MULTIFUNCTION THESAURUS FOR RUSSIAN WORD PROCESSING

Igor A. Bolshakov
Institute of Russian Language
Russian Academy of Science
Volkhonka, 18/2, 121019, Moscow, Russia
E-mail irlras@irl.msk.su

Abstract

A new type of thesaurus for word processing is proposed. It comprises 7 semantic and 8 syntagmatic types of links between Russian words and collocations. The original version now includes ca. 76,000 basic dictionary entries, 660,000 semantic and 292,000 syntagmatic links, English interface, and communication with any text editor. Methods of delivery enriching are used based on generic and synonymous links.

1 Introduction

Thesauri for commercial text editors are reduced now to synonym dictionaries. Meanwhile, the users often need to know, how might the given meaning be expressed by other words, not obligatory strictly synonymous or of different parts of speech, and what words are steadily combinable with the given one in texts. So various semantic (i.e. synonymous, antonymous, derivative, generic, meronymic) and syntagmatic (combinatorial) links are of interest.

Systematization of these links by A. Zholkovsky & I. Mel'chuk [1, 2] as lexical functions did not solve problem of gathering specific LF values. This proved to be of tremendous complexity and solved by the school of Mel'chuk–Apresian with speed insufficient for immediate word processing applications. But grouping LF makes them simpler for a common user to comprehend and less laborious for a developer to compile.

To get a friendly reference facility on links between Russian words, we have developed a prototype thesaurus named CrossLexica.

2 Directions of thesaurus use

In non–Russian community, our thesaurus is for students of universities with Slavonic departments, professional translators and teachers of Russian. A competence of such users in Russian may be various. So in the abroad version, hard–copy documentation, commands names, on–line help, error messages, and built–in translation dictionary were supplied in English.

Modes of use are the same for all conditions and comprise references out of or within context. In the first mode, the user types in a keyword by himself and gets, say, a set of its governing verbs. In the second mode, a query is formed within a conventional text editor, with return of the available information to the editor. In perspective, there exist many other ways of use of thesaurus DB, e.g. for filtering in syntactic parser.

The user might get through thesaurus following information: (1) synonyms; (2) antonym(s); (3) hyperonym; (4) hyponyms; (5) holonym; (6) meronyms; (7) common attributes for a given key; (8) words typically attributed by a given key; (9) semantic derivatives, i.e. the group of words conveying the same meaning through words of diverse parts of speech or through the same p.o.s., reflecting another participant of the situation; (10) verbs, (11) nouns, (12) adjectives, and (13) adverbs managing and steadily combinable with a given key; (14) managing model (case frame) for a given key, with all examples available; (15) a complementary element of a steadily coordinated pair (e.g. *prava i svobody* 'rights and liberties'). Consistently using this information, the user reaches valid and idiomatic texts.

3 Compilation of linguistic DB

The linguistic kernel of thesaurus is a dictionary consisting of words and phraseological collocations. It is between them the semantic and syntagmatic links are established.

When choosing elements of the dictionary, noun lexemes as a whole seemed unacceptable, since many nouns have diverse sets of attributes and/or managing verbs for the two numbers. So, as a rule, the numbers (if exist) were taken separately. Similary it is for two aspects of Russian verbs and verbs with reflexive particle *-sja*. Participles and adverbial participles are considered independently from their verbs, as exhibiting properties of adjectives and adverbs, correspondingly.

Homonyms, as usually, were numbered and supplied with short clear explanations. We deal similarly with polysemantic words such *tee* (drink *Vs.* grocery). The division took into account differences between sets of related words.

Compiling the dictionary, we took words covering Russian texts not less than to 90 percent and widely used words from sci–tech field. When acquiring new word combinations, new constituents appeared.

Methods of acquisition of word combinations were much more laborious:

Adoption from printed material. We disposed of only one dictionary of Russian word combinability with 2500 keyword entries, though.

Introspection, i.e. purposeful recollection of all stable combinations including the given word.

Analogy, i.e. matching a given entry with keywords significantly intersecting by meaning.

Systemity, i.e. engaging both noun numbers, both verb aspects, verbs adjoining this noun both as an object and a subject, etc.

Automated scanning of texts, i.e. the use of a program, moving a "window" along the text, and counting frequencies of joint falling into it of two or more relevant words [3]. This method is universal, even with a manual post–editing. Regretfully, we lack large corpora of Russian texts.

Calculation of LFs, i.e. intensive analysis, if there exist their explications for this key.

Manual scanning of texts turned to be the most productive. Different sci–tech papers, books, and abstracts on radar, electronics, computer science, automatic control, business, and applied linguistics were taken. Different Russian periodicals for 1988–1992 were also used.

4 Generation of on–line DB

The source files of the linguistic DB contain formatted texts, such as for managing verbs:

zabota	'care'
~ *okruzhaet*	'surrounds'
~ *projavljaetsja*	'is shown'
blagodarit' za1 ~u	'to thank for'
brat' na sebya ~u	'to take on oneself'
... ...	

We restricted marking of these texts to numbers of dictionary and preposition homonyms and to episodic part–of–speech labels.

At work, words/combinations should be automatically processed on entering to computer (normalization of inflectional forms) and on output (valid formation of gender, number, case, etc.). Thus, the dictionary entries should be supplied with morphological parameter(s).

Usually, construction of a morpho–dictionary considered as a separate task to be solved beforehand, thus necessitating permanent updating and morphological classification of new acquisitions. We took another way. Several complex utilities were written for translation of the source files to an on–line form and automatic constructing morpho–dictionary. These comprise automatic morpho–classification of words based on their final letters and short lists of peculiar lexemes, stems and prefixes, inserted directly to texts of the utilities.

Special codes were given to preposition–case combinations. All prepositions, including composite ones, were gathered and sorted. A Russian case (nominative, genitive,...) corresponds to each of them, forming a pair (preposition string, required case). Usual cases are formally among them as pairs (empty string, required case). The entries of the united pair list were named generalized cases. Their total number reaches 250. With a nonempty preposition, encoding of a word combination was thus evident, otherwise several heuristics were applied. Separate verb–noun combinations reflect subject–predicate pairs. For them, personal verb forms are used.

5 Delivery forming and enrichment

The thesaurus is destined for 15 main functions, basically described above: 1) **Synonyms**, 2) **Antonyms**, 3) **Genus**, 4) **Species**, 5) **Whole**, 6) **Parts**, 7) **SemGroup**, 8) **Attributing**, 9) **Attributed**, 10) **MngVerbs**, 11) **MngNouns**, 12) **CaseFrame**, 13) **Doublet**, 14) **MngAdjs**, 15) **MngAdvs**. In original version, the first twelve functions are implemented.

Each query to the system is a pair (main function, relevant key). A sequential use of delivery elements for next queries is a navigation within linguistic DB, that could lead arbitrarily far away from an initial key. The idea of the system implies, that none of its element could be an isolated node of the navigation network.

To perform specific functions, not only data of separate subsystems can be independently used (for direct delivery), but numerous links between subsystems (for enrichment of delivery), for example:

- If DB doesn't contain managing verbs, managing nouns, or attributes for the given noun, then sequentially, till finding nonempty contents, there are examined: other number of the same noun; its synonymous dominant; the nearest described hyperonym. E.g. there is the word combination *pick up berries* in DB, but not *pick up gooseberries*. So, using the hyperonymic link *gooseberries* $\rightarrow$ *berries*, needed combinations are delivered.

- As attributes for a given word, additionally to directly kept attributes, all passive participles are output, recorded in DB as predicates at the given noun subject. So for *abzats* 'paragraph', besides *bol'shoj* 'large',... words like *vydelennyj* 'chosen',... will be output.
- If there is no data for this aspect for a given verb in the DB, then those of the same verb in another aspect are taken.

6 Software implementation

As an operating environment, MS Windows ver. 3.1 with Russifier (font former) was taken. The IBM–compatible computer must have processor 386 or higher, main memory 2 MB or more and 6.5 MB of free disk space.

In the upper part of a working window, there is a menu of auxiliary functions. These are **Edit** (link with editors), **WordForms** (morphological paradigm of the key), **History** of current session, **Dictionary** (its fragment beginning by word closest to the input buffer contents), and **Help**. Below, the buttons with main functions are posed. Their inscriptions have three variants of contrast: (1) direct delivery is available for this function; (2) indirect delivery is possible; (3) delivery is empty.

Lower, the selected function and the input editing buffer are presented. An English translation of a highlighted word and a box for explanations of a homonymous key are also here. The input may be directly typed, as well as be taken from the **Dictionary** fragment, **History** list, a previous delivery, or text **Edit**or message.

The delivery, widely varying in size, is given at the lower part. For **CaseFrame**, it is split to zones corresponding to relevant generalized cases and supplied with questions, to which their entries response.

If an input string (as such or after automatic normalization) proved to be a dictionary entry, it is accepted as a component of a query. But if it is not reducible to a single entry, it is subject to simple parsing, with extaction of both potential parts and maybe a preposition. If both parts are in the dictionary and the link between them is also known, a query is formed automatically.

Though the thesaurus was developed for Russian, all its functions, run–time routines and the interface equally suit to other European languages. Only utilities for encoding of DB heavily depend on a specific language.

7 Quantitative features

The total size of the source text files of DB (without grammar tables) exceeds now 6.8 MB, while the volume of the dictionary is approximately 76,000. Semantic links are sized as follows:

derivative	44,200	526,100
synonymous	23,500	119,600
meronymous	3,200	6,400
hyponymous	2,200	4,400
antonymous	1,700	3,400
Total:	73,800	659,900

The second column counts all subsystems elements only once, the third one takes stock of all reverse and mutual links.

The current numbers of word combinations are:

managing verbs	149,800
managing nouns	56,100
attributes	85,600
coordinat.pairs	1,000
Total:	292,500

The coverage of open texts (in percents to a total occurrence number) was roughly estimated for verb–noun combinations (without enrichment feature). It is given below for several development steps, including the current (3rd) one and prognosis (4th) based on Zipf distribution.

St.	Num. of ent.	Mean ent.size	Text cov.,%	Num.of combs	Source vol,KB
1	2,670	9.8	40	26,500	419
2	3,870	17.1	50	66,100	1051
3	6,270	23.9	55	149,800	2408
4	7,000	30.0	60	210,000	4000

Laboriousness of acquisition of new DB elements is monstrous. But for users with not too deep knowledge in Russian, all necessary means for expression of the broadest specter of meaning through word combinations are already at hand.

Acknowledgements. I would like to thank Dr. P. Cassidy, USA, for sponsoring software development and primary system testing.

8 References

[1] Zholkovsky, A.K., I.A. Mel'chuk. *On semantic synthesis.* Problems of Cybernetics (in Russian). Moscow, Nauka Publ.- 1967.- v. 19.- pp. 117–238.

[2] Mel'chuk, I.A. *Semantic bases of linguistic description (Meaning–Text theory).* The Twelfth LACUS Forum, 1985.- Lake Bluff, Ill.: LACUS.- 1986.- p. 41–87.

[3] Calzolari, N., R. Bindi. *Acquisition of lexical information from a large textual Italian corpus.* Papers of 13th Int. Conf. Comp. Ling.- Helsinki.- 1990.- v.3.- pp. 54–59.