

**Information-Analytical System “Manuscript”:
technologies and tools of creation of electronic collections
of ancient and medieval documents**

Victor BARANOV

Linguistics Department
Izhevsk State Technical University
7 Studencheskaya Str., 426069 Izhevsk, RUSSIA
Laboratory of Computer-Aided Philological Research
Udmurtia State University
1 Universitetskaya Str., 426034 Izhevsk, RUSSIA
Email: baranov@udm.ru

0. The information-analytical system “Manuscript” (IAS “Manuscript”) is being created by a group of programmers and linguists of two universities of Izhevsk city. The developers are: Andrey Votintsev, Pavel Votincev, Roman Gnutikov, Andrey Lapin, Aleksey N. Mironov, Aleksey L. Mironov, Irina Mel’nikova, Sergey Oshchepkov, Vitaliy Romanenko.

0.1. The goal of work

Provision for the specialists in the field of ancient and medieval texts of a multi-functional tool enabling creation of electronic transcriptions of documents with complex structure and composition, publication of them on the Internet and use for basic research.

0.2. Requirements for the tool

The tool should enable processing deeply structured information on the texts, manuscripts, their objects and relationships between the objects and also their properties and values.

The tool (a) should support the multifold¹, multilingual², and multitext³ character, (b) should ensure storage of metadata⁴, analytical⁵, linguistic⁶ data, commenting and refer-

¹ Multifold character – use of symbols of various national alphabets.

² Multilingual character – the ability to store the information on units (properties, values, commentaries) and availability of an interface in various languages.

³ Multitext character – the possibility of creation of a query with the indication of texts/manuscripts where search and retrieval of units from the entire corpus of manuscripts/texts or from the part of corpus chosen by the user will be done.

⁴ Metadata – the properties or values of manuscripts/texts. Metadata should ensure (a) creation of a query with the indication of the properties and values of the texts/manuscripts where search will be done and also (b) retrieval of units from the manuscripts/texts possessing the values specified in the query.

⁵ Analytical data – the properties and values of manuscript/text fragments. The analytical data should ensure the possibility of creation of a query with the indication of the properties and values of the fragments where search will be done and also retrieval of units from the fragments possessing the values specified in the query.

⁶ Linguistic data – information on the form, properties and values of the textual linguistic units.

ence information, (c) should have means of work with the dictionaries⁷, unit relationships⁸ and graphic objects⁹.

The electronic collection or a library created with the help of this tool should have realized: (1) demonstration of manuscripts and texts and (2) use of reference materials, which implies: (a) organization of queries, (b) search and (c) ordering of data, (d) visualization of retrieval results, (e) creation of comparative lists, (f) creation of the critical comparison of the manuscripts of the same text, (g) creation of the critical apparatus of the edition, (h) show of various types of text.

1. *The technological solutions of the information-analytical system “Manuscript”* (URL: http://manuscripts.ru/index_en.html; Fig. 1)



Fig. 1: Title page of the portal “Manuscript”

IAS Manuscript comprises some modules with different functions interacting with the common database.

1.1. *The main principles of construction of IAS Manuscript*

⁷ Dictionaries – the components of the database containing the information on the form, properties and values of the dictionary units.

⁸ Relationships – information on the relationships between the texts/manuscripts objects and their units and also about the properties and values of the relationships.

⁹ Graphic files are used to store the scan-copies of the manuscript pages and graphic text elements.

1.1.1. The model of the database

The units of the database are analogs of the manuscripts, texts and their objects. The examples of units-objects are symbols, blocs, pages, sheets, word forms, phrases, fragments, manuscripts, texts, dictionaries, dictionary units etc. Three types of unit - symbol, manuscript and text - are obligatory for any document.

The symbol is a representative of the document symbols; the manuscript is the analog of the material object of the document and the text is the analog of the work.

The relationships existing in reality between the objects of the manuscripts and texts are represented in the form of relationships between the database units.

Units and relationships modeling the same object field are organized into a hierarchy that is more often a tree. The basic hierarchies are geometrical and linguistic. The first organizes the types of units corresponding to the real hierarchic relationships of the objects in the manuscript and text (Fig. 2) and the second according to the hierarchic relationships existing between the word forms, combinations of words, syntagmas, phrases and the text.

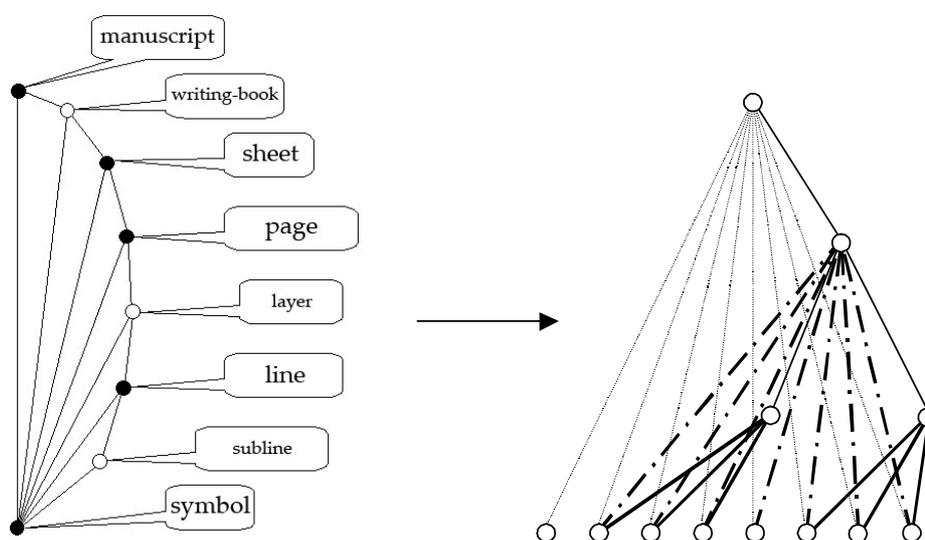


Fig. 2: Model of the geometrical hierarchy

According to the subject field some hierarchies, for example linguistic, look like a net (Fig. 3).

The availability of *unit values and properties* and also *relationship values and properties* allow storage of the characteristics of the objects of the manuscripts, texts and their objects and also information on the characteristic of the relationships between them in the database.

Hence, the database stores the symbols united into various units that, in their turn, are organized into hierarchies or subnets where the symbol is the lowest level and the manuscript or text is the highest. It is clear that one unit can include sub-units located close or at a distance.

To represent the electronic transcription maximum adequately to the original, an expanded set of old Slavonic symbols including the low-case, capital and superscript symbols of the alphabet, superscripts, titlos, ligatures, symbols of number, non-letter symbols in the line, damage symbols is created in Private Use Area. To represent the authors' versions of symbols, there are additional fonts in which glyphs have the same codes as in the main one.

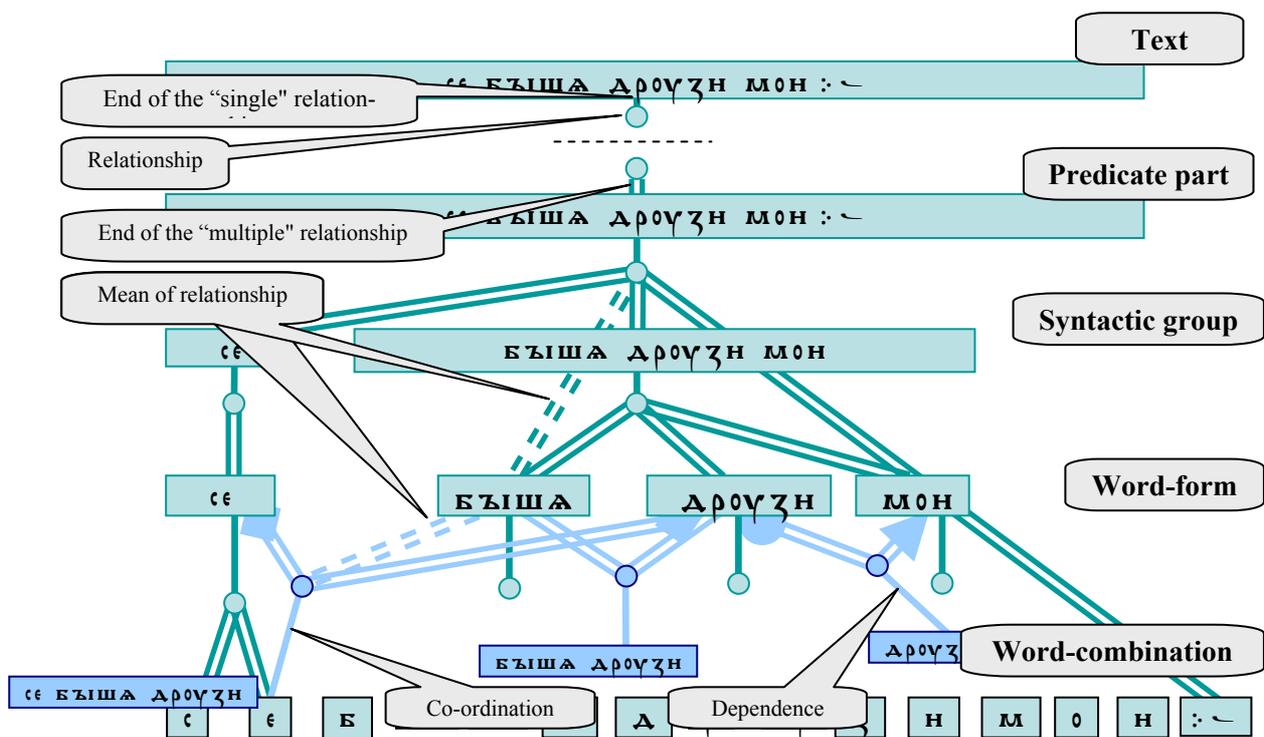


Fig. 3: Net of linguistic relationships

1.1.2. Additional solutions

The decision was made that, to store the meta- and analytical information on the manuscripts, texts and their fragments, the technologies of Open Archive Initiative (OAI) and Text Encoding Initiative (TEI) will be used. The technology OAI should ensure indexing of the existing texts and their fragments; the formats TEI the storage of the meta- and analytical information and also the transmission of this information and texts themselves into other similar systems.

1.2. The Modules of IAS Manuscript (see Section "Tools" of the portal) (Fig. 4):

- specialized editor;
- module of electronic publications;
- module of query and retrieval;
- module of dictionaries and lemmatizer;
- module of data exchange with other systems of a different format;
- module of storage and processing of meta and analytical data;
- module of authorized access;
- module of download of texts and reference materials for printed publications;
- module of storage and exchange of documents.

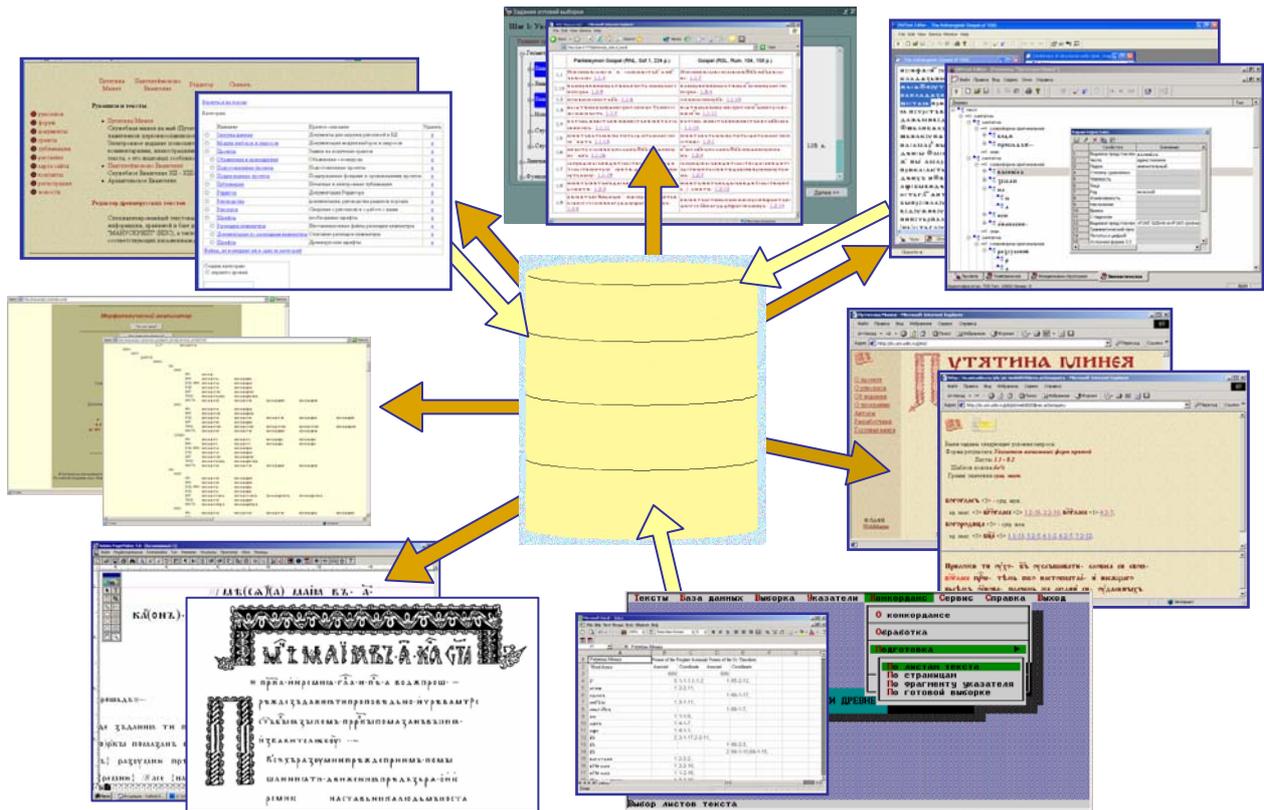


Fig. 4: Model of the Manuscript system

The modules of the manuscript system are connected between them into a common technological chain that ensures the preparation of the electronic collection satisfying many of the requirements listed in Section 0.2.

1.2.1. *The specialized editor OldEd (see Section “Tools” of the portal).* The editor is intended for: (a) the input and editing of the manuscript units, (b) fragmentation of the manuscript and text into units, (c) establishment of hierarchic and other relationships between units, (d) assignment to the units of the properties and values that they possess, (e) establishment of relationships of the manuscript units with the units of dictionaries, (f) lemmatization of texts and other operations (Fig. 5).

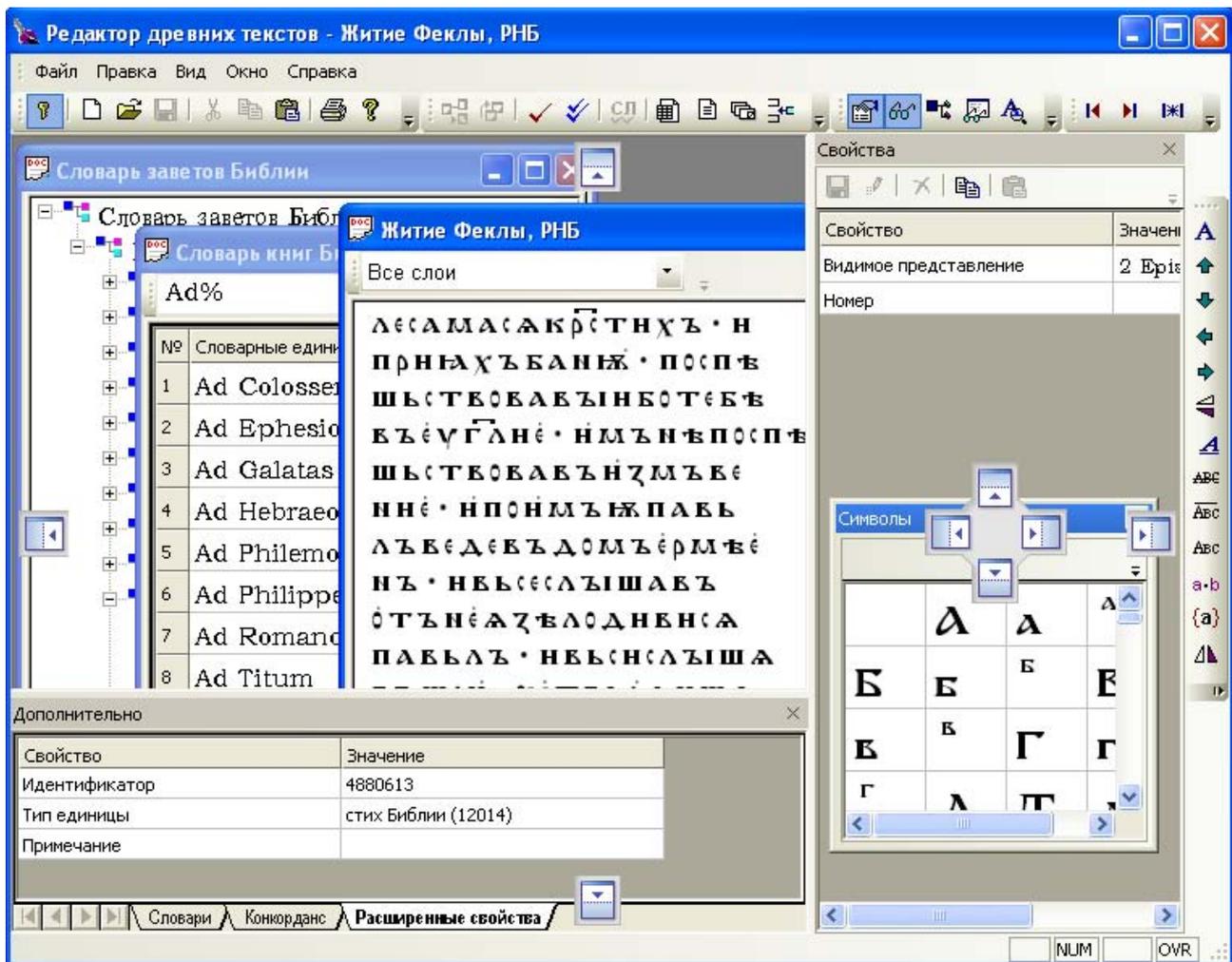


Fig. 5: Editor OldEd: main panels

The specially developed code-font system allows a maximum approximation of the text transcription to the original and representation of any graphic variants of symbols. The available dictionary of symbols ensures the storage and use of graphic and phonetic (phonological) properties and values of each of the symbols at making queries.

The editor ensures fragmentation of the manuscripts and texts and assignment of values to the highlighted objects. If the fragment is an object that exists or can exist in many manuscripts or texts, the editor ensures the assignment of the value not to the text object itself, but to the units of the corresponding dictionary (Fig. 6).

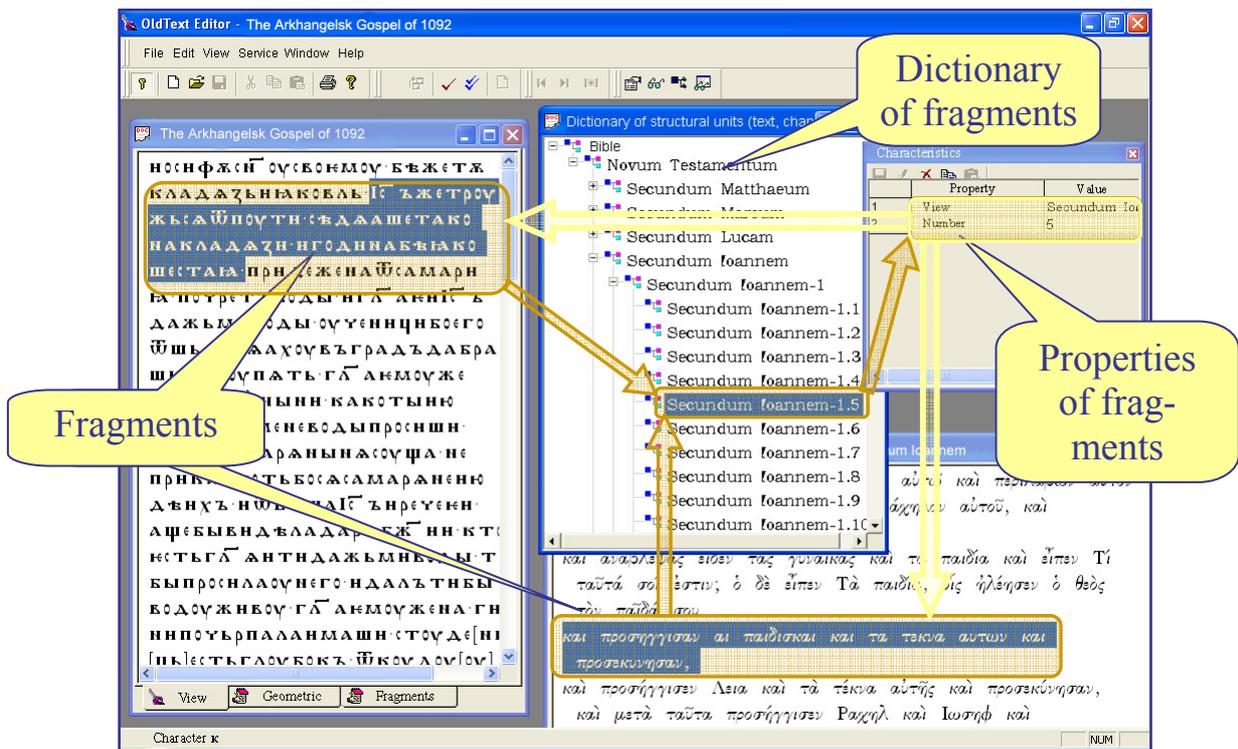


Fig. 6: Editor OldEd: Fragmentation of the manuscript texts into units and relationships with the dictionary units

The important property of the editor is the possibility of establishing and viewing the relationships between units. At present viewing is realized as a tree where all unit relationships are shown (Fig. 7).

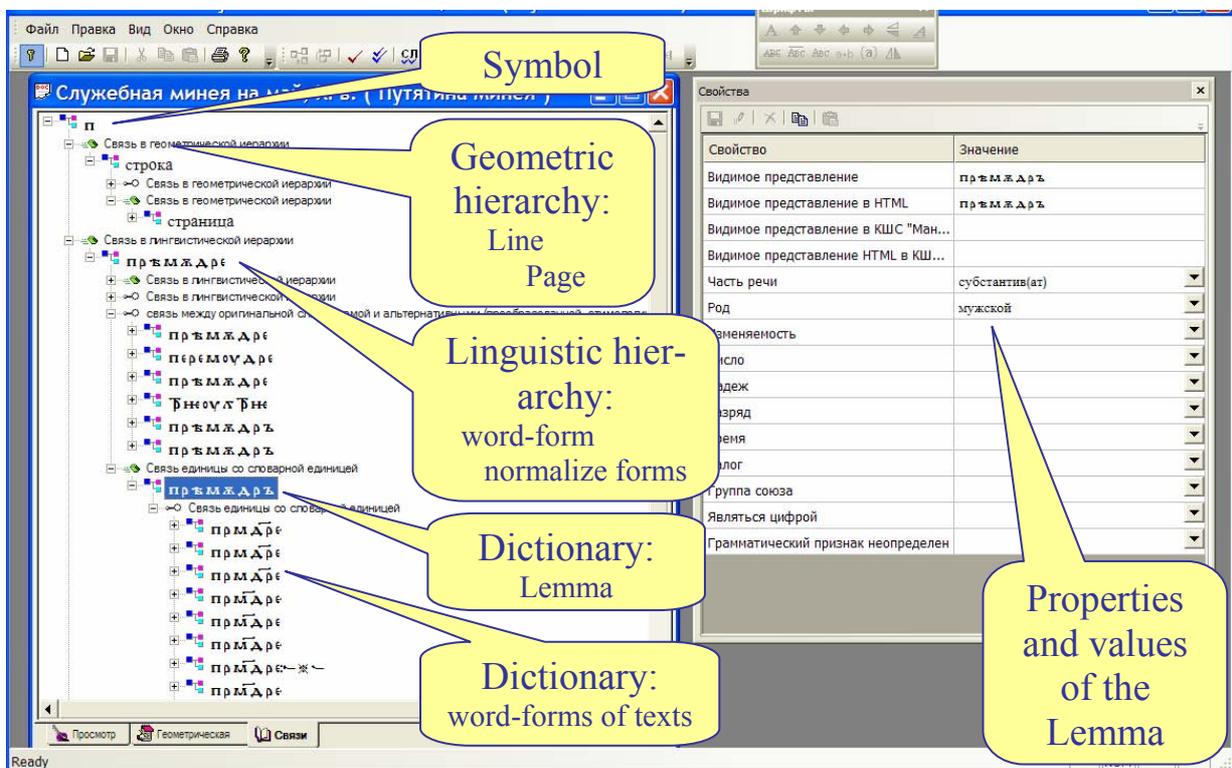


Fig. 7: Editor OldEd: Visualization of unit relationships

One of the important operations provided by the editor is the preparation of the dummy copy of the future electronic edition. The preparation of the breadboard model of the edition allows the location of the text fragments on the page according to the original (Fig. 8), which ensures the necessary visualization of the page in the electronic edition on the site (Fig. 9).

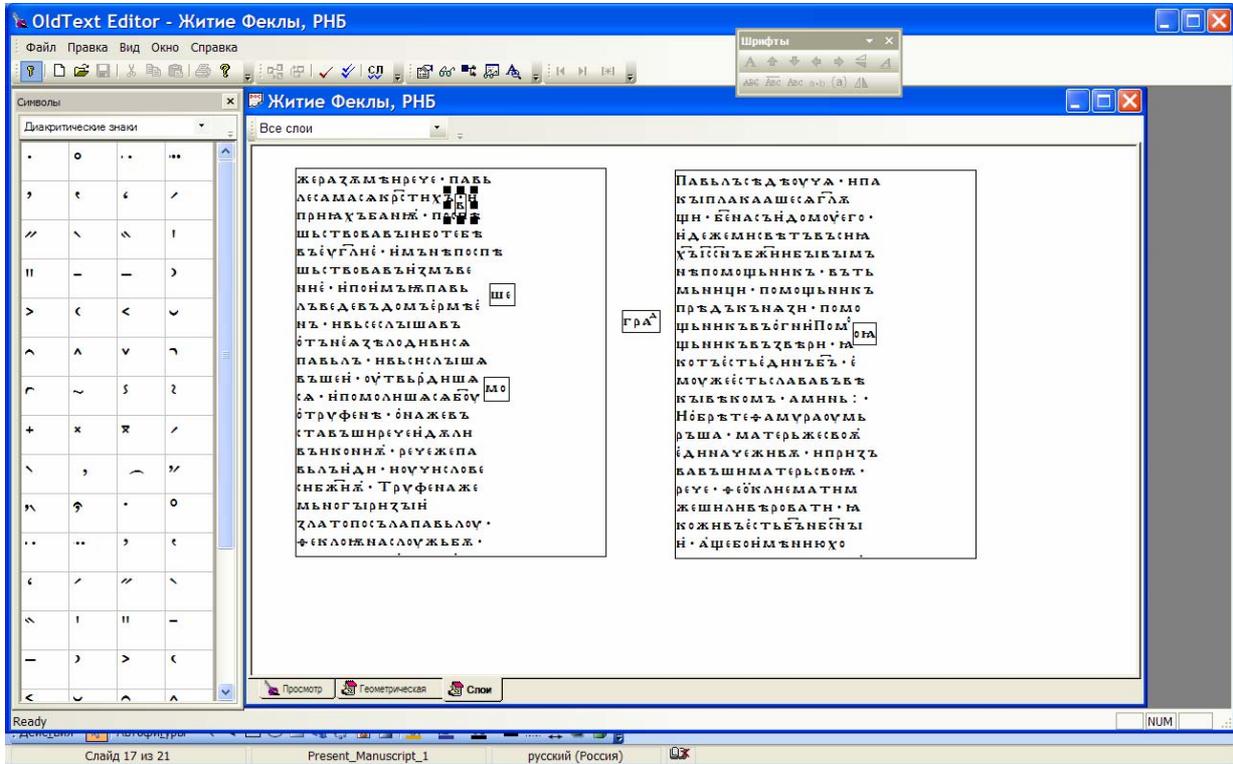


Fig. 8: Editor OldEd: Page layout

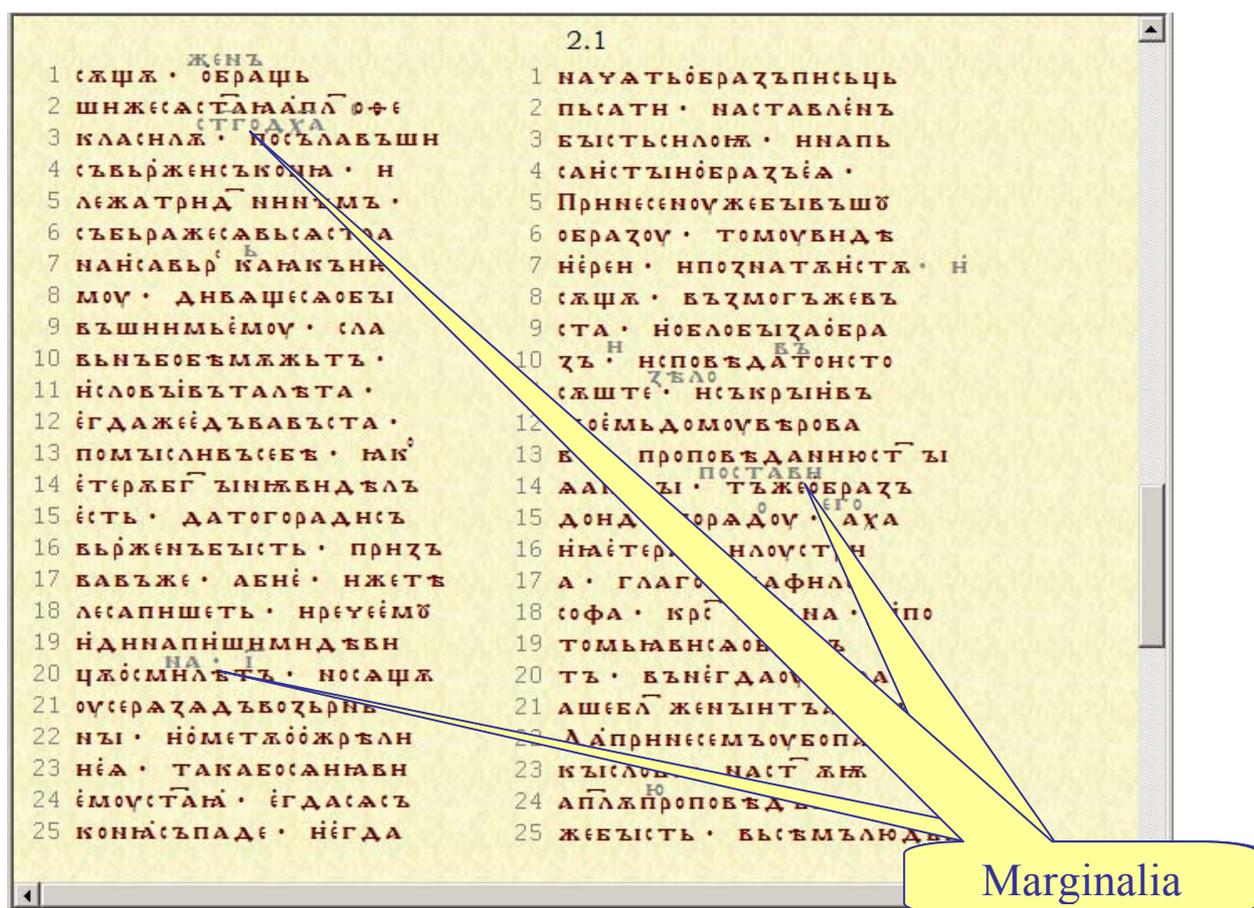


Fig. 9: Electronic edition: Result of page layout

The important property of the editor is the transfer of the edited text/manuscript transcription into the electronic edition. Together with the provision of the distributed and distant editing of texts in the database the editor gives a possibility of seeing and evaluating by the independent users the results of editing immediately in the course of work.

Hence, the editor enables the text input and editing, its fragmentation, assigning values to the text/manuscript units, establishing relationships between the objects, establishing relationships between objects and dictionary units (Fig. 10), establishing the relationship between the word form and its lemma and assigning them grammar values in the automated mode, transmission of the text/manuscript to the site.

Форма	В...	Форма	Грам. признаки	Вероят.
СВѦТОУЮ	<input type="checkbox"/>	СВѦТЬА ОУЮ (СВѦТЬАЪ)	прил., местоим., дв., муж., род	0
Н	<input type="checkbox"/>	СВѦТЬА ОУЮ (СВѦТЬАЪ)	прил., местоим., дв., муж., местн	0
СВѦТЬЛОУЮ	<input checked="" type="checkbox"/>	СВѦТЬА ОУЮ (СВѦТЬАЪ)	прил., местоим., дв., средн., род	60
НѦАЮ	<input type="checkbox"/>	СВѦТЬА ОУЮ (СВѦТЬАЪ)	прил., местоим., дв., средн., местн	0
ПАСЪХЫ	<input checked="" type="checkbox"/>	СВѦТЬА ОУЮ (СВѦТЬАЪ)	прил., местоим., ед., жен., вин	40
	<input type="checkbox"/>	СВѦТЬА ОУЮ (СВѦТЬАЪ)	прил., местоим., дв., жен., род	0
	<input type="checkbox"/>	СВѦТЬА ОУЮ (СВѦТЬАЪ)	прил., местоим., дв., жен., местн	0

Fig. 10: Editor OldEd: Lemmatization of word-forms

1.2.2. The web-module of electronic publications (URL: http://manuscripts.ru/mns/portal.main?p1=1&p_lid=2). The main component of the electronic edition is the query page that allows selection of a collection of manuscripts, the collection text and specifying the query criteria. The query result can be (a) various indexes: direct and reverse indexes of the initial forms and word forms, quantitative indexes, indexes of the incipites; (b) texts of various types: original, normalized, diplomatic, transliterated. The main among the linguistic indexes is the complete index of words and word forms which provides the grammar signs of units, their number and addresses (Fig. 11). I would like to stress that the search fields provided for the user give a possibility of searching effectively enough the necessary information in large volumes of texts and manuscripts (Fig. 12).

The screenshot shows a web interface for 'Electronic Editions' with the following elements:

- Header:** 'Electronic Editions', 'Home', 'version: RUS ENG', 'Slavonic manuscript heritage' logo.
- Search Criteria:**
 - Коллекция:** Коллекция русских житий
 - Рукопись:** Житие Александра Невского
 - Forma результата:** Указатель начальных форм
 - Листы от:** 67 до 68
 - Образец поиска:** А% (Selected: Словоформа)
 - Грам. значения:** %АБВБГ ДЄЖЗНІКЛМНОПРСТ ОУѦФХЦЧШЩЪЫЬЪЮЯѦѦѦ
- Search Results:**
 - А 3 - союз
 - А 68.2.1-32, 28, 13
 - АБНН 3 - нар.
 - АБНН 67.1.1-18, 29, 18
 - АБЛАБНН 4 - личн. муж.
 - ед. им. 1 АБЛАБНН 67.1.1-3 зват. 1 АБЛАБНН 67.2.1-24 род.-вин. 1 АБЛАБНН 67.1.1-ню 67.2.1-23
 - АБЛАБНН 1 - личн. муж.
 - ед. им. <А>АБЛАБНН 67.2.1-34

Annotations in yellow boxes point to:

- Collections & Manuscripts:** Points to the 'Коллекция' and 'Рукопись' dropdowns.
- Search criteria:** Points to the search form fields.
- Search result:** Points to the list of search results.

Fig. 11: Electronic edition: search page

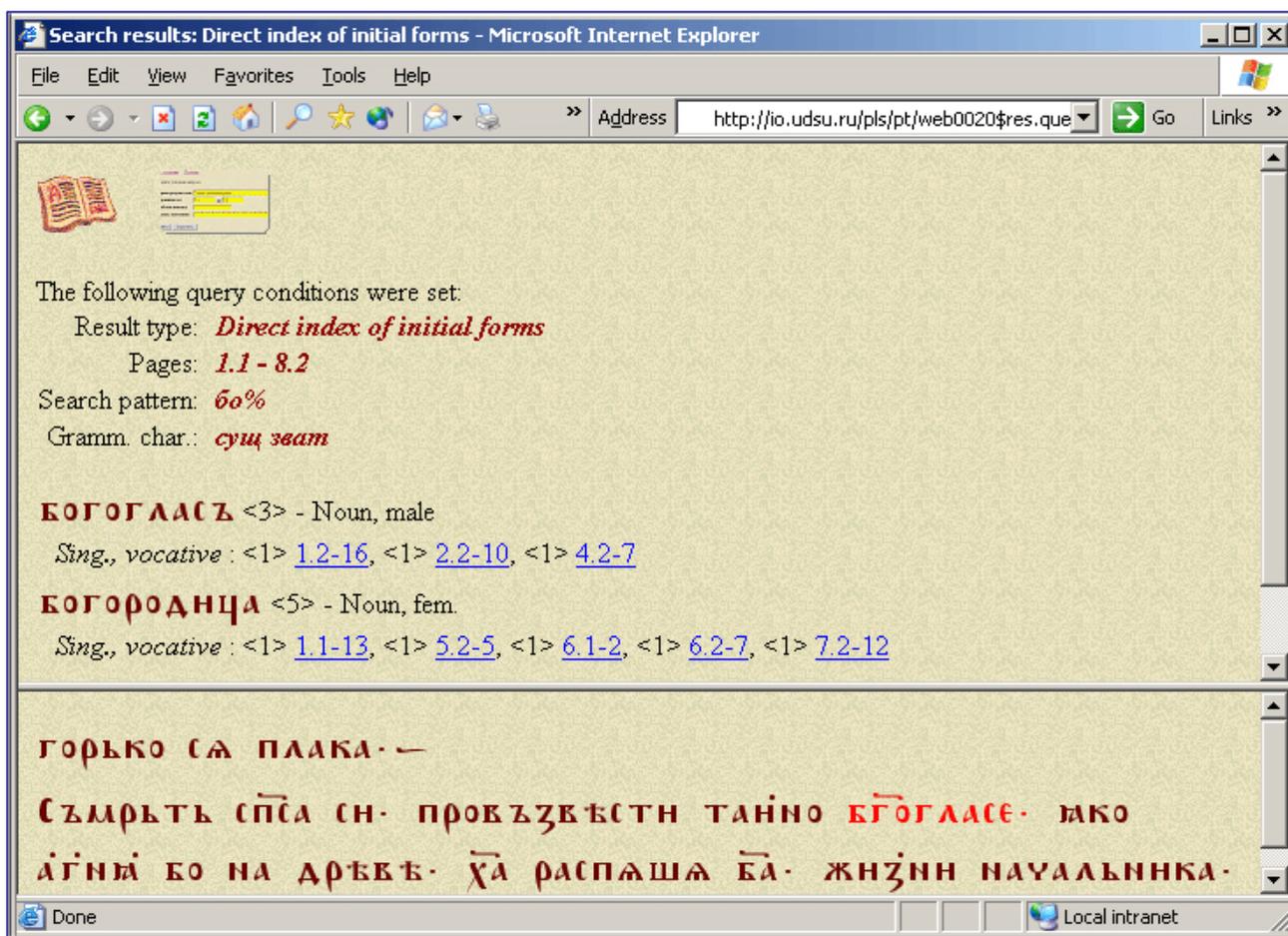


Fig. 12: Search results: word index and concordance

1.2.3. *The module of retrieval and queries* (see Section “Tools” of the portal). If in the previous module the possibilities of query are predetermined and the user cannot change them or expand, the module of retrieval and query is intended for the formation of a free query and enables (a) selection of units, their properties and values; (b) setting the relationships between the units; (c) setting the form of output; (d) saving the query and (e) use the query result for the creation of the next retrieval.

The work on the preparation of the retrieval includes (a) selection of the hierarchy, (b) selection of the unit, (c) setting the unit properties and values, (d) setting the properties and values of the dictionary units, (e) saving the query, (f) specifying the composition of the query result, the priority of its units and their sorting, (g) displaying the result (Fig. 13, 14, 15).

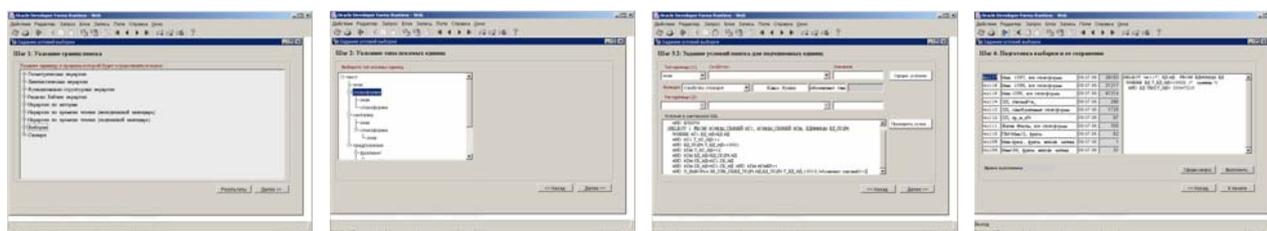


Fig. 13: Steps of the preparation of the query for getting the retrieval

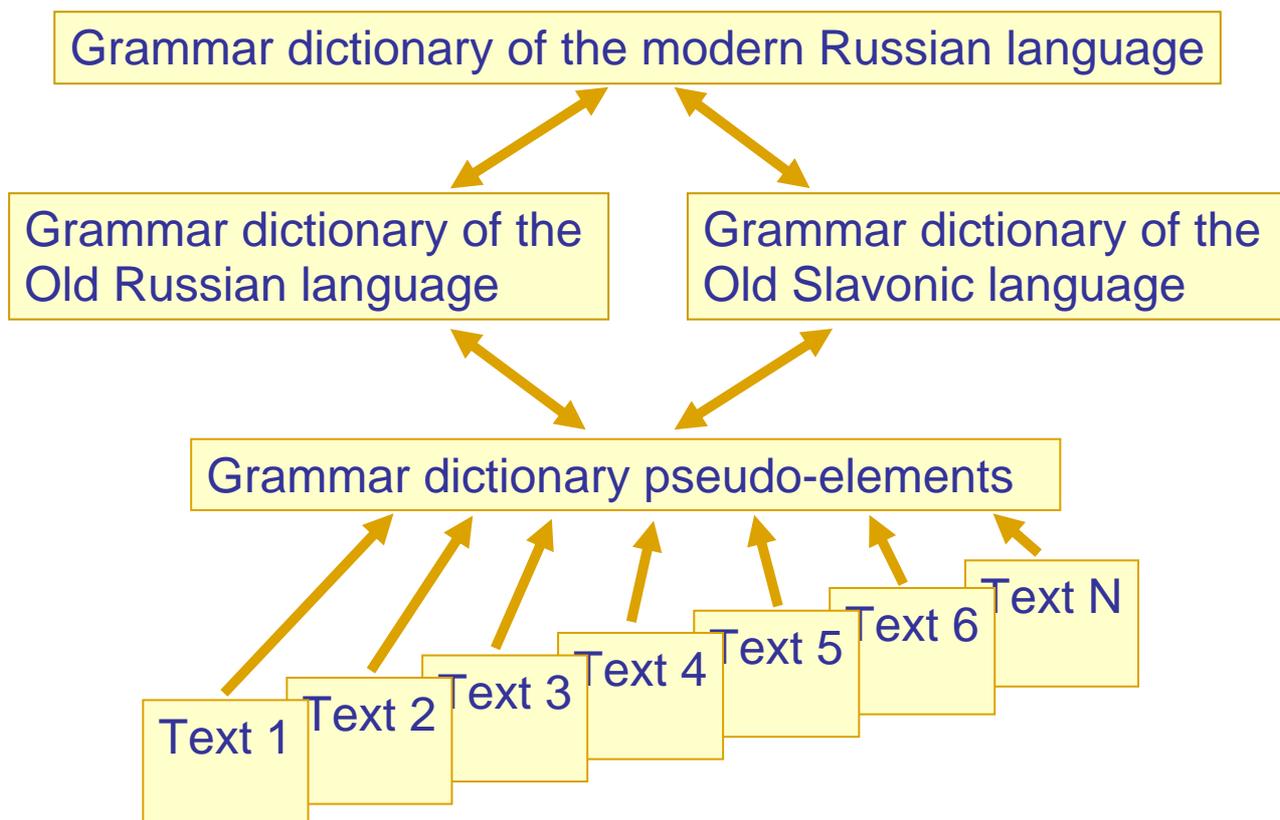


Fig. 16: Model of grammar dictionaries

The main functions of the module (*Fig. 17*) are now (a) the morphologic analysis (lemmatization) - bringing the Old Russian word forms to the initial forms of words (lemmas) (*Fig. 18*) and (b) the morphologic synthesis - obtaining paradigms of lemmas (*Fig. 19*).

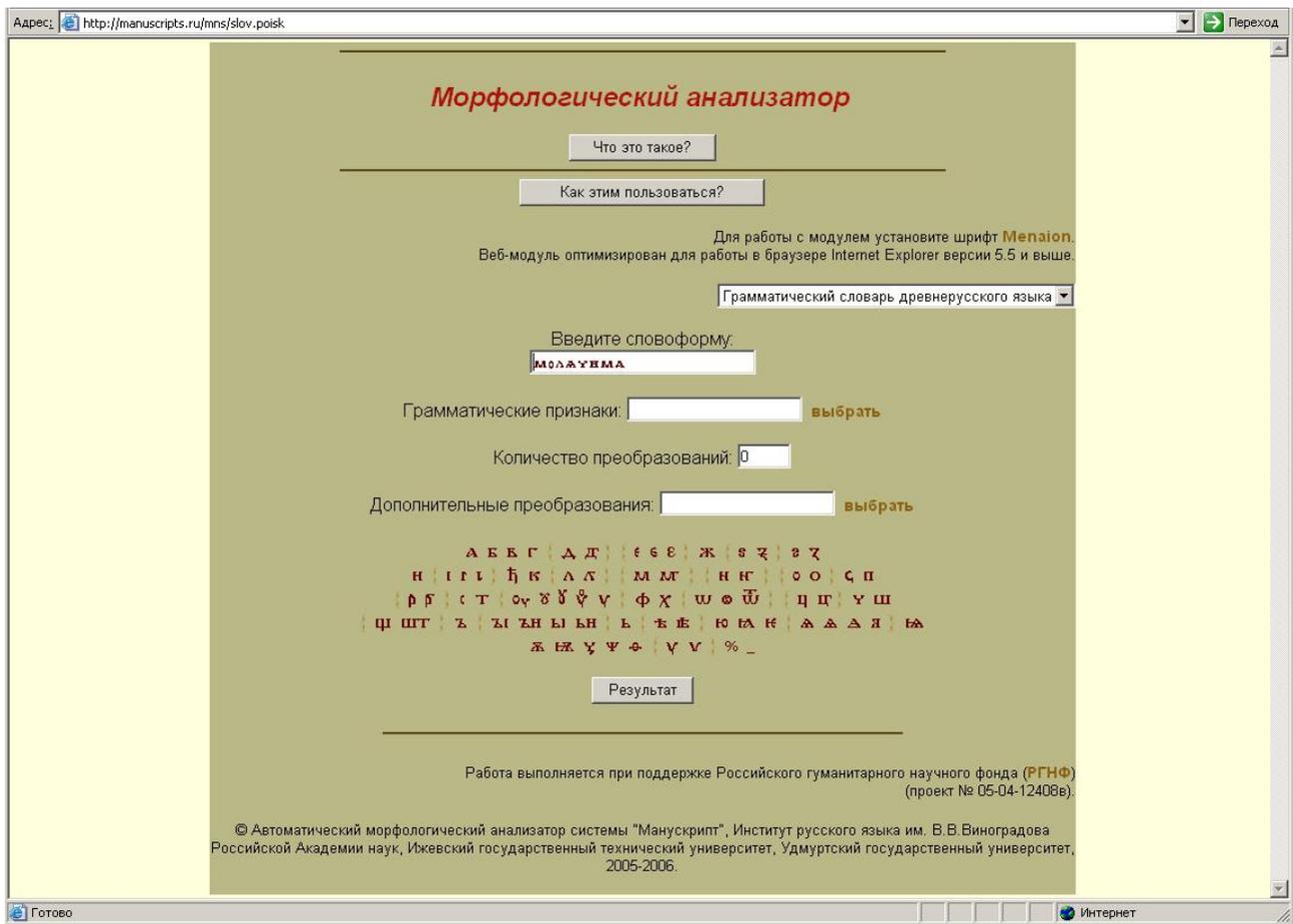


Fig. 17: Lemmatizer: the page of input of the word form and parameters for its recognition

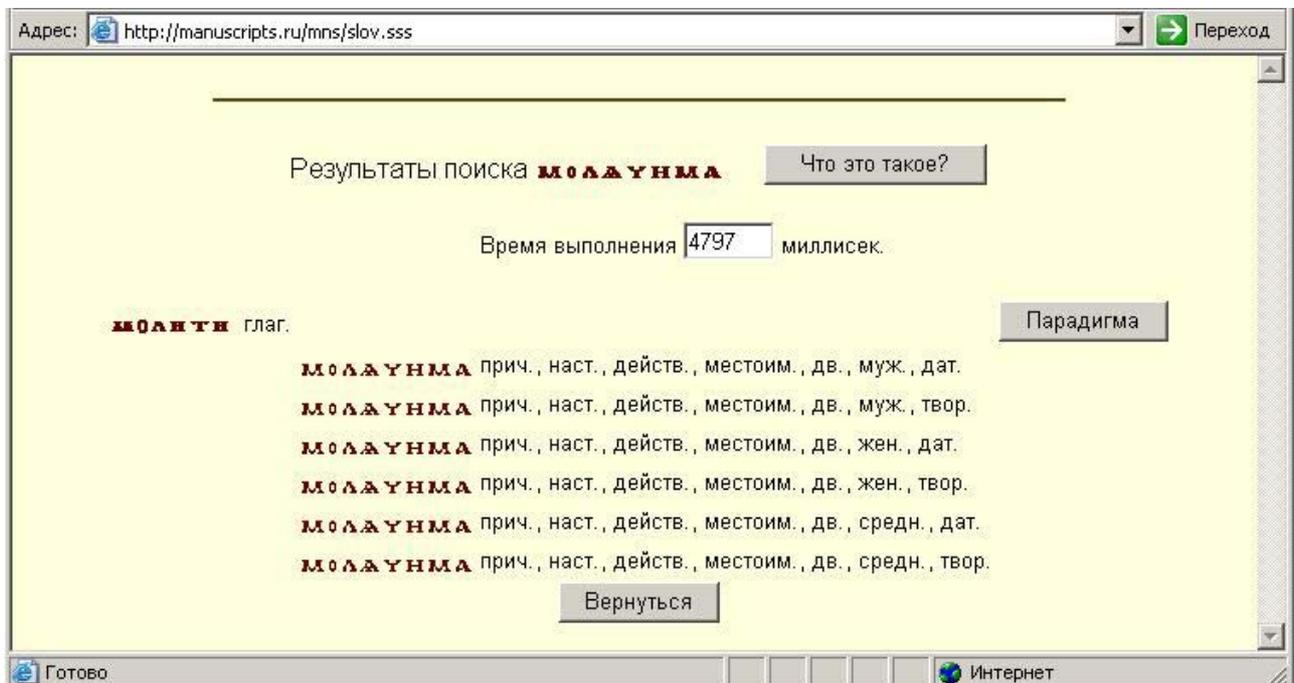


Fig. 18: Lemmatizer: finding a lemma “молити” and determination of the grammar signs of the homofoms

Адрес: http://manuscripts.ru/mns/view_paradigm?m_id=1&d_id=1&osn_id=38627645 Переход

2_п. **МОЛАНТА**

прич.	наст.	действ.	имен.	ед.						
				муж.						
				им.	МОЛАА					
				вин.	МОЛААУЪ	МОЛААЦЬ				
				род.-вин.	МОЛААУА	МОЛААЦА				
				род.	МОЛААУА	МОЛААЦА				
				дат.	МОЛААУОУ	МОЛААЦОУ				
				твор.	МОЛААУЪМЪ	МОЛААЦЬМЪ				
				местн.	МОЛААУН	МОЛААУЪ	МОЛААЦН	МОЛААЦЪ		
				жен.						
				им.	МОЛААУА	МОЛААЦА				
				вин.	МОЛААУОУ	МОЛААЦОУ				
				род.	МОЛААУА	МОЛААУН	МОЛААУЪ	МОЛААЦА	МОЛААЦН	
				дат.	МОЛААУН	МОЛААЦН	МОЛААУОН	МОЛААУОЮ	МОЛААЦЮ	
				твор.	МОЛААУН	МОЛААУЮ	МОЛААУОН	МОЛААУОЮ	МОЛААЦЮ	МОЛААЦЮН
				местн.	МОЛААУН	МОЛААУЪ	МОЛААЦН	МОЛААЦЪ		
				средн.						
				им.	МОЛААУѣ	МОЛААУо	МОЛААЦѣ	МОЛААЦю		
				вин.	МОЛААУѣ	МОЛААУо	МОЛААЦѣ	МОЛААЦю		
				род.-вин.	МОЛААУА	МОЛААЦА				
				род.	МОЛААУА	МОЛААЦА				
				дат.	МОЛААУОУ	МОЛААЦОУ				
				твор.	МОЛААУЪМЪ	МОЛААЦЬМЪ				
				местн.	МОЛААУН	МОЛААУЪ	МОЛААЦН	МОЛААЦЪ		
				мн.						
				муж.						
				им.	МОЛААУН	МОЛААЦН				
				вин.	МОЛААУН	МОЛААЦН				
				род.-вин.	МОЛААУЪ	МОЛААЦЬ				
				род.	МОЛААУЪ	МОЛААЦЬ				
				дат.	МОЛААУЪМЪ	МОЛААУОМЪ	МОЛААЦЪМЪ	МОЛААЦЮМЪ		
				твор.	МОЛААУН	МОЛААЦН				
				местн.	МОЛААУНХЪ	МОЛААЦНХЪ				
				жен.						
				им.	МОЛААУА	МОЛААУН	МОЛААУЪ	МОЛААЦА	МОЛААЦН	

Готово Интернет

Fig. 19: Construction of the paradigm of the present participle of the verb “молити”

1.2.5. 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.

1.2.6. The module of authorized access enables the differentiation of the rights of users at work with the system modules. The differentiation of the rights of access is used in the module of electronic publications (provision of the possibility of work with a part of document or with the entire document), the module of retrieval and queries (provision of the possibility of work only with those objects that are determined by the system administrator and the developer of the electronic copy of the document), in the editor (provision of the editing rights only for specified objects of the database) and in some other cases.

1.3. Hence, at present IAS Manuscript, having developed means of input, editing, analysis, prototyping and publication of documents and also creation of queries and processing retrievals satisfies many requirements for the multifunctional full-text systems intended for the creation of electronic collections and libraries.

2. Preparation of electronic publications and use of them

2.1. At present the collections present some tens of ancient Russian written treasures of the XI-XVII cent. (see Section “Manuscripts” of the portal), namely: the texts for divine service of Gospels, Mineia, Triodia, chronicles, hagiographies and Old Russian translations.

The main part of publications were prepared and are being prepared by the teachers, post-graduate and graduate students of Udmurtia State and Izhevsk State Technical Universities (Oksana V. Zuga, Elena V. Ryabova, Irina M. Nekipelova, Anna A. Korotayeva, Svetlana B. Gerashtshenko and others).

Together with our colleagues from the Institute of the Russian Language after V.V. Vinogradov (the work leaders Prof. Aleksandr M. Moldovan, Corresponding Member of the Russian Academy of Science and Candidate of Philological Sciences Anna A. Pichkhadze) we prepare a collection of Old Russian translations (URL: http://manuscripts.ru/mns/portal.main?p1=11&p_lid=2).

2.2. The electronic publications are on various stages of preparation and, as a result, accessibility to the external user. For example, the completed work is the electronic edition of the service Menaion for May of the XI cent. (Putyatina Mineya) (URL: http://manuscripts.ru/ptm_en/) that includes the text in various forms, a number of commenting and reference materials, enables getting various retrievals over the text and its units. The work at the similar edition of the Panteleymon Gospel of the XII-XIII cent. comes to the end (the author Oksana V. Zuga) (URL: http://manuscripts.ru/mns/portal.main?p1=21&p_lid=2&p_sid=1) (*Fig. 20*). The main distinction of these publications from others in our collections is the presence of full grammar word and form indexes made completely manually.

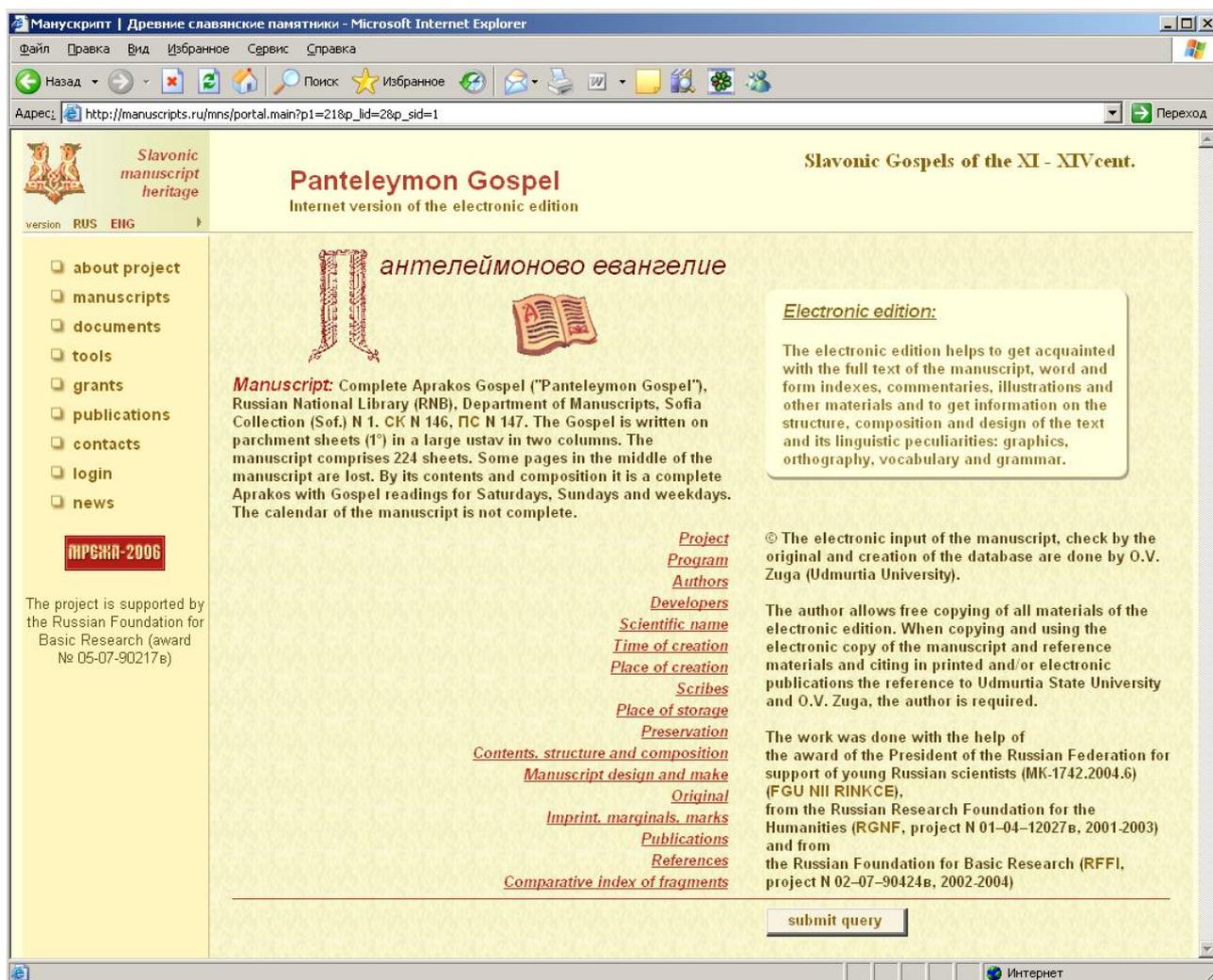


Fig. 20: Electronic edition of the Panteleymon Gospel

Another stage of preparation of the publication is presented by the publications where in addition to work with texts and form indexes the users have a possibility of work with fragments. For example, with the help of the module of retrieval and queries it is possible to get a comparative index of fragments of the May Service Menaion that, in particular, allows establishing discrepancies in the input texts in some copies of the XI-XIV cent. from various time periods (Fig. 15).

And finally, the initial stage of the preparation of texts and manuscripts in IAS Manuscript is presented by the publications where the input and proofreading were done by the original or its copy. Already this type of publication gives to the user flexible possibilities of use of materials for the acquaintance and study of the manuscript text. For example, the electronic transcriptions of the manuscripts of the XI century give a possibility, having done a corresponding query in the module of electronic publications, to get acquainted with the texts in two forms: without separation into word forms and with separation into word forms. In addition, the user can obtain the ordered retrieval of word forms, having specified, if necessary, the range of sheets and the masque of the word form in the relevant fields of the query form. The result is given in the form of the direct or reverse index of word forms and also in the form of quantitative index. After getting the index through the address of use of the word form it is possible to pass to the context where the given word form is used (Fig. 21).

References

- BARANOV, V.A., VOTINTSEV, A.A., GNUTIKOV, R.M., MIRONOV, A.N., OSHCHEPKOV, S.V., ROMANENKO, V.A. (2004) Old Slavic Manuscript Heritage: Electronic Publications and Full-Text Databases // EVA 2004 London (Electronic Imaging, the Visual Arts Conference & Beyond). Conference Proceedings. University College London. Institute of Archaeology. Principal Editor: James Hemsley. London, 2004. 11.1-11.8.
- 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) Elektronnyje izdanija drevnikh pis'mennykh pamjatnikov i tekhnologija sozdaniya polnotekstovykh baz dannykh (Electronic Editions of Old Manuscripts and Technology of Creation of Full-Text Databases). Krug idej: elektronnye resursy istoricheskoy informatiki, Moscow, pp. 234–260.
- BARANOV, V.A., VOTINTSEV, A.A., GNUTIKOV, R.M., MIRONOV, A.N. and ROMANENKO, V.A. (2003) Spetsializirivannyj tekstovyy redactor "Manuscript" Sistemy obrabotki drevnikh rukopisej (Specialized Text Editor Manuscript of the System for Processing Old Manuscripts). Informatsionnyj bjulleten' assotsiatsii "Istorija i komp'yuter 31: 159-165.
- SOVREMENNYJE INFORMATSIONNYJE TEKHNologii I PIS'MENNOJE NASLEDIJE: Ot drevnikh rukopisej k elektronnym tekstam (Modern Informational Technologies and Written Heritage: From Ancient Manuscripts to Electronic Texts). Materials of the International Conference. Izhevsk State Technical University. Principal Editor: Victor A. Barabov. Izhevsk, 2006. 196 p.