[From: *International Journal of Translation*, vol.13, no.1-2, Jan-Dec 2001]

A Large Database of Collocations and Semantic References: Interlingual Applications

IGOR BOLSHAKOV, ALEXANDER GELBUKH
Center for Computing Research
National Polytechnic Institute, Mexico

INTRODUCTION

Perhaps one of the most difficult problems in translation, both manual and automatic, is that of word combinability. Say, the word *strong* has a dozen different translations in Russian (as well as in Spanish, German, etc.); which one is combined with the Russian equivalent of *tea?* and with *man?* and with *wind?* and with *argument?* In each of these cases a different translation is to be chosen! Wrong choice of the translation variant is perhaps the most common error made by foreigners who pretty well know the language.

The information on relations between words is collected in special dictionaries. There are two major types of such dictionaries, both developed in the recent decade. The well-known databases of the first type, which we refer to as WordNet-like systems [1, 2, 3], are essentially thesauri. Their nodes are synonymy word groups (synsets) linked into a united hyperonymy (genus-species) hierarchy. The nodes are additionally interrelated with other types of semantic links: of antonymy, meronymy, role, cause, etc.

The databases of the second type are still rare and scarcely familiar. Their basic structure is the network of collocations (common word combinations) occurring in texts, with or without gaps between the components in their linear order, e.g., *fold → [in] arms, way → [of] speaking, deep ← admiration, kiss → passionately.* Collections of collocations in the electronic form are touched upon in [4, 5, 6]; the dictionary [7] a collection in printed form for English. Though some semantic correspondences between components of collocations do exist, the idea of a collocation is just co-occurrence

of its components in texts, as a chain of immediate syntactical dependences.

The databases of the second type can be also supplied with a thesaurical part that interlinks words that constitute various available collocations. The thesaurus plays auxiliary role herein, e.g., as a tool of enrichment of the collocations. However, a united database of collocations and semantic references (DBCSR) acquires new facilities that have no parallels in WordNet-like systems. The main objective of a DBCSR is to facilitate interactive text editing. However, it can be also used non-interactively, for parse filtering, lexical disambiguation, text segmentation, etc. All these applications imply a single-language database. Meanwhile, the introduction of a simple translation dictionary, from any external language to the basic language of DBCSR, makes it possible to use the system for various interlingual applications. Among them, we can mention translation of common collocations, computer-aided translation of idiomatic constructions of arbitrary complexity, and advanced learning of foreign language in general. The interlingual applications of DBCSRs are the subject of this paper.

Below, we first describe the most important parts of a DBCSR, with special stress to the collocation network. This aims to better distinguishing between DBCSRs and WordNet-like systems. Then interlingual applications of DBCSRs are described in more detail. Finally, we explain what additional demands to the subsystems of DBCSR the interlingual applications bring in.

All our prototypical considerations are based on the *CrossLexica* system preliminary described in [4]. This system operates with Russian as the basic language and English as the external one. However, examples of collocations are usually given in English (and thus are not taken from the real database) to be understandable for the reader.

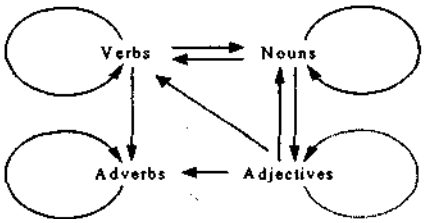


Fig 1. Various types of syntactic links

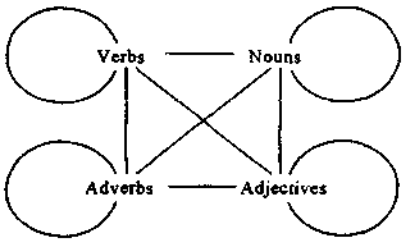


Fig.2 Various types of semantic links

2. Contents of a DBCSR

A DBCSR is a network with the nodes being words taken mainly from the general lexicon of a given language. Some idiomatic, terminological, or commonly used uninterrupted word groups are additionally included to the system dictionary as headwords, e.g., *mass media.* The types of relations between dictionary entries cover the majority of possible relations between words, both syntactic and semantic. Most of these relation types are present in all major European languages (i.e., of Germanic, Romance, and Slavic families).

2.1 *Syntactic Relations*

In the system database, these relations connect words of different parts of speech (POS); see Fig. 1. We consider only four main POSs: nouns N, verbs V, adjectives Adj, and adverbs Adv, in their specific syntactic roles. Each arrow in Fig. 1 represents an oriented syntactic link (we consider the links to be oriented from the syntactic governor to the dependent, such as eat → potatoes or hot ← potatoes). The system can retrieve the relation (a word pair) by either of the two words.

For the types of the nodes we consider, a noun group always plays the syntactic role of its head noun (mass ← media), verb group plays the role its head verb (is → closed), while a prepositional group as a whole plays the role of an adjective or an adverb, e.g., man → (from the South) or speak → (at random).

A syntactic relation between two words can be realized in the text, depending on the language, through (1) a "weak" preposition

in between, (2) a specific grammatical case of the noun or adjective (e.g., in Slavic languages), (3) a specific finite form of a verb, (4) a grammatical agreement, (5) a specific word order of the linked words, or (6) a combination of any of these ways. All these features are reflected in the system dictionary. Since substantive forms of different numbers can correspond to different sets of collocations, these forms and sets are represented in the system independently.

All types of commonly used collocations are registered in the system: free combinations like white ← dress or (to) see → (a) book; lexically restricted combinations like strong → tea or to pay → attention (cf. the notion of lexical functions in [8]); and idiomatic (phraseologically fixed) combinations like kick the bucket or hot dog. The restricted and phraseological combinations are introduced in the system when they belong to the mentioned classes, whereas the criterion of inclusion of a free combination is its "commonness" that is a rather diffuse notion. Nevertheless, as the experience of developing CrossLexica revealed, the semantics of collocation components significantly restricts the combinability of words even in free combinations. For example, one may expect only either an event or a living being whose appearance constitutes an event by itself. In Russian, the mean number of words combinable with any given word is in a rather narrow interval from 11 to 18, though the specific number is statistically distributed.

We consider only binary collocations. Of course, this underspecifies the description of the multi-valencied constructions. For reflecting somehow ternary collocations, we use ellipsis sign "..." for an omitted but obligatory valency (e.g., give a book..., give ... to the boy).

The following specific syntactic relations could be primarily taken for the collocation networks:

1. Has Attributes is a set of collocations in which a given word - a noun, adjective or verb - is attributed with some other word - an adjective or an adverb. These are relations of Noun → Adj, Verb → Adv, Adj → Adv, or Adv → Adv types. For example, the noun *act* can be attributed with *barbaric, courageous, criminal;* the noun *period,* with *incubation, prehistoric, transitional,* etc.

2. Is Attribute Of is the relation reverse to the previous one. For example, the adjective *national* can be an attribute for the nouns *autonomy, economy, institute, currency;* the adjective *economic,* for the nouns *activities, aid, zone,* etc.

3. Predicates is **a** set of collocations of Verb → Noun type in which a given noun is the grammatical subject and various ruling verbs are its commonly used predicates. I.e., this relation is of Verb → Noun type. For example, the noun *heart* commonly uses predicates *sinks, aches, bleeds;* the noun *money* uses *burns, is close, is flush,* etc.

4. Managing Verbs 1 is a set of collocations of Verb → Noun type in which a given noun is a complement and a common verb is its governor. E.g., the noun *head* can have governing verbs *bare, beat (into), bend, shake;* the noun *enemy* can have *arrange (with), attack, chase,* etc.

5. *Managing Verbs II* is a set of collocations of Verb → Verb type in which a given verb rules various other verbs, like in *to prepare → to sleep / use / read.*

6. Managing Nouns I is a set of collocations of Noun → Noun type in which various other nouns rule a given noun. E.g., the noun *clock* can be ruled by *hand (of), regulation (of),* etc.

7. Managing Nouns II is a set of collocations of Noun → Verb type in which a given noun rules different verbs, like in *readiness → to go / sleep / use.*

8. Managing Adjectives I is a set of collocations of Adj → Noun type in which a given noun is ruled by different adjectives. For example, the noun *rage* can be ruled by *mad (with).*

9. Managing Adjectives II is a set of collocations of Adj → Verb type in which a given adjective rules different verbs, like in *ready → to go / sleep / use.*

10. Managing Adjectives III is a set of collocations of Adj → Adj type in which a given adjective rules various other adjectives, like in *most ← profitable.*

11. Government Patterns represent schemes according which a given word (usually verb or noun) rules other words (usually nouns), and also give the lists of specific collocations for each sub-pattern. In the case of verbs, these are just their subcategorization frames without taking into account the word order in the pair. For example, the verb *have* has the pattern *"what / whom?"* with examples of dependents *capacity, money, family;* the pattern *"in what?"* with examples *hand, pocket;* and pattern *"between what / whom?"* with examples *lines, eyes.* Conceptually, this function is inverse to ***Managing Nouns I, Managing Verbs I, Predicates*** and several other relations. The system forms the patterns automatically, through the inversion of functions mentioned above.

12. Coordinated Pairs represent the coordination relation of Noun → Noun, Verb → Verb, Adj → Adj, or Adv → Adv type. It gives a word complementary to the keyword if the both constitute a stable coordinated pair like *back and forth, black and white, body and soul, good and evil, now and then, come and go,* etc.

Comparison of the relations enumerated above with dependency relations defined in Meaning <=>Text Theory by Mel'čuk [8, 9] shows that the former amalgamate some of the latter. Only auxiliary dependencies of the MMT are ignored as playing purely syntactic role. The relations in a DBCSR are those that directly link words in semantic representation of the same utterance. Comparison of the collocation mentioned above with lexical functions by Mel'čuk [8] shows that some of lexically restricted collocations are just these functions, in their basic definition or in a compound use (functions of functions). However, lexical functions represent only a small part of all possible collocations. We will not deep here in the corresponding theoretical issues.

2.2 *Semantic Relations*

These are well-known relations that include:

1. ***Synonyms,***
2. ***Antonyms,***
3. ***Genus*** *(*= superclass),

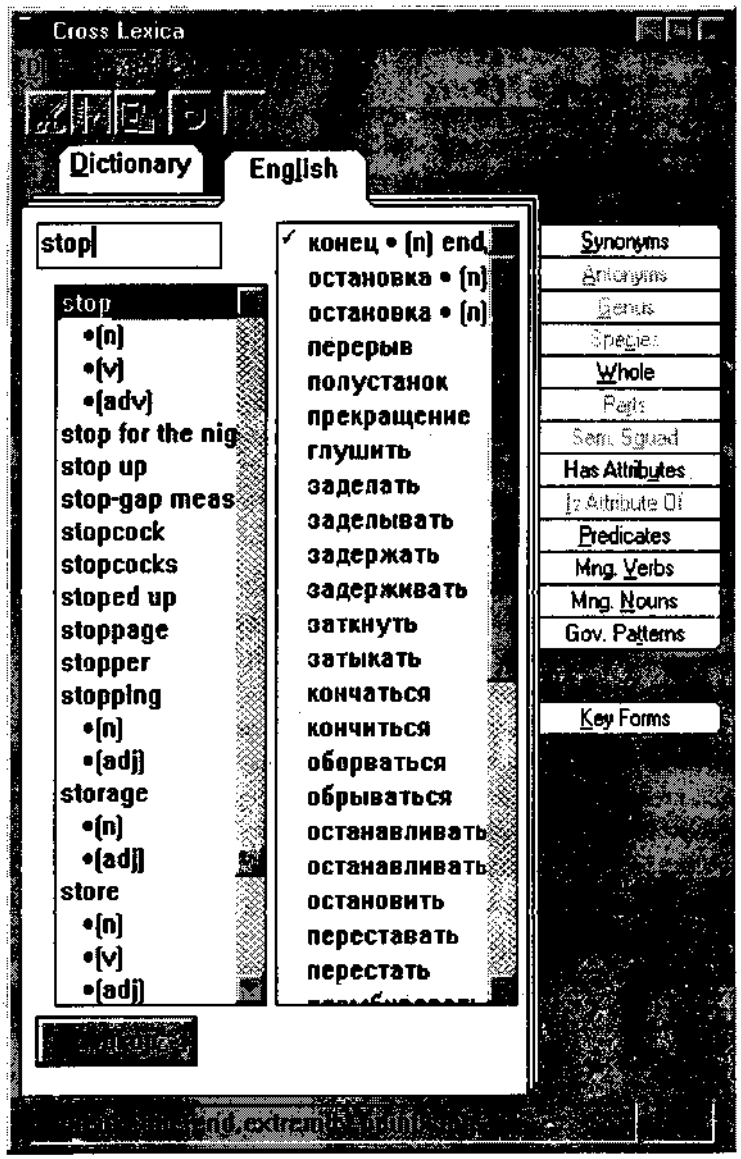
4. ***Species*** (= subclasses; inverse to ***Genus),***
5. ***Whole,***
6. ***Parts*** (inverse to ***Whole),***
7. ***Semantic derivates*** (see below).

Semantic derivates depict the meaning of a given word from other points of view and perhaps by other POSs. For example, when the user searches the system dictionary by any word of the following list

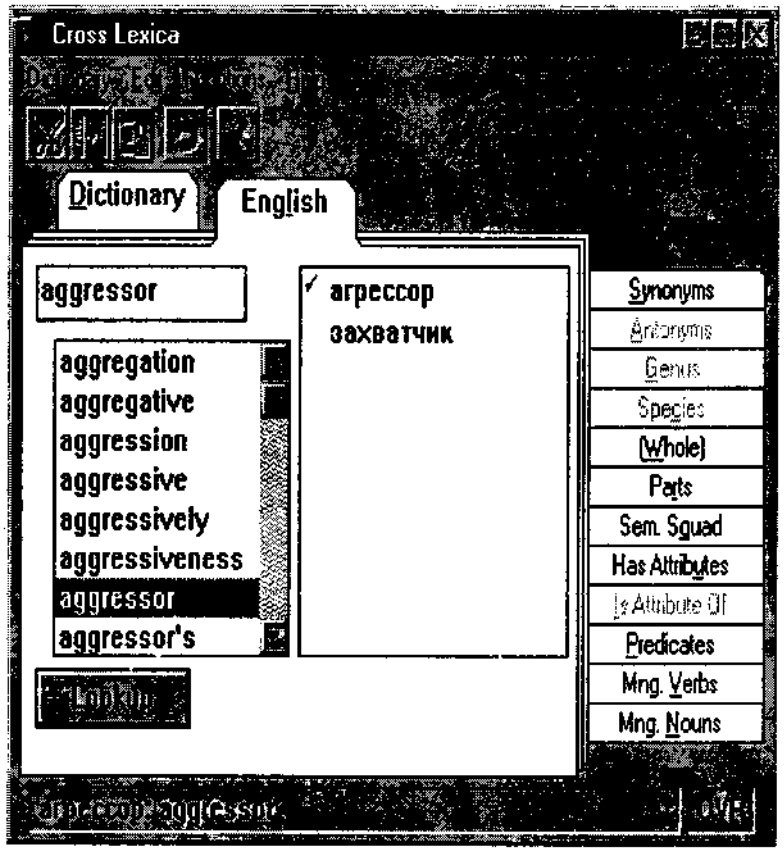
N:	*possession*	Verb:	*possess*
	property		*be possessed*
	possessor		*appropriate*
Adj:	*possessive*	*Adv:*	*in possession*
	possessing		
	possessed		

the other words of the same list are shown as its semantic derivatives. Only semantic derivates link words of different POSs in any-to-any manner, as shown in the full graph of Fig. 2. Here we assume that most meanings can be expressed in any given language by words or word combinations of all four main POSs, which is rather evident for verbs. On the other hand, for nouns reflecting living creatures, things, and artifacts this is hardly true. Indeed, what is the verbal equivalent for the meanings a fly or a stone? For such meanings, the corresponding partition in the set of four POS groups as in the list above is empty.

The set of semantic relations is very similar to that of WordNet-like systems. The main differences with the current version of WordNet can be formulated as follows.



a: The variants of translation for *stop*. The one correct for the phrase *to stop the pretensions of the aggressor* is beyond the bottom of the visible part of the list. The synonym English translations for the active Russian variant (marked with u) that are in the bottom line facilitate sense disambiguation.



b: The process starts with selection of the Russian word for *aggressor.* A double-click on the desired translation арессор in the right-hand window brings to Fig. 4.

* The species-genus hierarchy in a DBCSR does not have to be a mono-hierarchy. Deep mono-hierarchies give arbitrary (and often unexpected) answers to the questions like *Is vacuum cleaner a subclass of electrical device or of domestic appliances?* Poly-hierarchies cope better with such questions. They can be constructed quite naturally from separate genus-species pairs. It is necessary to take into account that natural language has a good method to verbally test such pairs: the notions *A, B,...* are species of the notion *C,* is the text "*A, B,..., and other Cs"* sounds semantically and grammatically reasonable (e.g., *radio, TV, and other mass media).*

* The nodes of the poly-hierarchies are usual words of natural language rather than theoretical concepts like *object* or *artifact* that do not occur in normal texts; otherwise the feature of self-enrichment of the system (described below) would not be realizable. This means that the nodes of poly-hierarchies scarcely can serve as nodes of ontologies well known in modern knowledge representation.

3. External Language

The inclusion in the DBCSR of a simple translation dictionary from an external language to the basic language of the system (i.e., the language for which the collocation and semantic links are given) permits the user to search the database it in the external language and receive the answer in the basic one. Thus, the primary merit of a DBCSR with additional language is its ability to perform word-to-word translation from external language to the basis one and vice versa, see Fig. 3 (explanations for this figure concern the example that will be given below).

For the needs of foreign users it would be ideal to include in the translation dictionary also translation equivalents of all collocations in the basic language present in the system dictionary. However, this is not realistic for the foreseeable future because of the huge amount (and cost) of work necessary for this. Instead, our *CrossLexica* system can propose to the user only out-of-context translations of collocation components, with the exception only for idiomatic phrasemes like *hot dogs,* which we do include in the dictionary.

However, in the direction 'external language → basic language' bi-partite collocations can be translated quite idiomatically. It seems quite easy for free combinations like *white dress*. However, for lexically restricted combinations this is not so simple because of multiple translations present in the dictionary.

In *CrossLexica,* the user automatically obtains translations of word combinations by just typing the English combination in the English-to-Russian dictionary entry field. Internally, the coincidence filter is used to build only idiomatic combinations. The translation is made in two steps. First, for each of the two words, all its out-of-

context translations are looked up, which gives two sets of words, A and B. Then, for all pairs of words $a \in A$ and $b \in B$ all pairs $a - b$ and $b - a$ are looked up in the combination dictionary, applying if necessary the enrichment technique described below. The pairs found as collocations are presented to the user.

For example, English collocation *strong tea* retrieves the only idiomatic Russian combination *krepkiy chay,* though for the word *krepkiy* out of context there exist numerous translations: *firm, robust, strong,* etc. Another example: English collocation *Russian President* is translated by the system only as *rossijskij president,* though for *Russian* both translations *russkij* (lit. 'of Russian-speaking people') and *rossijskij* (lit. 'of Russia') are present in the dictionary (the collocation **russkij president* is incorrect and thus absent in the collocation base).

In some cases the sets of separate translations of both components intersect twice or more. For example, Eng. *strong woman* is idiomatically translated as *sil'naja zhenschina* and *krepkaja baba.*

From the developer's point of view, the task is to include in the system the maximum of translation equivalents to each entry of basic dictionary, perhaps without any other information about words in external language (e.g., about their distinguishing to homonyms, polysemantic variants, etc.). The coincidence filter proved to be very strong.

4. Translation of Syntactical Combinations of Arbitrary Complexity

The word-by-word translation with filtering out idiomatic bi-partite combinations constitutes a rather powerful tool for computer-aided translation of syntactical combinations of arbitrary complexity from external language to the basic one. The goal is to obtain quite idiomatic text in the basic language.

As an example, let us consider expressing in Russian the idea that can be expressed in English roughly as *(We should) very very energetically stop the pretensions of the aggressor.* In fact we

even do not need to formulate it in an idiomatic English since we will choose the Russian words from the available repertoire.

CrossLexica system provides the user with an English-to-Russian dictionary. However, the naïve idea of its direct use for a word-by-word translation fails immediately because of polysemy and homonymy, see Fig. 3. Obviously, the choice of translations of some words in this phrase depends on the other (actually, these are lexical functions [8]).

The translation process is shown in the figures from 3b to 6. We start our translation from the only independent word in the phrase: aggressor (Fig. 3b). Of the two variants of the translation, actually both are suitable, though the first one - агрессор - is actually a transliteration of the English word and thus is expected to correspond closer to it, which in this case is true. With a double click on this variant, we get its collocations (Fig. 4), of which to narrow our search we choose nouns with the tabs located in the right-hand part of the window. Then we try each word, from the start of the noun section, and choose the one whose translation better reflects our idea (in this case, “pretensions”).

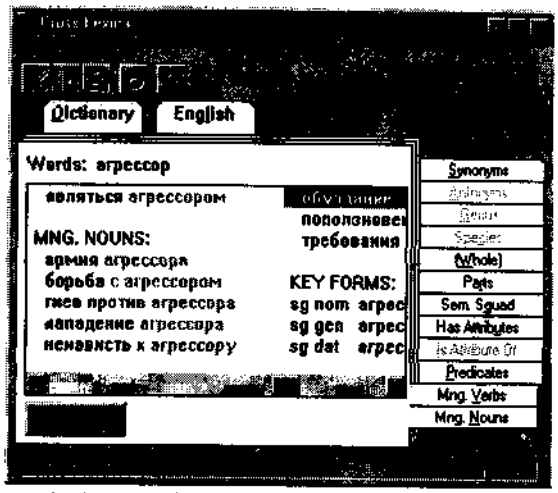


Fig. 4. Word combinations for *aggressor,* choosing the translation for *pretensions.* The headword агрессор in the appropriate grammatical case and with a proposition appears in gray color. The collocation on the left picture is rejected basing on its translation in the bottom line. The collocation in the right picture is accepted, so we have a part of the desired chain поползновения агрессора (note the case). A double click on the desired collocation brings to Fig. 5

If more than one word has the desired translation, usually any of them can be chosen. Expanding step by step the chain, we finally get the desired translation: райне решительно пресечь попол зновения агрессора that by its very construction is quite idiomatic.

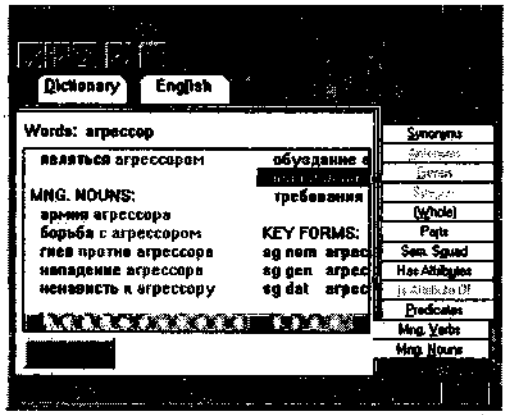


Fig. 5,

Note that the system automatically suggested the genitive case in the collocation пополэновения агрессора. If necessary, the system also suggests the appropriate preposition, пополэновения агрессора as in Fig. 5.

Why then the process is not completely automatic? First, sometimes the choice of a variant can be improved if the user has an intuition on the stylistic nuances of the available options. Second, frequently the available variants suggest improvements to the originally planned phrase. For instance, in Fig. 6 the English word *extremely* proves to better express the desired idea than the originally planned *very.* In this way, the system serves as a tool that encourages creativity of the user. On the other hand, we believe that an automatic (or a more automatic) procedure can give good results. This is a topic of future investigation.

5. REQUIREMENT: SOME NOTIONS ARE TO BE BROADENED

To improve the usefulness of the DBCSR for foreigner users and for translators, the system is improved in various ways in comparison

with a monolingual version. Most of these improvements have been successfully implemented in *CrossLexica* and has shown their effectiveness.

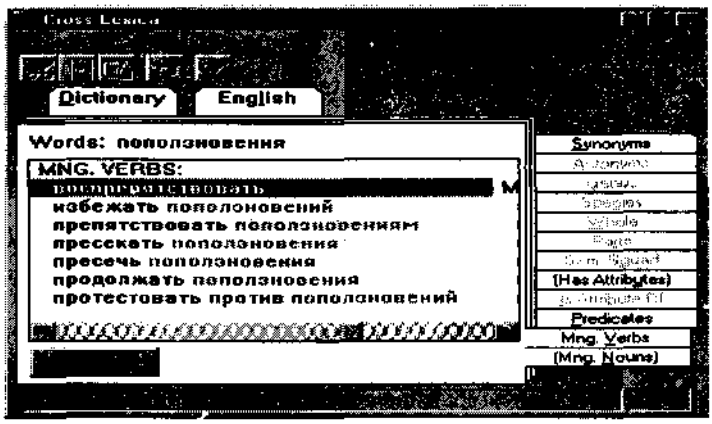
First of such improvements is the enlargement of its lexical base with a set of special functions useful for foreigners.

Broadened synonyms are the synonyms less similar to the headword than usual synonyms. They allow for greater flexibility in choosing the required meaning when the direct translation of a given word does not exist, or does not combine with the other words of the phrase, or the necessary collocation is not available in the system dictionary.

Broadened antonyms are, in the similar way, the words with the meaning similar to that of antonyms.

Paronyms represent the list of words of the same part of speech and the same root, but with potentially quite different meaning and collocations. For example, *sensation* is a representative of the paronymous group: *sensationalism, sense, sensitivity, sensibility, sensuality, sentiment.*

Homonyms and quasi-homonyms. Each homonymous word in the database forms a separate entry of a system dictionary. Each entry is supplied with numeric label and a short explanation of meaning. User can choose the necessary entry or observe them in parallel. It is important, that each homonym have its specific syntactic and semantic links.



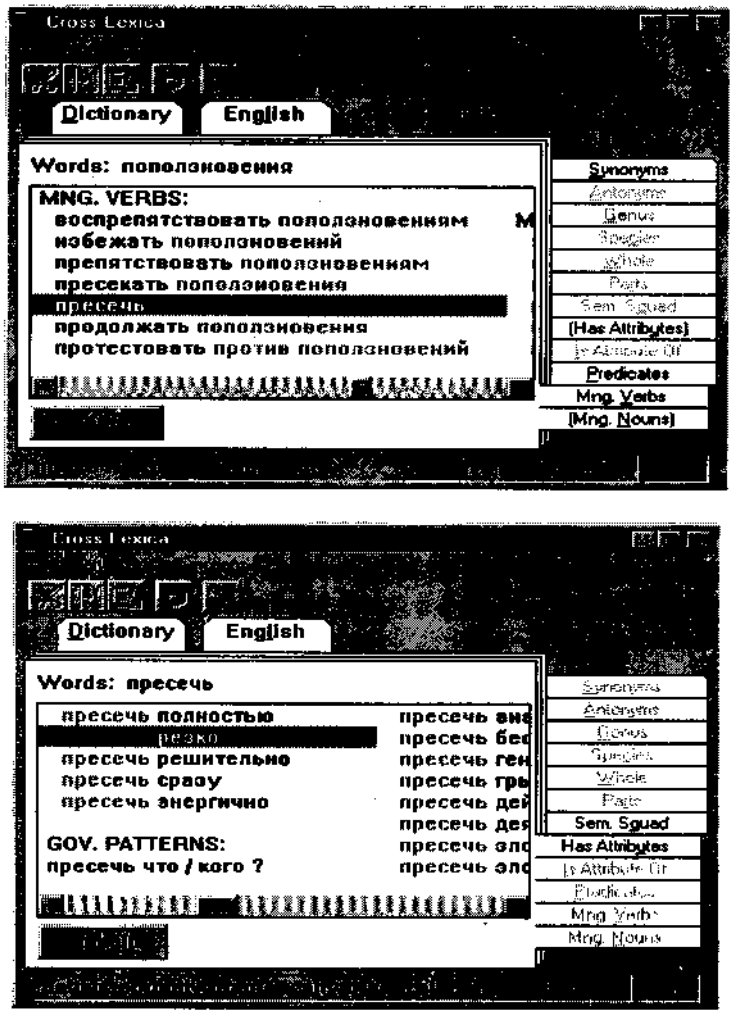
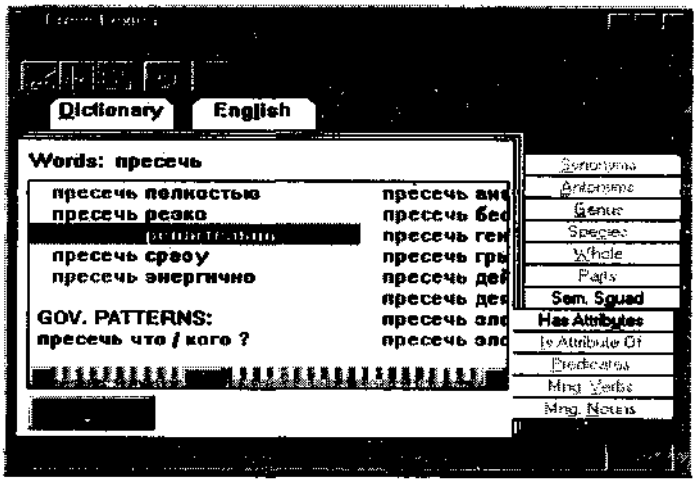


Fig. 5. Continuation of the process: The collocation in the top-left figure is rejected; the one in the top-right figure continues the chain: пресечь поползновения агрессора (cf. Fig. 3a). A double click brings the bottom figures; the right one continues the chain: решительно пресечь поползновения агрессора. A double click brings to Fig. 6.



6. REQUIREMENT: INFERENCE ABILITY

Another improvement made to *CrossLexica* in order to increase its usability by foreign users is its on-the-fly inference ability to enrich its collocation base, which is a unique property among this class of systems.

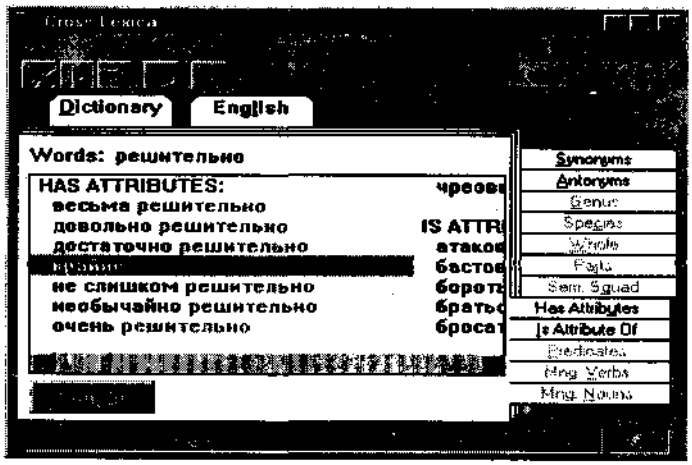


Fig. 6. The process finishes with the choice of the translation for *very,* which gives the desired chain: крайне решительно пресечь поползновения агрессора

The idea is that if the system has no information on some type of relations (e.g., on attributes) of a word but does have it for another word somehow similar to the former one, then the available information can be transferred to the unspecified or underspecified word. The inherited information is visually indicated on the screen (with a pale color) as not guaranteed, to warn the user that the system does not have any reliable information on the suggested collocation.

The types of the word similarity are as follows.

Genus. Suppose the complete combinatorial description of the notion *refreshing drink* is available. For example, verbs that combine with it are known: *to bottle, to have, to pour,* etc. In contrast, the same information on *Coca-Cola* is absent from the system database, except that this notion is a subclass of *refreshing drink.* In this case, the system transfers the information connected with the superclass to any its subclass that does not have its own information of the same type. Thus, it is determined that the verbs mentioned above are also applicable to *Coca Cola.*

Synonyms. Suppose that the noun *coating* has no collocations in the database, but it belongs to the synonymy group with *layer* as the group dominant. If *layer* is completely characterized in the database, the system transfers the information connected with it to all group members lacking the complete description. Thus, a user can recognize that there exist collocations of the type *cover with a coating.*

Note: these two types of self-enrichment are applied to all syntactic relations except ***Gov. Patterns,*** since this transfer reflects semantic properties not always corresponding to syntactic ones.

Enrichment of antonyms. Besides of antonyms recorded in common dictionaries, synonyms of the antonyms and antonyms of the synonymous dominant of the word are automatically added to quasi-antonyms. This is the only semantic relation subject to the enrichment.

Some inferences are nevertheless wrong. For example, *berries* as superclass can have nearly any color, smell and taste, but its

subclass *blueberries* are scarcely *yellow*. Hence, our inference rules have to avoid at least the most frequent errors. The following two filters are included in the system for this goal.

Classifying and non-classifying adjectives. The adjectival attributes sometimes imply incorrect inferred combinations like **European Argentina* obtained through the inference chain *(Argentina ⇒ country) & (European ← country)*. To avoid them, the system does not use the adjectives marked as classifying for the enrichment. Such adjectives reflect properties that convert a specific notion to its subclasses, e.g., *country ⇒ European / American / African country*. In contradistinction to them, non-classifying adjectives like *agrarian, beautiful, great, industrial, small do* not translate the superclass *country* to any subclass but instead indicate properties, so that collocation *beautiful Argentina* is considered by the system valid for enrichment.

Idiomatic and scope labeled collocations are not used for enrichment either. It is obvious that the collocation *hot poodle* based on the chain *(poodle ⇐ dog) & (hot ← dog)* is wrong.

Even with all these precautions that result in rather good quality of inference, the hundred percent correct inference is still impossible without further semantic research.

REQUIREMENT: VERY BROAD COVERAGE

The development of *CrossLexica* system, i.e., a Russian prototypic version of a DBCSR, is instructive both for unilingual and interlingual applications. Though there are no reasonable limits for the size of database and its dictionary, we consider *CrossLexica* to be now near to its final version. The dictionary contains now ca. 110,000 entries (headwords) that form about 2.5 million links. Namely, the database contains 1,437,600 syntactic links (counted unilaterally), of them:

Verbs - their noun complements	342,400
Verbs - their subjects	182,800
Nouns - short-form adjectives	52,600
Attributive collocations	595,000
Nouns - their noun complements	216,600
Verbs - their infinitive complements	21,400
Nouns - their infinitive complements	10,800
Copulative collocations	12,400
Coordinated pairs	3,600
Total	1, 437,600

and 1,047,600 unilateral semantic links, of them:

Semantic derivatives	804,400
Synonyms	193,900
Part/whole	17,300
Paronyms	13,500
Antonyms	10,000
Subclass/superclass	8,500
Total	1,047,600

In its current state, the system covers from 43% (abstracts) to 65% (advertising) of unprepared text with the inference capability turned off. The inference capability improves coverage not more than 5% rise in coverage so far. The reason is that the underspecified subclasses turn to be rather rare in texts and possibly because the thesaurical part of the system is not yet perfect.

For a foreigner, the coverage of approximately 60% of text is not sufficient. Thus, to improve its interlinguistic applications, the database should be significantly broadened.

Will such broadening lead to huge increase of the size and cost of the database? According to our data, it will not. Indeed, currently each word is associated on average with approximately 15 words with the most fertile links, such as verb → noun and noun → noun (the figure varies from 12.8 to 17.9, depending on the specific relation). Thus even including into the database the collocations

considered linguistically quite free gives on average only ca. 20 different words syntactically connected with each given word, in each category of collocations (both dependent and ruling syntactic positions are considered). The evident reason of this constraint is semantics of words, so that the total variety of collocations in each specific case does not exceed some limits.

8. CONCLUSIONS

DBCSRs prove to be very useful tools not only for monolingual text editing but also (and maybe in an even greater degree) for both computer-aided translation and advanced learning of a foreign language (which is a necessary stage of preparation of a human translator). They address the issues traditionally most difficult for translators: idiomatic combinability of words in foreign language and word choice in non-free combinations.

On the other hand, they help in translation from the foreign language to the native one since text editing - for which they were initially developed - is an essential part of high-quality translation work.

As an example of a fully implemented DBCSR we have considered our system *CrossLexica.* This system allows the user to automatically translate word combinations such as *strong tea* and helps the user to express long word chains in Russian, the system's basic language.

From the developer’s point of view there are issues to be addressed in the developing a DBCSR intended to foreign users and to translation tasks. Of these, we have considered widening the set of semantic and syntactic relations included in the dictionary and enlarging the dictionary to improve its coverage. A very interesting feature implemented in *CrossLexica* for translation purposes is its inference ability.

Though we have discussed the use of DBCSRs such as *CrossLexica* only as tools for computer-aided manual work, we believe that the same use scenarios are applicable for automatic translation systems. Implementation of these functions in a fully automatic system is a topic of future research.

REFERENCES

M. Benson, M. et al. 1989. *The BBI Combinatory Dictionary of English.* Amsterdam, Philadelphia: John Benjamin.

Bolshakov, A. 1994. Multifunction thesaurus for Russian word processing. *Proceedings of 4lh Conference on Applied Natural language Processing,* Stuttgart, 13-15 October, 1994, 200-202. Fellbaum, Ch. (ed.). 1998. *WordNet as Electronic Lexical Database.* MIT Press.

Ferret, O., B. Grau, N. Masson. 1998. Thematic segmentation of texts: two methods for two kinds of texts. *Proc. of Coling-ACL-1998,* v. 1, p. 392-396.

Koyama, Yasuo et al. 1998. Large Scale Collocation Data and Their Application to Japanese Word Processor Technology. *Proc. Intern.* Conf. *Coling-ACL '98,* v.I, p. 694-698.

Mel'čuk, Igor A. 1988. *Dependency Syntax: Theory and Practice.* Albany, N.Y.: State University of New York Press.

Steele, J. (ed.).1990. *Meaning - Text Theory. Linguistics, Lexicography, and Implications.* University of Ottawa Press.

The Spanish WordNet. Version 1.0, July 1999. Euro WordNet, LE2-4003 & LE4-8328. CD ROM (distributed by ELDA).

Vossen, Piek (ed.). 2000. *EuroWordNet General Document.* Vers. 3 final. http:// 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 Publ.