A Very Large Dictionary with Paradigmatic, Syntagmatic, and Paronymic Links between Entries

Igor A. Bolshakov and **Alexander Gelbukh**
Center for Computing Research, National Polytechnic Institute,
07738 DF, Mexico
igor@cic.ipn.mx, gelbukh@gelbukh.com; www.gelbukh.com

Abstract

A very large Russian dictionary is described. It contains currently 3.6 million links between its 120,000 entries. The links are syntagmatic (collocations), paradigmatic (WordNet-like), or paronymic (words similar in letters or in morphs). The entries of the dictionary are single- or multiwords belonging to four main POS. The entries represent so-called grammemes rather than lexemes: e.g., nouns are represented as singular and plural; verbs are split into ‘finite forms + infinitive’, ‘participle’, and ‘gerund’. The multiword entries in turn can be collocations—idiomatic free—whose parts are also entries of the same dictionary.

1 Introduction

The entries of modern dictionaries are usually lexemes, and semantic relations between lexemes are only morphological derivates of the entry. E.g., Webster’s Universal College Dictionary of English gives with *manufacture* also *manufacturable, manufactural,* and *manufacturer*, having no entry.

A rather specialized exception is WordNet (Fellbaum, 1998). It gives paradigmatic semantic links: synonymous (common gloss), antonymous (common presupposed meaning, opposite assertion), XPOS links (same meaning, different POS), etc.

Other exceptions are BBI (Benson *et al*., 1986) and much vaster OCDSE (OCDSE, 2003), both existing only in printed form. They give collocations, i.e. stable and idiomatic word combinations connected by both syntactic (immediate or through auxiliary words) and semantic links, specifically of syntagmatic type (e.g., a verb and its valence fillers, a noun with its modifiers, etc.).

The entries are usually single words, whereas general linguistics also suggests multiwords, which may be both inseparable idiomatic concepts like *hot dog* or *point of view* and closely tied and frequently used terms like *routs of communications*.

This paper shortly outlines a very large electronic dictionary with the following main features:

- It adheres to the ***grammemic principle*** of selection of entries. Grammemes are subsets of morphological paradigms of lexemes, i.e. they are between lexemes and wordforms. E.g., noun lexemes are represented by two grammemes: singular and plural; verb lexemes are split into finite forms + infinitive, participle, and gerund.
- Its entries are only content (semantically autonomous) single and multiwords; the latter can be additionally represented by links between their components (this can be called ***decomposition principle*** of entry formation). Auxiliary words are stored as an auxiliary part of the dictionary, e.g., as a list of prepositions.
- It includes links of three vast classes: semantic ***syntagmatic*** (as in OCDSE), semantic ***paradigmatic*** (as in WordNet), and ***paronymic***. The latter class is new in dictionaries: it connects entries similar in letters (e.g., *sigh* vs. *sign*) or in morphs (e.g., *sensible* vs. *sensitive*).

The dictionary is the result of CrossLexica project (Bolshakov, 2004; Bolshakov and Gelbukh, 2000) and is mainly oriented to collocations.

Specifically, the objectives of this paper are:

- To describe possible options and features for the dictionary entries;
- To characterize types of the links: syntagmatic, paradigmatic, and paronymic;
- To prove language independence of the CrossLexica structure;
- To statistically characterize CrossLexica dictionary in its present state.

We use English examples for the illustrations.

2 Types and Features of Entries

Wordforms are grouped together into dictionary entries basing on several important features:

Parts of speech It can be any of the four main POS. The POS is defined according to the syntactic role: participles are considered adjectival; a prepositional phrase can be adjectival or adverbial, e.g., *in substance* is adjectival in *unconstitutionality in substance* (≈ ‘substantial’) and adverbial in *to verify in substance* (≈ ‘substantially’). We consider such two-functional entries homonymous.

Grammemes For a Russian noun, the singular and plural have their own collocations. Printed dictionaries mark this as *mainly plural* or the like. So we split the morphoparadigm of a noun to singular and plural, calling such sub-paradigms grammemes.

Based on their syntactic roles, we divide morphoparadigms of verbs into the grammemes of participles (adjectival), gerunds (adverbial), and personal forms plus infinitives (predicates).

Russian verbs have two aspects differing in their combinability: the perfect tends to collocate with singular nouns, the imperfect being indifferent to number; the perfect is usually modified with 'concentrated' adverbials like *suddenly, at once* or *straightway*, the imperfect preferring 'spread' adverbials like *gradually, continuously* or *repeatedly*. So we split verbs into aspectual grammemes.

Homonyms We consider various homonyms separately. Their combinatorial differences are especially useful for word sense disambiguation.

Idioms Idiomatic collocations like *point of view* are entries, since combinability of an idiom is always different from that of its head. Their components can, though, be also entries on their own.

Multiwords If a non-idiomatic multiword has a single-word synonym, we treat it as an entry, since its combinability differs from that of its head. E.g., Rus. *puti soobščenija* 'routes of communications' has a synonym *kommunikacii.* Cf. a similar problem in EuroWordNet (Vossen, 2000).

Absolute synonyms, abbreviations, and morphological variants Absolute synonyms (*sofa* = *settee*) are very rare in any language, but there are other types of equivalence: abbreviations (*United States of America = USA = United States*) and the so-called morphological variants (e.g., Rus. *nul'* = *nol'* 'zero' or *mučat'* = *mučit'* 'to torture'). Since all their collocations are the same, we store them as one entry, selecting one of them as a representative.

Paste-ups Many Russian noun-headed concepts are used in two equivalent forms: (1) a bigram consisting of a modifier with the stem S_1 plus its head noun with the stem S_2, or (2) a single noun containing the stems S_1 and S_2, or their initial parts, or only S_1: *električeskij tok* 'electrical current' = *elektrotok*; *fizičeskij fakul'tet* 'physical faculty' = *fizfak*; *komičeskij akter* 'comical actor' = *komik*. The number of the paste-ups grows, especially in the newswire and everyday speech, but in dictionaries they are scarce. Our dictionary stores about three thousand of them in both forms.

Compound pairs Russian has numerous stable pairs of nouns separated by a dash, usually with both parts declinable in parallel: *strana-učastnica* 'participant country', *letčik-ispytatel'* 'test pilot', *zavod-izgotovitel'* 'manufacturing plant'. A compound pair is considered an entry.

Coordinated pairs Dependency links within multiwords can be of stable coordinative type: *mother and father*, *safe and sound, sooner or later.* We consider such pairs as both collocations (with syntagmatic links) and separate entries. E.g., each bracketed item of the term [[[*probability*] [*theory*]] *and* [[*mathematical*] [*statistics*]]] is an entry.

Synonyms, hyperonyms/hyponyms, and antonyms These are semantically paradigmatic links. We take their participants as entries.

Proper names We consider as entries those names that are a part of everyday life and encyclopaedic knowledge: names of geographic objects, countries, famous persons, large organizations, etc. They are linked to their hyperonyms: *country*, *mountains*, *island*, *writer*, *organization,* etc.

Semantic derivates These are lexemes of any POS with same basic meaning, e.g., *to marry, marriage, bride, bridegroom,* and *matrimonial* (XPOS in WordNet). We take such words as entries.

Idiomaticity in general All complete idioms are included as collocations, e.g., *sest'* | *v galošu* 'to get | into a fix', lit. 'to sit | into a galosh'. In rarer cases of tripartite idioms the dichotomy was merely a practical step; e.g. in *byt'* | *bez carja v golove* 'to be stupid', lit. 'to be | without the Tsar in one's head', we regard the right part as a modifier. Two marks are used: ***idiom*** and ***possible idiom***, the latter for collocations with both figurative and direct senses, e.g., *sest' v lužu* means 'to get into a mess' or 'to sit down into a puddle'.

Usage marks ***Special, bookish*** or ***obsolete***: the use in writing is recommended if the meaning is clear to the writer; ***colloquial***: the use in official writing is not recommended; ***vulgar***: both written and oral use are prohibited; and ***incorrect***: used sometimes but contradicts language norms.

3 Types of Syntagmatic Links

We define a *collocation* as a syntactically connected and semantically compatible pair of content words, like *full-length dress, well expressed, to briefly expose, to pick up the knife* or *to listen to the radio* (collocation components are underlined).

Syntactical connectedness is understood as in dependency grammars (Mel'čuk, 1995) (maybe through an auxiliary word), not as co-occurrence (Bentivogli and Pianta, 2002); the components can be distant in the sentence. We consider collocations from absolutely free to purely idiomatic. The following are collocation types.

Modifiers These are modifying or attributive components: *great* ← *country*; *man* → *of letters*; *eat* → *quickly*; *enormously* ← *big*; *very* ← *well.*

Verbs with their subjects The subject is a specific dependent of a predicate verb: *soldier* ← *died*; *bus* ← *arrives*. A specifically Russian type of the subject-to-predicate link is a predicate containing the copula *byt'* 'to be' (omitted in Russian in present tense) and an adjectival in short form: *god* ← *zaveršen* (participle) 'the year is over'; *vek* ← *korotok* (adjective) 'the lifetime is short'.

Verbs with their noun complements Noun complements of a verb are all nouns dependent on the verb as direct, indirect, or prepositional object: *to read a book*; *to strive for peace*. We also consider as complements circumstantial phrases like *to travel by train*. A word can have several complements; each collocation reflects one of them, while the omission of other obligatory complement(s) is marked with the ellipsis: *to give ... to the boy.*

Nouns / adjectivals with their noun complements All POS can have noun complements, e.g., nouns *the capital of the country*, *the struggle against poverty*; adjectives *blind with rage*, *mentioned by the observer* or *going to the cinema*.

Verbs / nouns / adjectivals with their infinitive complements E.g. *to stop to talk* or *to permit to enter*; *permission to enter* or *cream to protect*; *forced to return* or *ready to appeal*.

Adverbials with their infinitive complements These are purely Russian collocations: *xolodno idti* 'it is cold to go', lit. 'coldly to go'; *reshiv* (gerund) *idti* 'after having decided to go'. They are possible only with some predicative adverbs or gerunds.

Adverbials with noun complements Purely Russian: *xolodno* (adverb) *bez pal'to* 'it is cold without a coat', lit. 'coldly without a coat'; *pobyvav* (gerund) *v centre* 'after visiting the center'.

Verbs / adjectivals with their adjectival complements E.g., *to remain silent* or *to consider... stupid*; *remaining silent* or *considering ... stupid*.

Coordinated pairs E.g. *mom and dad*, *safe and sound*, or *sooner or later*, cf. (Bolshakov *et al.*, 2003b) for details.

4 Types of Paradigmatic Links

These semantic (WordNet-like) links are:

Synonyms Unlike WordNet synsets, our synonymy groups have a dominant member and may include member(s) marked as its absolute synonyms. Synonyms can be periphrastic multiwords or even short definitions: *to help* ≈ *to give help*; *fall* ≈ *quick descent*; *suffocation* ≈ *lack of fresh air*. Non-absolute synonyms can be used for heuristic inferences of new collocations from those existing in the dictionary (Bolshakov and Gelbukh, 2002a).

Hyponyms *vs.* hyperonyms Hyperonyms are also used for such inferences.

Antonyms Together with standard antonyms (*good—bad, vanguard—rearguard*), we consider opposite notions: *missiles—antimissiles*.

Meronyms *vs.* holonyms E.g. *finger—hand*, *motor—car*.

Semantic derivates They connect parts of the morphoparadigms split into grammemes. Also, they describe the same idea from various aspects, thus compensating for the lack of glosses.

5 Types of Paronymic Links

The types of such links are as follows:

Literal paronyms They are at the distance of few editing operations (replacement, omission, insertion, permutation of adjacent letters) from each other. E.g., for *sign*: *sigh, sin, sing*. They are useful, e.g., to correct the malapropisms (Bolshakov and Gelbukh, 2003a).

Morphemic paronyms They are of the same POS and radix but have different prefixes and/or suffixes, e.g., *sens-ation-al, sens-ible, in-sens-ible, sens-itive, sens-less, sens-ual*. Foreigners' malapropisms are often confusion of morphemic paronyms, so that we can immediately propose candidates for correcting such errors.

Auxiliary parts of CrossLexica is a Russian-English-Russian dictionary (e.g., by two English words, the user can find a fluent Russian collocation), and a generator of all morphological forms.

6 Interlingual Structural Universality

The system operates with two main data structures: a list of entries and a set of links between them. An entry contains a list of its morphological categories. This structure is language-independent.

The specific links between entries can, however, be language specific. Let us outline grammatical peculiarities of Russian that influence these links.

Nouns and adjectivals declinable In English this problem does not exist.

Too few tenses Russian verbs have only three tenses, whereas English has many.

No articles For other languages, it is important to specify the forms of articles in collocations.

Nouns cannot modify nouns In English the collocations like *book review* are quite common. A special attributive type of syntagmatic links should be introduced for such English collocations.

Thus the Cross-Lexica structure is (almost) linguistically universal.

7 CrossLexica Statistics and Some Discussion

As of April 2004, the dictionary contains more than 120,000 entries. Collocations are divided into three classes: primary, secondary, and inferred.

The primary collocations are collected manually. The secondary collocations result from automatic morphological transformations of the primary ones. For example, verbs with their noun complements are transformed into adjectivals with their noun complements, e.g., *to participate in the meeting* gives *participating in the meeting*.

Table 1 shows the statistics of the collocations.

Type of collocations	Primary	Secondary
Words with modifiers	281,000	99,500
Verbs + subjects	106,400	27,500
Verbs + noun complements	204,900	
Nouns + noun complements	129,000	
Adjectivals w/ noun compl.	13,800	173,300
Adverbials w/ noun compl.	100	148,900
Other types	29,100	
Total	764,300	449,200

Table 1: Statistics of collocations

The inferences are performed with constraints (Bolshakov and Gelbukh, 2002a), e.g., the source collocation cannot be an idiom, to avoid the inference like (*hot dog*)$_{\text{idiom}}$ & (*poodle* IS_A *dog*) → *(*hot poodle*). The total of the inferred collocations never exceeded 6 to 8% of the primaries and is declining because the rare species are getting a full description within the primaries.

In Table 2, other link statistics are given. All links are counted, e.g., *n* antonyms pairs give 2*n* unilateral links, and a group of *n* synonyms gives $n(n{-}1)/2$. The total is more than 1.2 million. Thus, the total of links of the three classes is 3.6 million.

No.	Type of links	Amt. of links
1	Semantic derivates	821,600
2	Synonyms	226,200
3	Meronyms *vs*. holonyms	20,800
4	Hyponyms *vs*. hyperonyms	13,100
5	Antonyms	10,500
6	Morphemic paronyms	86,600
7	Literal paronyms	24,200
	Total	1,203,000

Table 2: Statistics of paradigmatic/paronymic links

8 Conclusion

A dictionary of a new type was developed. Its main features are:

– Entries belong only to nouns, verbs, adjectives and adverbs; they are grammemes (i.e. parts of lexemes) and can be multiwords.

– The links between entries are of semantic (syntagmatic or paradigmatic) or paronymic class.

– Its structure is interlingually universal.

Such dictionaries have a vast specter of applications: language learning (Bolshakov and Gelbukh, 2002b), word processing, syntactic analysis, word sense disambiguation; semantic error detection and correction (2003a); text translation (2001b), generation and segmentation (2001a); revealing text cohesion (Gelbukh *et al*., 2000), steganography.

9 Acknowledgements

Work was done under partial support of Mexican Government (CONACyT, SNI) and CGPI-IPN.

References

M. Benson, E. Benson, R. Ilson. 1986. *The BBI Combinatory Dictionary of English*, JBP.

L. Bentivogli, E. Pianta. 2002. Detecting Hidden Multiwords in Bilingual Dictionaries. *EURALEX-2002*, p. 14–17.

I. A. Bolshakov. 2004. Getting One's First Million… Collocations. *CICLing-2004. LNCS* 2945:229–242, Springer.

I. A. Bolshakov, A. Gelbukh. 2000. A Very Large Database of Collocations and Semantic Links. *NLDB-2000. LNCS* 1959:103–114, Springer.

I. A. Bolshakov, A. Gelbukh. 2001a. Text segmentation into paragraphs based on local text cohesion. *TSD-2001. LNAI* 2166:158–166, Springer.

I. A. Bolshakov, A. Gelbukh. 2001b. A Large Database of Collocations and Semantic References: Interlingual Applications. *International J. of Translation* 13(1-2):167–187.

I. A. Bolshakov, A. Gelbukh. 2002a. Heuristics-Based Replenishment of Collocation Databases. *PorTAL-2002. LNAI* 2389:25–32, Springer.

I. A. Bolshakov, A. Gelbukh. 2002b. Enseñando idiomas extranjeros con una base de colocaciones. In: *La telemática y su aplicación en la educación a distancia y en la informatización de la sociedad*, p. 632–638. Félix Varela.

I. A. Bolshakov, A. Gelbukh. 2003a. On Detection of Malapropisms by Multistage Collocation Testing. *NLDB-2003. LNI* 41:28–41. Bonner Köllen.

I. A. Bolshakov, A. Gelbukh, S. Galicia-Haro. 2003b. Stable Coordinated Pairs in Text Processing. *TSD-2003. LNAI* 2807:27–34, Springer.

A. Gelbukh, G. Sidorov, I. A. Bolshakov. 2000. Dictionary-based Method for Coherence Maintenance in Man-Machine Dialogue with Indirect Antecedents and Ellipses. *TSD-2000*. LNAI 1902:357–362, Springer.

Ch. Fellbaum (Ed.). 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

I. Mel'čuk. 1995. Phrasemes in Language and Phraseology in Linguistics. In: *Idioms: Structural and Psychological Perspectives*, p. 169–252. Lawrence Erlbaum Publ., UK, 1995.

OCDSE. 2003. *Oxford Collocations Dictionary for Students of English.* Oxford University Press.

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