Journal of the Applied Mathematics, Statistics and Informatics (JAMSI), **3** (2007), No. 2

# The Ideology and Technology Of Creating Online Full-Text Digital Collections of Ancient and Medieval Slavonic Literary Texts

VICTOR A. BARANOV

**Abstract**

One of the main tasks for creators of both large- and small-scale digital collections of literary texts is to provide users a convenient means of navigation to allow them to quickly find information. If the materials in the collection are unique documents (in particular, ancient and medieval), the development of a means of storage of digital copies of manuscripts is made significantly difficult by the fact that the system needs to help the user not only locate a document but also solve a concrete research goal.
Currently the creation of full-text collections and libraries of ancient and medieval manuscripts of literary texts uses technology based on database interfaces and markup languages for deeply encoded documents. For both technologies there are three tasks: (a) the unification in one information system of capabilities for handling not only findings from the texts and manuscripts but also findings on small (down to the character level) and abstract (semic structure of words, characteristics of objects of thematic classes, etc.) objects; (b) the creation of systems for remote access with developed facilities for search and, in certain circumstances, for manipulation of data; (c) the necessity to create a multifunctional and user-friendly interface.
An example of the creation of a full-text, multifunctional, specialized system is the Manuscript project (http://manuscripts.ru/index_en.html), based at the Udmurtia State University and the Izhevsk State Technical University. The system represents a complex of editorial, search, and research modules based on a full-text database and intended for entering, editing, storing, manipulation, and analyzing transcriptions of ancient and medieval documents with complex structure as well as for preparing queries and manipulating retrieval to prepare glossaries and concordances and publish these texts online.

**Mathematics Subject Classification 2000**: 68N99

**General Terms**: On-Line Digital Collection
**Additional Key Words and Phrases**: Complex Technology of Creation of Electronic Collections of Slavonic Manuscripts on the Internet: Multilingual Full-Text Database, Tools for Filling, Access and Publication

## 1. AIMS AND TASKS

1.1. One of the main tasks for creators of both large- and small-scale digital collections of literary texts is to provide users a convenient means of navigation to allow them to quickly find information. Today this task is being successfully solved in online digital libraries, established for storing digital texts in contemporary languages.

If the materials in the collection are unique literary texts (in particular, ancient and and medieval), the development of a means of storage of digital copies of manuscripts is made significantly difficult by the fact that the system needs to help the user not only locate a document but also solve a concrete research goal. Since literary texts can be used to solve such a wide range of problems, today's technology comes in many forms: from special systems for accessing scanned page images to multi-function systems for input, editing, manipulation, and electronic publishing of texts, plus their metadata and analytical, semantic, and other characteristics of digital objects.

1.2. The construction of multi-functional digital repositories is made difficult by the incongruous demands of scholars from various disciplines, including different requirements for the level of detail of a document and its faithfulness to the original, for the uniformity of digital objects and metadata, and for the data model.

The solution to this problem is to create a text- and document-oriented platform for research that would allow the user to manipulate all objects in manuscripts and texts at the point of preparation, manipulation, and use. There are a few types of objects subject to manipulation in a full-text system: (a) formally displayed units of electronic transcription (characters, wordforms, clauses, significant elements, texts and manuscripts); (b) entities such as dictionaries, indices, glossaries and concordances, and authority files; (c) properties and values

of these units and entities (metadata about manuscripts and texts, analytical data about fragments, linguistic information about the semantics and morphology of textual units, information about dictionary entities); (d) the relationships between manuscripts, texts, and dictionary entities and their objects (relationships between meta- and analytical objects, information on links, syntactic links between units of text, etc.).

A principally important feature in building such a system is the scholarly apparatus, built on the structuro-formal, essentio-semantic, and thesauro-commentarial systematization of all objects in the collection. In particular, thesauri, indices, and dictionaries created using manuscriptual evidence for various units, allow for effective navigation and searching by the end user. A semantic and grammatical dictionary and thesaurus must be the central components of the linguistic modules of the system, and dictionaries for meta- and analytical information must be central for modules that handle manuscripts, texts, and fragments.

Since the object of study is often collections of unique manuscripts containing variants of a single text (or fragments thereof), the system must be able to compare similar objects and allow for navigation between corresponding parts of manuscripts: fragments, citations, allusions, periphrases, and such. This requirement is related to the user's need to consult a certain fragment of the work in various witnesses (written in different languages, times, and places). In essence, the problem is one of creating an digital analog of a critical edition of a text.

Therefore, the use of a digital library or collection -- not only digital copies of manuscripts themselves (in various forms, including transcriptions and scanned images) but also additional information about them -- allows the user, regardless of goal or task, to move not only from dictionary entities to their realization in texts and/or manuscripts but also from a concrete context (subjects, semantics, relationship between objects) to a system of analogous objects, presented as variants from other witnesses, dictionary entities, or objects from authority files.

1.3. Currently the creation of full-text collections and libraries of ancient and medieval manuscripts of literary texts uses technology based on database interfaces and markup languages for deeply encoded documents. For both technologies there are three tasks: (a) the unification in one information system of capabilities for handling not only findings from the texts and manuscripts but also findings on small (down to the character level) and abstract (semic structure of words, characteristics of objects of thematic classes, etc.) objects; (b) the creation of systems for remote access with developed facilities for search and, in certain circumstances, for manipulation of data; (c) the necessity to create a multifunctional and user-friendly interface.

1.4. Regardless of the fact that the Slavic written heritage of the 11th-17th centuries is so far only presented online in fragments (as electronic publications of separate manuscripts and texts or as small collections of varying quality and technological implimentation), we can say, based on current projects, that (a) soon the number of the online electronic publications of Slavic manuscripts will grow significantly, (b) a digital grammatical diction of Old Russian will soon be created, (c) research teams will be able to exchange full-text collections, and (d) work will soon be begun on a digital thesaurus of Old Russian.

## 2. INFORMATION-ANALITICAL SYSTEM “MANUSCRIPT”

2.1. One example of the creation of a full-text, multifunctional, specialized system is the Manuscript project (http://manuscripts.ru/index_en.html), based at the Udmurtia State University and the Izhevsk State Technical University. The system represents a complex of editorial, research, and informational-search modules intended for entering, editing, storing, manipulation, and analyzing ancient and medieval literary texts with

complex structure as well as for preparing queries and manipulating retrieval to prepare glossaries and concordances and publish these literary texts online.

The system includes the following modules, which interact with a single database containing a hierarchically structured data: a specialized editor, a module for electronic publishing, a query and retrieval module, a dictionary module, a lemmatizer, a module for exchanging data with other systems in other formats, a module for storing and manipulating meta- and analytical data, a module for loading text and reference material for print publication, and a module for storing and exchanging documents.

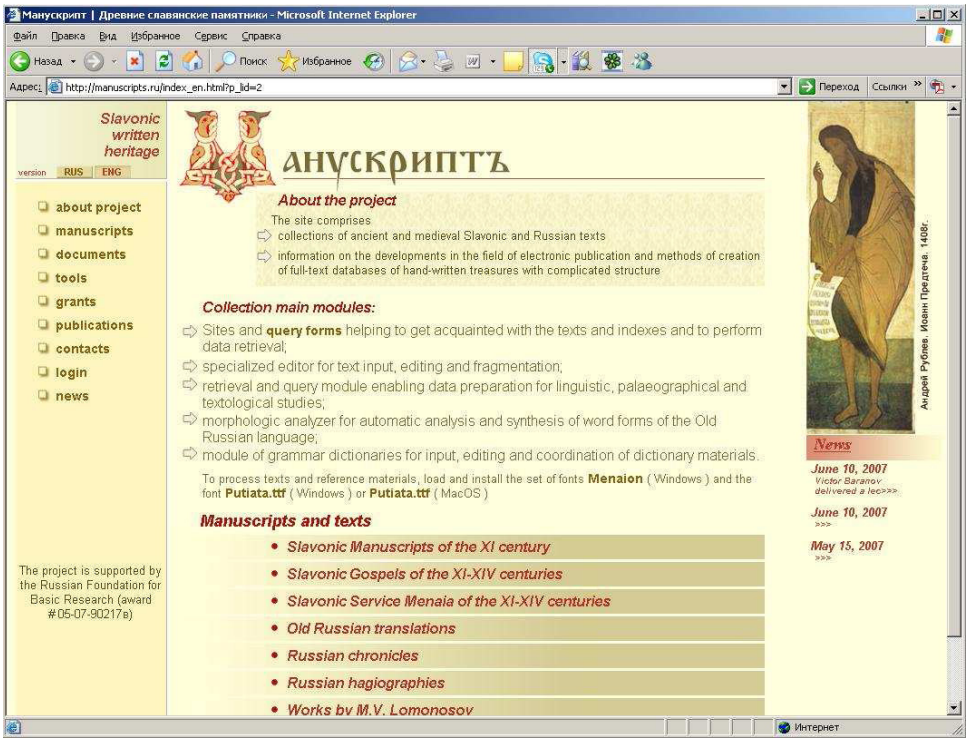


Fig. 1. Title page of the project Manuscript portal.

2.2. Database model: Findings on manuscripts and texts are stored as hierarchically linked objects of a certain type: objects describing the arrangement of units in the manuscript (a folio, page, column, line), textological units (a fragment created later than the manuscript, a fragment created by another scribe), and linguistic units (a word-form, syntagm, phrase, text, etc.).

The minimal unit of the hierarchy is always the character; the maximal, a text or manuscript. A specially developed character encoding scheme, including a special extended block of Cyrillic characters, allows the transcription to correspond maximally to the original and manipulate data without needing to approximate characters. More complicated is the linguistic hierarchy (see Figure 2), which includes objects needed for adequately representing the structure and semantics of a text, for describing the relationship between units of text, and for working with these objects and their relationships. Structural relationships between units are stored in a tree, and semantic relationships in a network.

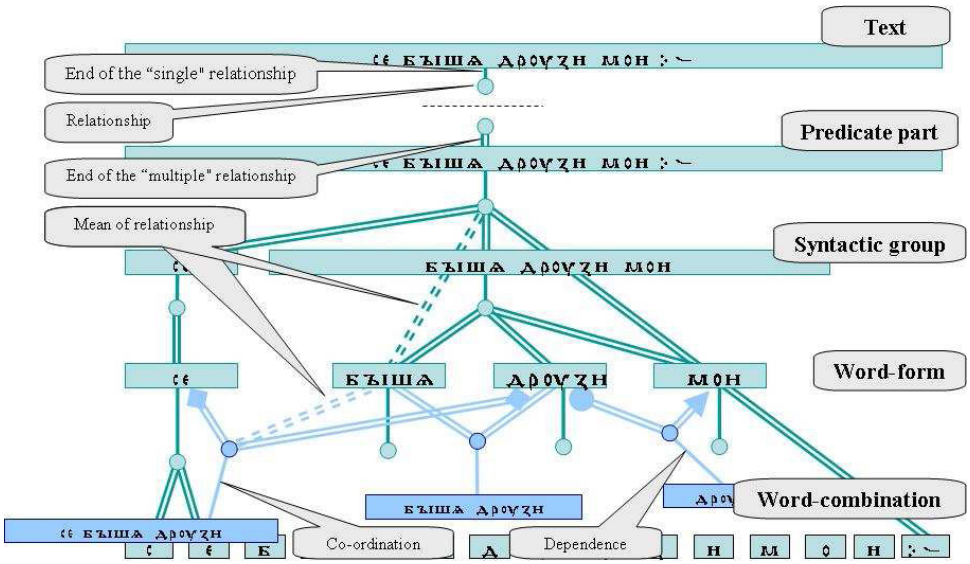


Fig. 2. Net of linguistic relationships.

2.3. OldEd specialized editor: The editor is intended (a) for entering and editing units of a manuscript, (b) for fragmenting a manuscript and text into units, (c) for building hierarchical and non-hierarchical relationships between units, (d) for assigning properties and their values to units, (e) for linking between units of the manuscript and units of a dictionary (see Figure 3), (f) for lemmatizing a text, (g) for creating a protograph of a text, and (h) for modelling electronic publishing and other operations.

The editor assists in fragmenting manuscripts and texts and assigning value to selected objects. If the fragment represents a variant of an object, which exists or might exist in many manuscripts or texts, the editor allows the user to connect fragments with their invariant stored in the corresponding dictionary. The editor assists distributed and distance work on digital transcriptions of text.

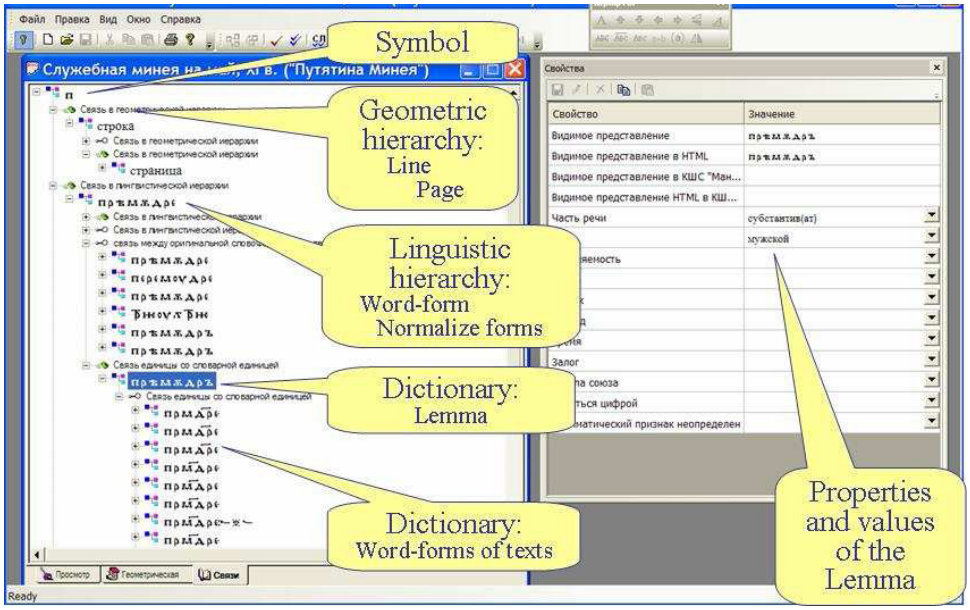


Fig. 3. Editor OldEd: Visualization of unit relationships.

2.4. Web-module for electronic publishing: The main component of an electronic edition is the search page, which allows the user to choose a collection of manuscripts and a text in a collection and to define the search criteria. The search results can contain lists of (a) direct and inverted indices of stems, frequency indices, and incipit indices and (b) various texts (original, normalized, diplomatic, transliterated) (see Figures 4, 5). There is rapid growth in collections of of ancient Slavic minaias, triodeas, and gospels, as well as a collection of 16th-century Old Russian manuscripts and a collection of Old Russian translations.

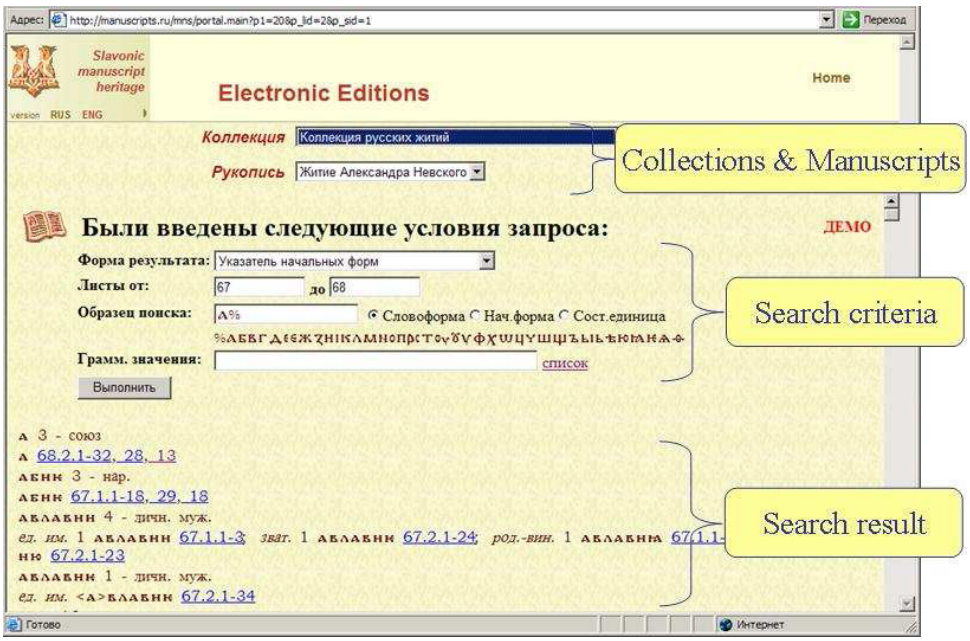


Fig. 4. Electronic edition: search page.

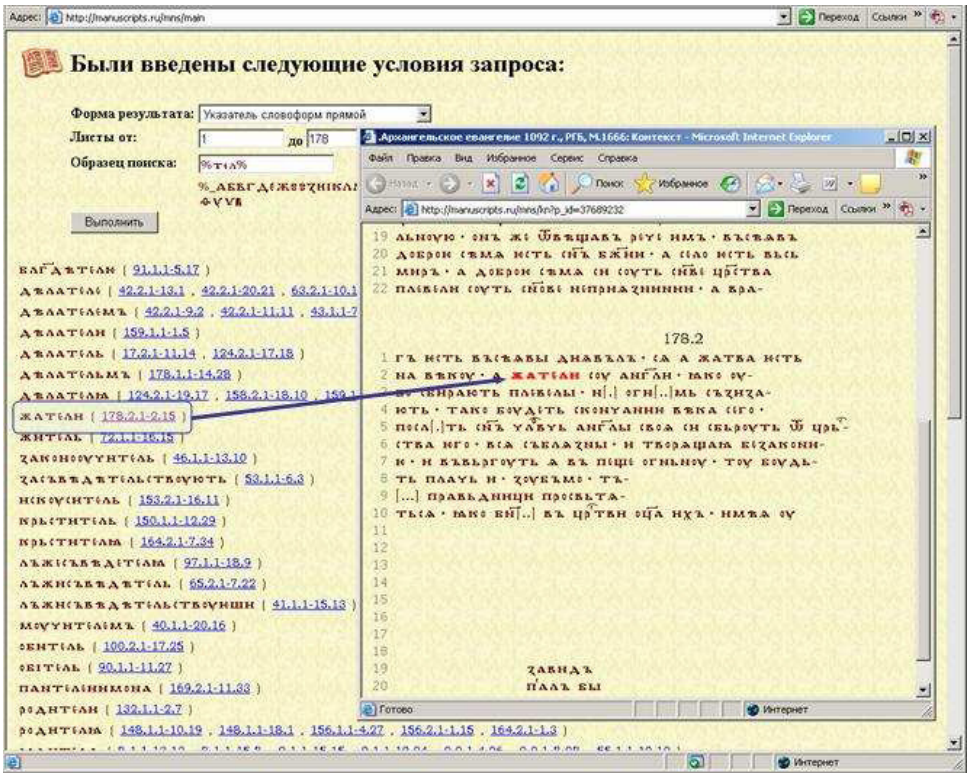


Fig. 5. Electronic edition: Word-form index and concordance.

2.5. The retrieval and query module is intended for keyword queries on the database and allows the user to (a) choose units, properties, and their values (see Figure 6); (b) specify relationships between entities; (c) specify the composition of the query result, the priority of the entities, and their sort order; (d) retrieve results; (e) perform operations on the results; and (f) use query results for future queries.

On the whole, the module of retrieval and queries provides to the user flexible possibilities for the independent formation of queries on the basis of any unit properties and values available in the database. In the future this module should allow expanding boundless the possibilities of the Manuscript system for research of scientists from various fields.

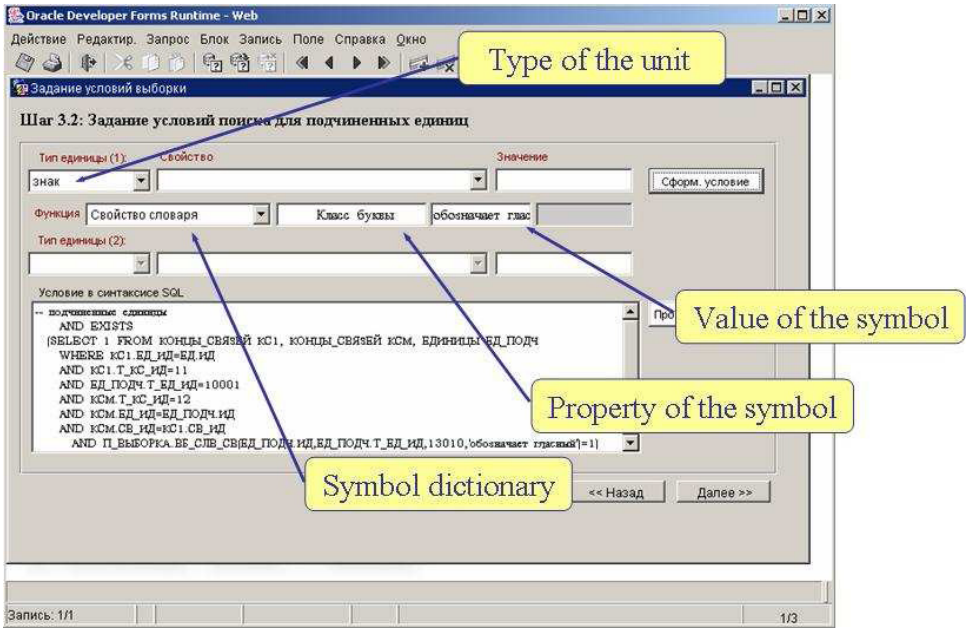


Fig. 6. Module of retrievals: setting the entity properties and values.

2.6. Grammar dictionaries and lemmatizer: The central module of the manuscript system should be the module of dictionaries that is under construction. The module should ensure the use of the common dictionaries – the authoritative files in fragmentation and analysis of texts and manuscripts.

The most complicated system, from the point of view of implementation, is a system of linguistic dictionaries that is oriented to automatic morphologic analysis of ancient and medieval Slavonic texts. The creation of grammar dictionaries is based on the principle of operation with the variable and invariable parts of the word form – endings and stems (see Figures 7). The computer linguistics has several approaches to the division of the word form into these components: with and without storage of interchanges in the stem. Our system uses both the approaches: for normalized stems and their endings the first approach, for variant components the second. This is done to get a comprehensive solution of the main task of the automatic lemmatizer which is the task of analysis of a very changing graphic and morphologic structure of the ancient Slavonic word form (see Figures 8).

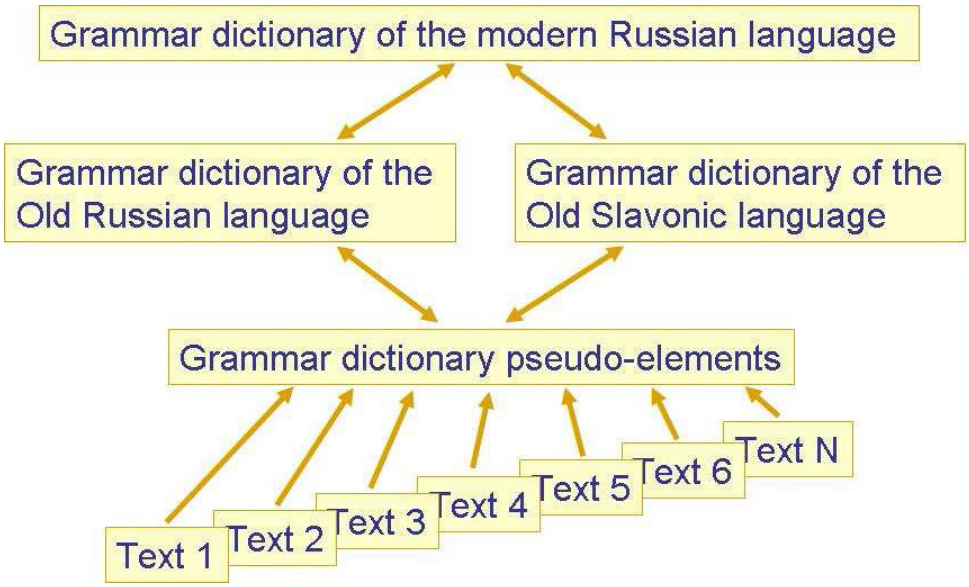


Fig. 7. Grammar dictionaries model.

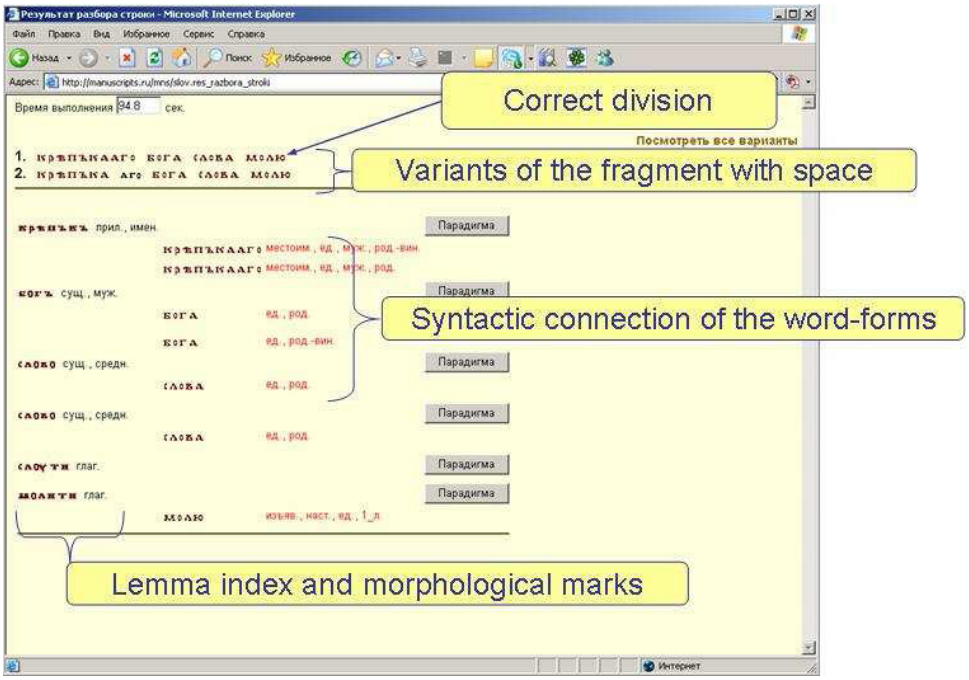


Fig. 8. Lemmatization: word index and word-form morphological marks.

2.7. The module of data exchange is intended for loading external data into the system and downloading the information from the Manuscript system in the format providing for the transmission of data into similar systems. The module of storage and processing of meta- and analytical data is intended for work with

archeographic, textologic, and other types of information on the manuscripts, texts and their fragments themselves and also for the formation of queries and retrievals on the basis of that information. The connection with the database is provided by the use of the unique identifiers of all objects of the system.

Both the modules are developed on the basis of the format of description and tagging of XML documents and recommendations of the TEI consortium. The use of those standards is the necessary condition of data exchange between the collections that are being created both in Russia and abroad.

2.8. At present the manuscript system has developed means for input, editing, analysis, making-up and publication of documents complicated by their structure and composition and also means of analysis for creation of queries and operations over retrievals thereby satisfying many requirements for the full-text multi-functional analytical systems of remote access. The use of the system modules in linguistic and textological research helps getting materials for the analysis of the Old Russian texts and their objects and drawing important fundamental conclusions.

## ACKNOWLEDGMENT

The work on the creation of IAS Manuscript is being carried out with the support from the Russian Foundation of Basic Research (Grant # 05-07-90217в); the work on the creation of the automated morphologic analyzer with the support of the Russian Foundation for the Humanities (Grant # 05-04-12408в).

## REFERENCES

[1] BARANOV, V.A., VOTINTSEV, A.A., GNUTIKOV, R.M., ZUGA, O.V., MIRONOV, A.N., NIKIFOROVA, S.A., OSHCHEPKOV, S.V., ROMANENKO, V.A. AND RYABOVA, E.V. 2003 Electronnyje izdanija drevnikh pis'mennykh pamjatnikov i tekhnologija sozdanija polnotekstovykh baz dannykh (Electronic Editions of Old Manuscripts and Technology of Creation of Full-Text Databases). In *Krug idej: electronnye resursy istoricheskoj informatiki*, Moscow, Russia, 234–260.

[2] BARANOV, V.A., VOTINTSEV, A.A., GNUTIKOV, R.M., MIRONOV, A.N., AND ROMANENKO, V.A. 2003. Spetsializirivannyj tekstovyj redactor "Manuscript" Sistemy obrabotki drevnikh rukopisej (Specialized Text Editor Manuscript of the System for Processing Old Manuscripts). In *Informatsionnyj bjulleten' assotsiatsii "Istorija i komp'juter 31*. Moscow, Russia, 159-165.

[3] BARANOV, V.A., VOTINTSEV, A.A., GNUTIKOV, R.M., MIRONOV, A.N., OSHCHEPKOV, S.V., AND ROMANENKO, V.A. 2004. Old Slavic Manuscript Heritage: Electronic Publications and Full-Text Databases. In *Proceedings of the EVA 2004 London (Electronic Imaging, the Visual Arts Conference & Beyond)*, London, Great Britain, July 2004, University College Institute of Archaeology. London, GB, 11.1-11.8.

[4] SOVREMENNYJE INFORMATSIONNYJE TEKHNOLOGII I PIS'MENNOJE NASLEDIJE: Ot drevnikh rukopisej k elektronnym tekstam. 2006. (Modern Informational Technologies and Written Heritage: From Ancient Manuscripts to Electronic Texts). *Proceedings of the* International Conference. Izhevsk, Russia, June 2006, 196 p.

[5] BARANOV, V. Information-Analytical System "Manuscript": technologies and tools of creation of electronic collections of ancient and medieval documents. In *Proceedings of the Dagstuhl Seminar 06491: Digital Historical Corpora - Architecture, Annotation, and Retrieval.* Schloss Dagstuhl Seminar 06491, Germany, December 2006; ISSN 1862-4405, Electronic resource: http://drops.dagstuhl.de/portals/index.php?semnr=06491.

[6] BARANOV, V., AND GNUTIKOV, R. Up-to-date means of access to full-text databases. In *Proceedings of the* 19th Joint International Conference of the Association for Computers and the Humanities, and the Association for Literary and Linguistic Computing, Urbana-Champaign, USA, June 2007. University of Illinois, 2007, 74–76. (Electronic resource: http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=199).

Victor Baranov
Department of Linguistics,
Izhevsk State Technical University,
Izhevsk, Studencheskaya Str. 6, Izhevsk, 426069, Russian Federation
baranov@udm.ru

Received September 2007