquired Heuristically

Taking into account the facilities of broadened computer storage, the authors have created in the 90's a database of ca. 800,000 Russian word combinations of various types and classes, including free combinations [2, 3]. For phrasemes of various types, no special rules were necessary. As to free word combinations, the following extremely mild heuristic rules were applied for their inclusion:

1. If a given word combination is not yet represented in our dictionary and we find it in another dictionary—educational, academic, or special—then this combination is included unconditionally.
2. If a new combination is encountered in a text, then it is included if the linguist expects it to appear in other texts, except for the following cases:
   - From the combinations with absolute synonyms, only those with the "main" synonym are included. The combinations with any other member of the synset are to be constructed by the program at runtime, using already available combinations and synonymy links. If, e.g., in the pair *dwellings* / *residences* the first word is considered the main synonym, the DB should store its combination such as *to search* (*one's*) *dwellings,* whereas *to search* (*one's*) *residences* is constructed in runtime. Though true absolute synonyms are rare, we consider as equivalent to them also the abbreviations like *USA* = *United States* = *United States of America* or *Coca* = *Coca-Cola*, orthographical variants like Russian *tunnel'* = *tonnel'* (both meaning 'tunnel'), and dialectal variants like American English *favor* = British English *favour*.
   - The combinations of the kind *to drink Mirinda* / *Fanta* / *7Up*, etc. or *to travel across Bolivia* / *Honduras* / *Nepal*, etc. are not included. These species of drinks, 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 *drink, country, fruits*, etc. are easily transformed to combinations with their 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*, *Russia*, or *Afghanistan*,

cerns word combinations) countries as *USA*, *Russia*, or *Afghanistan*, or for such peculiar drinks as *vodka*.

– Combinations like *tears of the inspector, dress of the probationer, president's sperm*, etc. that combine parts or secrets of human body, standard human reactions or clothes with nouns characterizing individuals of narrow human groups are not included. They can be constructed at runtime using the genus—species hierarchies with the WordNet-type chains [12] 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, to nail a piece, to break the grip, to commit a series*, ... are not included. At runtime, the three-component combinations *to drink a barrel of water, to nail a piece of board, to break the grip of a blockade, to commit a series of murders* can be composed, using the combinations *to drink water, to nail a board, to break the blockade, to commit a murder* and the semantic links of the species—genus type: *barrel of beer*—*beer*, *piece of board*—*board, grip of a blockade*—*blockade* (these links are also lexical functions, though of a different, paradigmatic type).

Concerning the second rule, such combinations as *to loom large, reshuffled cabinet, economic malaise, stalemated talks* (examples from Internet) are to be included. Meanwhile, some such combinations seem doubtful to a non-native speaker. In any case, the ultimate decision on whether a combination is to be included or not must be made by a well-read native speaker.

6 The Statistical Approach is Limited in its Scope

Subjectivity of the rules given above is obvious. Additionally, purely manual methods of word combination acquisition seemingly bring the problem out of the scope of computational linguistics. Let us imagine, though, applying purely statistical methods 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 $P(k) = (\ln N)^{-1}\ 1/k$, where k = 1, 2, 3, ... 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 needed corpus size, we suppose that any two words with ranks $k \neq m$ form a word combination co-occurring statistically independent. Taking the words with the ranks in the middle of the frequency list $k \approx N / 2$, $m \approx N / 2$, we have $C_{up} \approx (N \ln N / 2)^2 \approx 3.3\times10^{11}$ words.

To estimate the lower bound, we suppose that a word generates a limited number M of combinations of a given type, and M is about 14 (cf. below the experience of CrossLexica). Supposing that a given context type occurs only in a half of cases and taking the Zipf law within the group, we have $C_{low} \approx (N \ln N)(M \ln M) \approx 4.3\times10^6$ words.

A word-form with a delimiting space covers about 10 bytes in texts in 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 and we do not know which one is nearer to the truth. A formidable obstacle that should be taken into account is that the corpus should be utmost polythematic, in order to include the total vocabulary mentioned, and should be tagged by syntactic dependencies between words.

Possessing such facilities, we could apply the strict statistical criterion of representative co-occurrence of any two content words. The criterion can be expressed in terms of mutual information [6]:

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

where $P(k, m)$ is an empirical frequency of the co-occurrence and $P(k)$ and $P(m)$ are frequencies for the combined words calculated independently. 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 these facilities were unreachable for us, we had applied statistical methods for searching 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 our computerized results were often excessive.

Our pessimistic conclusion based on the experience with CrossLexica dictionary [2, 3] is as follows: we 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 indispensable.

With such, at least partially manual, acquisition, it was expected that the mean combinatorial productivity of a word for any types 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., will grow infinitely. The practice of creation of our database did not affirm this anticipation.

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 searched through, only the number of words obtaining the full characterization as to their word combinations has been steadily increasing. This growth cannot be avoided if we want to guarantee rather high (at least 65%) coverage of an arbitrary unspecialized text.

This stabilization can be explained rationally. On numerous examples we have ascertained that word combinations considered as free are significantly restrained by mutual semantic selectivity of combinable 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 a significant growth of combinatorial productivity of content words because of their mutual semantic selectivity.

To our experience, one cannot get rid of manual methods of acquiring of combinations or at least post-editing automatically acquired ones.

References

1. Benson, M., E. Benson, and R. Ilson. *The BBI Combinatory Dictionary of English*. Amsterdam / Philadelphia: John Benjamin, 1989.
2. Bolshakov, I. A. *Multifunction thesaurus for Russian word processing*. Proceedings of 4th Conference on Applied Natural language Processing, Stuttgart, October 13-15, 1994, p. 200-202.
3. Bolshakov, I. A., A. Gelbukh. *A Very Large Database of Collocations and Semantic Links*. In: Mokrane Bouzeghoub et al. (eds.) *Natural Language Processing and Information Systems.* 5th International Conference on Applications NLDB-2000, Versailles, France, June 2000. Lecture Notes in Computer Science No. 1959, Springer, 2001, p. 103-114.
4. Bolshakov, I. A., A. Gelbukh, S. N. Galicia-Haro. *Electronic Dictionaries: For Both Humans and Computers.* In: V. Matousek et al. (eds.). *Text, Speech, and Dialog*. Second International Workshop TSD'99. Lecture Notes in Artificial Intelligence, No. 1692, Springer, 1999, p. 365-368.
5. Calzolari, N., R. Bindi. *Acquisition of Lexical Information from a Large Textual Italian Corpus*. Proc. of COLING-90, Helsinki, 1990.
6. Manning, Ch. D., H. Schütze. *Foundations of Statistical Natural language processing*. Cambridge, Massachusetts / London: The MIT Press, 1999.
7. Mel'čuk, Igor. *Fraseología y diccionario en la lingüística moderna.* In: I. Uzcanga Vivar et al. (eds.) *Presencia y renovación de la lingüística francesa.* Salamanca: Ediciones Universidad, 2001, p. 267-310.
8. Mel'čuk, I., A. Zholkovsky. *The explanatory combinatorial dictionary*. In: M. Evens (ed.) *Relational models of lexicon*. Cambridge, England: Cambridge University Press, 1988, p. 41-74.
9. Satoshi Sekine et al. *Automatic Learning for Semantic Collocation.* Proc. 3rd Conf. Applied Natural Language Processing, Trento, Italy, 1992, p. 104-110.
10. Sidorov, G. *Methods of analysis of combinability of words in Russian*. (in Russian) In: Taal en cultur (Language and culture). Materials of the conference "Belgium–Holland–Russia," Maastricht, Holland and Moscow, Russia, 1999, p. 294-302.
11. Smadja, F. *Retreiving collocations from text: Xtract*. Computational Linguistics. Vol. 19, No. 1, 1991, p. 143-177.
12. Vossen, P. (ed.). *EuroWordNet General Document*. Vers. 3 final. 2000, www.hum.uva.nl/~ewn.
13. Wanner, Leo (ed.). *Lexical Functions in Lexicography and Natural Language Processing*. Studies in Language Companion Series, ser. 31. Amsterdam / Philadelphia: John Benjamin, 1996.