A large electronic dictionary as polythematic guide and shaper of queries to the Web

Igor A. Bolshakov
Independent researcher, Moscow, Russia
Alexander F. Gelbukh
National Polytechnic Institute, Mexico City, Mexico

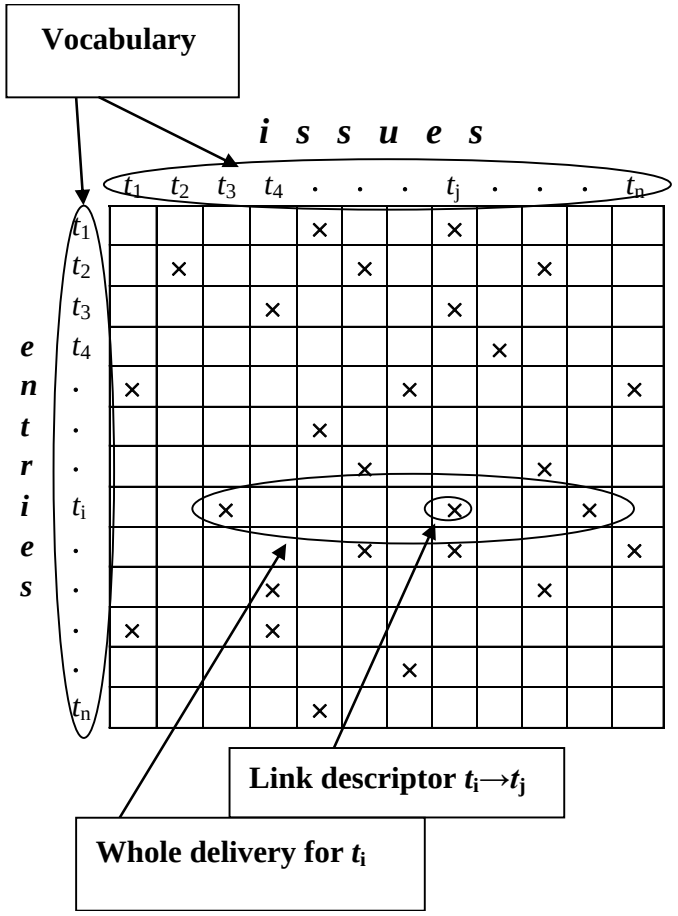
Summary

CrossLexica is revisited, a large electronic dictionary containing both fundamental information on Russian language (grammatical properties of words and their combinations, semantic and paronymous links between words), and extensive encyclopedic information on geoobjects, well known persons, organizations, and artifacts.

CL contains technical terms and basic notions of exact and natural sciences, of humanities, business, and economics.

CL allows to form and to immediately send to the Web queries of medical, commercial and travel nature.

Main idea of the dictionary: Huge matrix



CrossLexica =

{Vocabulary x Vocabulary}

with an element *D(ti, tj),* i.e. descriptor of a link between query title t_i and issued title t_j
i, j = 1, ..., 245000+

Important feature:

Because of language restrictions, the matrics is very sparse: only each 8600th descriptor is nonempty, although approximately 7 millions links are already found in it.

Links between dictionary titles

are divided to

- **Sintagmatic** ones that form collocations
- **Semantic** ones between words with meaning similarity
- **Paronymous** ones between words with letter (sound) similarity

Quantitative parameters

- Total vocabulary volume in title entries 245,000
  - Nouns 41%
  - Verbs 15%
  - Adjectives 26%
  - Adverbs 18%
- Total number of collocations 1.97 million
- Total number of semantic links 2.43 million
- Total number of paronymous links 0.60 million
- Grand total of links ` 6.97 million

CrossLexica is an integral whole

However the named sections of its deliveries taken together could be considered as subdictionaries, to compare them with the available dictionaries.

Let us divide the subdictionaries to

- Majorizing subdictionaries that exceed parameters of their analogues
- Subdictionaries of a new content not having analogues till now
- Other subdictionaries having a number of new properties

Majorizing subdictionary (1/5)

Subdictionary из **collocations**, partially terms of

- Economics and business
- Socio-political themes
- Technologies
- Exact and natural sciences
- Humanities, arts, religions
- Medicine

Terms reflect encyclopedic knowlende

Majorizing subdictionary (2/5)

Subdictionary of **Management Patterns** for

- Verbs (more than 22,000)
- Nouns (more than 14,000)
- Adjectives / participles (more than 4,000)
- Adverbs / (Russian) gerunds (more than 2,000)

It divides collocations to groups corresponding to the same «case»:

- конструировать что?
  волноводы, машину, механизмы, модель, прибор, приборы, программы,…
- конструкции из чего?
  из металла, из стали, из фанеры,…

Majorizing subdictionary (3/5)

Subdictionary of ***synonyms*** (1.22 million links, i.e. Twice as much as in Alexandrova's dictionary)

It indicates and uses absolute synonymy.

It include abbreviations and glues like

филфак или *заксобрание*

Majorizing subdictionary (4/5)

Subdictionary of ***antonyms:***

It exceeds all known dictionaries

It additionally contains (given in low contrast) approximate antonyms, i.e. synonyms of true antonyms and true antonyms of synonyms.

Majorizing subdictionary (5/5)

Morphological dictionary gives all inflectional forms for all vocabulary titles.

Multiword noun expressions have up to 6 elements, with up to 5 of them inflectional
(десять заповедей и семь смертных грехов)
(the ten commandments and the seven deadly sins)

Multiword verb expressions have up to 6 elements, with up to 3 of them inflectional
(жить-поживать и добра наживать)

Subdictionary with a new content (1/4)

- Subdictionary of ***semantic derivates*** contains groups with 4 POS sections in each:
  nouns, verbs, adjectives / participles, adverbs / (Russian) gerunds.
- Usually, the sections are connected morphologically:
  владелец, владение; владеть, овладеть; владеющий, владевший; владея, овладев.
- However there exist groups with noun section containing encyclopedic information on several tens of countries (on Russia the most, on top-7 less, on top-20 even less, etc.), on Russian cities and large areas, etc.

Subdictionary with a new content (2/4)

Dicrionary of associations in Web users' queries
(i.e. typical user's profile)

454	*pregnancy*	110	*culture1*
213	*health*	107	*smoking*
189	*alcohol*	106	*love*
139	*sports*	101	*diet*
118	*prices*	97	*business*
118	*human*	92	*politics*
115	*diabetis*	92	*psychology*
111	*children*	92	*ecology*

Subdictionary with a new content (3/4)

Subdictionary of ***literal paronyms*****,** i.e. words differinf in one letter:

кадка: *кака, ка**с**ка, ка**ч**ка, ка**ш**ка, к**л**адка*

Subdictionary with a new content (4/4)

Subdictionary of ***meronyms / holonyms***
(links «part vs. whole» or «whole vs. part»)

- For subsets: *writer – group of writers / writers association/ …*
- For divisible objects: *body1 – hands / feet / head1 / belly /…*
- For innumarable objects: *water – water drop / glass of water / pint of water /…*

Subdictionary with advanced features (1/4)

Subdictionary of ***morphemic paronyms*** (i.e. words with the same root and a small number of differing affixes:

бегающий, беглый, беговой , бегучий , бегущий,...

Subdictionary with advanced features (2/4)

Subdictionary of ***hyponyms / hyperonyms*** (the link IS_A or its reverse) is basically a polyhierarchy of notions where a notions can be member of various (may be intersecting) classes.
E.g., France is a member of classes:

European country,
Mediterranean country,
NATO member country,
European Union member country, etc.

Encyclopedic references available

- Names of continents, oceans, seas, moutain ranges, and other geoobjects of the world.
- Names of world's largest cities in relation to their countries.
- Information on 60 states (capital, monetary unit, way of governing, the title of Head of State, titular nation, state language, unit of administrative division, official religion, etc.)
- Names and other information on tens of Russia's cities and regions (names of city residents as well as names of all Moscow areas are given).
- About 300 male and female frequent names (mainly Russian) with their diminutives.
- Names of the most famous political, business, scientific and cultural persons of the world.
- Names of the largest organizations (corporations).
- Names of the most famous artistic works of the world (novels, operas, musicals, movies, etc.)

Optional facility:
English-Russian and Russian-English dictionaries

From English part it is possible

- To enter to Russian vocabulary and then to select a Russian word needed.
- To get an idiomatic Russian translation of English collocation (frequently in several options).

For each Russian title, it is possible to get all available English translations.

Computer-aded composition of a query to the Web

Procedure for making request:

- A key title for the query is quickly found in the vocabulary.
- If the key title is considered sufficient, it is immediately sent to the Web. In this way we can send queries *abortions and consequenses, respiratory organs*, *heart deseases, social and health insurance, Great Britain and European Union.*
- If the key title is not sufficient for effective search, its delivery is searched through, and a collocation needed is sent to the Web. For example, the key *shelf life* is taken, and in its delivery the collocation *shelf life of yogurt*.

Main advantages are grammatical correctness of the query – and extensive menus to select from.

Main disadvantage is inability to make queries including new technical names (brands) or concerning events recently occurred.

Conclusion

- Advanced features of CrossLexica are described, which is combinatorial dictionary with no analogues in volume and structure for any language.
- It is demonstrated that CrossLexica is a resource equally suited
  - for giving references on linguistics and encyclopedic issues
  - And for shaping queries to the Web.

Thank you for your attention!
Any question?

Igor A. Bolshakov
bolshakov34@mail.ru
iabolshakov@gmail.com