Development of the processing and visualization technologies for the linguistic information in the manuscript system: lemmatization

Victor A. Baranov[1], Aleksey N. Mironov[2], Aleksey N. Lapin[2], Irina S. Melnikova[2], Anastasiya A. Sokolova[2]

[1]Izhevsk State Technical University – Studencheskaya Str. 7 – 426069 Izhevsk – Russia

[2]Udmurtia State University – 426034 Universitetskaya Str. 1 – Izhevsk – Russia

Abstract

The article deals with an experience of development and creation of an automatic morphological analyzer of the Old Russian designed for automatic lemmatization of Old Russian texts. Special attention is given to the technological and linguistic solutions of the structurization of the dictionary units in the database and also to the methods of elimination of the graphic-orthographic variance of the word forms. The article contains the description of the web-modules of the system MANUSCRIPT providing for data search on the basis of the corpus of Old Russian texts and for visualization of the lemmatization results.

Keywords: morphologic analysis, data visualization, lemmatization, Medieval Slavonic, Old Russian.

1. Introduction

It is well-known that the development and creation of the means of automatic morphological analysis of texts may be grounded on various basic principles and considerably different linguistic material: (1) on the precedents accumulated during manual analysis of word forms, (2) on the algorithms and rules of generation of morphological forms, (3) on the full morphological (grammar) dictionary of the language.

In the case of construction of the information system for the automatic grammar analysis of the ancient Slavonic text in its original form or a from close to the original, the matter gets complicated by some circumstances: (1) the impossibility of collection of all word forms and their variants available in the manuscripts due to the today's absence of the satisfactory number of full-text electronic resources that would satisfy the linguistic requirements for their reproduction and due to very considerable variance of writing; (2) the absence of the morphological (grammar) dictionaries of the Old Slavonic and Old Russian languages; (3) the absence of the full electronic indexes of words of these languages.

At the same time, the active preparation of the full-text electronic resources on the basis of the ancient and medieval Slavonic written treasures requires the quickest creation of the automated systems of morphological analysis of this type of text, as without them neither syntactic nor semantic analysis is possible. This is why a group of developers and researchers of some Russian organizations now work at the project "Automatic Morphological Analyzer of the Old Russian Language" that envisages the development and creation of the automatic lemmatizer that is based on the full morphological (grammar) dictionary of the Old Russian language.

The analyzer must perform (1) the automatic analysis of the word forms – finding their lemmas, (2) the automatic synthesis of the word forms on the basis of the lemma and (3) the determination of the grammar characteristics of the lemmas and word forms.

The final goal of this work will be the creation of the tool (1) for the automatic lemmatization of the Old Russian texts and for the automatic construction of the paradigms of the dictionary of lemmas (2) for ensuring the distributed work on filling and editing the dictionaries.

The main project objectives: (1) the creation of the database of paradigms of the variable parts of speech; (2) the creation of the module of input end editing of the words and their paradigms; (3) the automatic elimination of the variance of the word forms; (4) the creation of the web-modules of lemmatization of the word forms and text fragments.

At present the lemmatizer is a module of the information-analytical system MANUSCIPT (http://manuscripts.ru/index_en) comprising the database of linguistic objects, the system of programs and web-interfaces.

2. Technological solutions

2.1. Components of the module of dictionaries

The database of dictionaries is a system of grammar dictionaries which units (elements) are interconnected. The main dictionaries are the following: grammar dictionary of the Old Russian language (GDORL), grammar dictionary of the modern Russian language (GDMRL)[1] and the grammar dictionary of pseudo-elements.

The software system is designed for the organization of input, storage and editing of grammar dictionaries, for the implementation of queries over the dictionaries, for establishing and supporting of the relationships of the text word forms with the elements of the dictionaries, the relationships between the objects of various dictionaries and the relationships between the elements inside the dictionaries.

The web-interfaces are intended for the distant and distributed work with the dictionaries and for input, editing and organization of queries and presentation of the results to the end user.

2.2. Units of dictionaries

The dictionaries comprise the following objects:

- *the stem* is the dictionary element possessing the characteristics of the word as the lexical-semantic unit: **ног-** / **нож-**;
- *the ending* is the dictionary element possessing the characteristics of the word forms: **-а** / **-ов** / **-ы** / **-ѣ**, etc.;
- *the pseudo-stem* is the element of the pseudo-dictionary possessing the characteristics of the word: **но-**;
- *the pseudo-ending* is the element of the pseudo-dictionary possessing the characteristics of the word forms: **-га** / **-гов** / **-гы** / **-жѣ**, etc.;
- *the type of inflection* is the dictionary element possessing the characteristics of the word as the lexical-grammar unit having a unique set of (pseudo)endings with their

[1] GDMRL was created on the basis of the electronic grammar dictionary of A. A. Zalizniyak and A. V. Sokirko.

morphological characteristics and fully describing the part of the word forms of the paradigm possessing the similar stem variant;

– *the stem variant* are the stems used for the formation of the word forms of the same word (**кoуп-** / **кoупль-**). The stem variants differ by the alternating components and have the directional relationship with the root stem;

– *the paradigm* is the set of word forms of the same word made of the stem (stem variants) and endings (endings variants);

– *the (sub)paradigm* is the paradigm which root stem has the directional relationship with another root stem and that can be used for the construction of the paradigm and having another stem as the main.

The dictionary elements have or may have the properties and characteristics corresponding to their linguistic analogs. For example, the properties and values of the noun stems are *the homonym number* and lexical-semantic and lexical-grammar characteristics, like, the *personal /impersonal, geographical, animate, collective* and some other, of the verb and adjective are only *the homonym number.* The properties of the noun endings are *the number* and *case*, of the adjective are *the gender, number* and *case*, of the verb are *the variability, time, number, person, gender* (for the forms of participles) and *case* (for the forms of participles).

The *inflection type* (IT) unit possesses the classifying grammar characteristics of the parts of speech. For example, the characteristic of the inflection types of the noun is *the gender*, of the adjective – *class*, of the verb – *the case*, *time* (for the participle as part of the verb), *voice* (for the participle). In addition, each inflection type possesses the characteristic of the part of speech.

The inflection types are one of the key objects of the base of dictionaries, as, possessing the classifying grammar signs and simultaneously having the relationship with the stem variants and the endings making part of the inflection type, they organize the paradigm. Each IT comprises *a set of endings* with their grammar signs and can be associated with the group of stems.

2.3. Model of data

The data meta model of each of the dictionaries making part of the system provides for storage of the information for the construction of the full paradigm of the word. The paradigm is built by concatenating the stem with the number of endings of the inflection type used with the stem.

The model is realized in the database of MANUSCRIPT as follows:

– the copies of the entities *the stem*, *ending*, *inflection type and dictionary* are stored in the table *the units*;

– the grammar signs and other necessary characteristics of the elements of the dictionaries are in the table *characteristics_units;*

– the main and auxiliary relationships between the dictionaries elements and between the dictionaries elements and the texts are in the tables *relationships, ends_relationships*.

To ensure the differentiation of the relationships, use is made of (1) the relationship of the main (root) variant of the paradigm stem with the subordinate variants (the sub-paradigm of

the subordinate pronominal forms of *добрыи* in the paradigm of the adjective *добръ*), (2) the relationship of the stem with the inflection type (the paradigm comprises the variant stems and their grammar characteristics and the inflection types associated with them), (3) the relationship of the inflection type with the endings (combines the endings with their grammar signs into the inflection type), (4) the relationship of the text units and the dictionary units (comprises 2 parts: determines the initial word form from the relationship with the stem and the set of grammar signs of the text form from the relationship with the ending (inside of the inflection variant)).

2.4. Program interfaces of work with the dictionary database

Processing of data in the grammar dictionaries is executed by several groups of procedures combined by the following functional peculiarities: the procedures ensuring (1) the creation, correction and elimination of the stems, endings, inflection types and their grammar signs; (2) the dynamic formation of queries over the databases of grammar dictionaries on the basis of the transmitted criterion; (3) the execution of retrieval of the dictionaries elements by the formed query; (4) filling the necessary structures with the data for their further processing and representation. All procedures are combined into a package and ensure the necessary transformations of the dictionary and text forms. The package provides the mechanisms of the simple expansion of the set of transformations. One of the sets of accessible transformations is presented on http://mns.udsu.ru/mns/slov.list_preobr.

3. Linguistic solutions

3.1. Lexical and morphological base

The analyzer basis is the database of grammar dictionaries which initial material are (1) the full-text electronic resources created in the Institute of the Russian Language after V. V. Vinogradov (work leaders A. M. Moldovan and A. A. Pichkhadze) and in Izhevsk Technical and Udmurtia Sate Universities (work leader V. A. Baranov) on the basis of the Old Russian translations and Old Russian copies of the divine service texts, (2) the description of the paradigms of the Old Russian in the form of the list of all forms and the initial indexation of all units of the index of words according to the indexes of paradigms (A. A. Pichkhadze).

3.2. Elimination of variance

One of the main linguistic problems arising during the creation of the grammar dictionary of the Old Russian is the considerable graphic-orthographic variance of the text word forms. Other problems immediately associated with that are (1) the necessity of ensuring the inexact search in the dictionary and (2) the necessity of unification of the linguistic units for the presentation of the lemmatization results and construction of the lists and indexes containing the units of the texts.

The module of morphological dictionaries has envisaged several levels of elimination of variance. The first level is the elimination of the graphic and orthographic variance with the use of the special rules of setting the functionally similar symbols and combination of symbols equal to one another. The second level is the use of the relationship of the text precedents with their normalized equivalents. The third level is the use of the relationship of the normalized units with their concordances in the grammar dictionary of the modern Russian.

3.2.1. Rules of unification (typification)

Due to the fact that the graphic and orthographic view of the word forms in the texts may differ from the samples in the database, the rules of unification (typification) of the text and dictionary units are necessary. By the typification we mean bringing the graphic-orthographic form of the linguistic units to their generalized variants that are invariants. The typification is performed by using the rules of transformation of the word forms (see below).

The elimination of the variance of the elements of the dictionaries can be done: (1) by the replacement of the main letter symbol by another (a) regardless of the dependence on the position, (b) depending on the preceding or following letter symbol and/or the position in the word form and also depending on the morpheme which part it is, (2) by the elimination of the symbols from the word form.

Let us name the positions where the rules of typification can be applied (the group of symbols may be set as the list of them or the list of their values): (1) independent, (2) dependent on the preceding symbol, the group of preceding symbols, the following symbol, the group of the following symbols, the preceding and following symbols, the groups of preceding and following symbols, (3) dependent on the position in the word form, (4) dependent on the application to the part of the word form – the stem, ending, (5) dependent on the application to the boundary of the parts of the word form – the stem and ending, the ending and postfix etc.

In connection with the above the rules now used in the system can be relatively divided into graphic, position and orthographic. Let us name some rules (the numbers indicate the sequence of the rules during the unification of the word forms).

Graphic rules

(1) The capital, small and superlinear letters are set equal to one another.

(3) The main and functional variants of the letter symbols are set equal to one another (**Д**–**д** etc.).

(4) The main variants of the letter symbols **о**–**ѡ**, **н**–**і**, **а**–**ѣ**, **ѕ**–**ѕ**, **ф**–**ѳ** and their variants are set equal to one another.

(9) The ligatures and digraphs and the corresponding combinations of letters **ѧ**-**аѵ**, **Ѣ**-**тв** etc. and also **ѿ**-**от**, **оѵ**–**оѵ** and other are set equal to one another.

Position rules

(2) The letters **ь** and **ъ** and their variants in the end of the word form are set equal to one another.

(7) The letters **а**-**ѣа**-**ѧ**-**ѣѧ**, **ю**-**оѵ**-**ѫ**-**ѣѫ** and **ѥ**-**є** and their variants in the position after the letters and combinations of letters **ш**, **щ**, **ж**, **жд**, **ѵ** and **ц** and their variants are set equal to one another.

(10) The letters **н** and **ь** and their variants in the position before the letters **н** and **і** and j-letters (yotized characters) and their variants and other are set equal to one another.

Orthographic rules

(5) The combinations "consonant + ѕ/ѧ + ь/ъ + consonant" and "consonant + ь/ъ + ѕ/ѧ + consonant" are set equal to one another.

(12) The combinations "consonant + ѕ/ѧ + ѣ + consonant" and "consonant + ѕ/ѧ + є + consonant" and other are set equal to one another.

Besides these transformations, to unify, it is necessary to perform some more changes in the composition and the order of the symbols in the word form, in particular, the spaces and diacritical marks are eliminated and the titles are unified and carried to one position.

The sequence of application of the rules for the creation of the unified variants of the word forms that should eliminate the excessive variance of the word forms and, in so doing, do not lead to the appearance of homographs, is important. For example, due to the fact that the inclusion in the list of the correspondence "(6) The letters ѣ–є and ѣ–н and their variants are set equal to one another" leads to the appearance of the unjustified amount of "noise", at present it is excluded from the rules.

3.2.2. Normalized equivalents

One of the important module components is the dictionary of normalized units. Exactly the normalized units are the samples to which the text word forms are compared after the application of the rules of variance elimination to them. The existence of this dictionary differentiates the module from other similar information systems.

By the normalization we understand bringing the graphic-orthographic form of the units to the view corresponding to the relative, traditionally applied graphic-orthographic rules of representation in writing of the units (morphemes) of the specific linguistic system. During normalization the existence of two variants of the same unit conditioned by the regular for this unit alternations and replacements of the letter symbols in its composition is possible. The presence of the variants presumes its ranking.

Due to the necessity of differentiation of the subtypes within their paradigms that differ by the graphics of their endings, the number of stem variants and the correlation of the alternating stem components, the secondary indexation of stems was done in the process of loading of the indexes of words in the dictionary database (V. A. Baranov).

Let us illustrate the above. For example, the paradigm of the nouns of the *o-stems of the hard variety* of the type **градъ** (м1а1) has five subtypes: three paradigms of the nouns with the stem on velars (м1а1_к, м1а1_г, м1а1_х), the paradigm of the nouns with other consonants in the stem end (м1а1) and the paradigm of the noun **христосъ** that differs from the preceding by the presence of two variants of the stem in the nominative case (м1а1*_хс). The paradigm of the verbs with **-нтн** has 35 subtypes that differ by the stem end and the presence / absence of the postfix: а1 – the verbs of the type of **молнтн** (the stems on **н**, **л**, **ѕ**), а1_ся – the verbs of the type of **молнтнсѧ** (the stems on **н**, **л**, **ѕ**), а1_си – the verbs of the type of **жалнтнсн** (the stems on **н**, **л**, **ѕ**), а1_ст – the verbs of the type of **поустнтн**, а1_ст_ся – the verbs of the type of **простнтнсѧ**, а1_ст_си – the verbs of the type of **пакостнтнсн**, а1_б – the verbs of the type of **любнтн** etc. At present the dictionary comprises 195 subtypes of verb paradigms. The similar split indexation enables describing unambiguously all lemma word forms.

The unambiguity (normalization) of the unit representation is used in the module for search of the dictionary units by their mask and display of the result both in the form satisfying the word form search condition and in the form of the paradigm.

3.2.3. Modern equivalents

The modern equivalents of the Old Russian lemmas are the GDMRL lemmas. The conformity is established as the relationship between the corresponding stems that can be the relationship of the type of "one-to-one" and "one-to-many". The latter relationship type is necessary if at present the Old Russian lemma is matched or may be matched by many words. For example, there are 3 stems *скотин-* with the grammar signs *fem. anim.*, *fem. inanim.*, *com.* in GDMRL. The Old Russian lemma has the relationship with all of them.

If there is no exact graphic-orthographic, grammar and/or word-formative correspondence for the Old Russian word in the modern language, the first is associated with the modern word with the similar meaning and morpheme composition but a different writing: **ж//жд** (**вожєннє** – **вождєннє** = вождение, **рожєннє** – **рождєннє** = рождение), **ѵ//щ** (**свѣща** = свеча) etc.

In some cases it is permissible to add the forms absent in the modern language to GDMRL. Let us give an example of the rules regulating such cases.

Pleophony – non-pleophony. There are (1) both (pleophonic and non-pleophonic) in the modern language: **сторона** – **страна**: **сторона** = 1. сторона, 2. страна; **страна** = 1. страна, 2. сторона; (2) only one variant: **болото** = 1. болото, 2. блато; **блато** = 1. блато, 2. болото; (3) there are neither the pleophonic nor non-pleophonic variants: we introduce both the variants in GDMRL and link each Old Russian word to each variant.

The similar assumptions allow considering the GDMRL units as the maximum unified relative to the Old Russian units and using this dictionary for search in the case when the user does not know the writing of the normalized form in the Old Russian language or its precedent in the text.

4. Web-modules

4.1. Module of editing of dictionaries

It is intended for input and correction of the elements of the grammar dictionaries and organization of the relationships of the elements inside the same dictionary and between the elements of different dictionaries (http://mns.udsu.ru/dic/dic2.main). The module uses the safety system of IRS Manuscript that ensures the access to the dictionaries only for the registered users possessing the necessary rights.

The main possibilities of the module: (1) input, editing and deleting of the dictionaries units and their values, (2) establishing, editing and deleting the relationships between the units of the dictionaries, (3) search of the stems on the basis of their formal and grammar characteristics, (4) loading the words with the simultaneous establishment of the relationship of the stem with the corresponding paradigm.

4.2. Module of lemmatization

The module of the morphological analyzer is intended for the analysis and synthesis of word forms, execution of search queries to the dictionaries of the database and representation of

query results. The module exists in three versions that differ by the functional possibilities is accessible from the portal "Manuscript" (http://manuscripts.ru/) on the address http://manuscripts.ru/mns/portal.main?p1=16&p_lid=1. The module is provided with the context user instructions.

4.2.1. First version

The main functions of this version are: (1) bringing the word form to the initial form (lemma), (2) getting data on the grammar (morphological) signs of the words and word forms, (3) building the word paradigm, (4) search of word forms by the mask (http://manuscripts.ru/mns/slov.prost_poisk?p_lang=EN).

Lemmatization can be realized only on the basis of the normalized view of the word form and search of the word forms satisfying the search conditions – only on the basis of the full coincidence of the input mask and the samples in the database. The result of the word form lemmatization is the lemma, list of homographs, their morphological values and the full paradigm of the lemma, the search result by the not full mask gives the list of word forms, their lemmas, grammar values and the possibility of viewing the paradigms.

4.2.2. Second version

This version demonstrates the possibility of lemmatization of the word forms that differ by the graphic-orthographic view from the normalized ones. The main distinctions from the first version: (1) bringing the graphic-orthographic variants of the word form to the lemma, (2) the possibility of selection of the typification rules and the number of the word form transformations, (3) the possibility of limitation of the search by the grammar signs of the word form(s) (http://manuscripts.ru/mns/slov.poisk?p_lang=EN).

The selection of the number of transformations gives the possibility of indicating how the input word form can differ from the normalized one: the higher is the number of transformations, the more variable may be the graphic view of the word form.

4.2.3. Third version

The main distinction from the previous versions is the possibility of lemmatization of the text fragment (http://manuscripts.ru/mns/slov.razbor_stroki?p_lang=EN). Due to the fact that the Old Slavonic text was not divided into the word forms, the lemmatizer has two modes: with the division of the fragment into the word forms and without.

The result of the fragment processing without spaces is the word division and morphological analysis of the fragment. All unreliable versions (the components that have no correspondences in the dictionary are determined) are eliminated. For example, during the analysis of the fragment from the Gospel **ндпрнзовнмужанрнндсѣмо** (*иди, призови мужа и приди сюда*) the user gets eight the most possible versions of word division, the remaining some tens of them – the less possible – are not shown to him. Besides, the analysis of the grammar signs of the concordant nouns and pronouns and their attributes allows keeping as the result only the forms that match each other by the morphological values: the lemmatization result **крѣпъкаагобогаслобамолю** (*крепкого Бога Слова молю*) contains only the forms of the genitive and genitive-accusative cases of the adjective and nouns.

4.2.4. Fourth version

The fourth version that is being built must provide for (1) the lemmatization of the text fragments, (2) the creation of the queries with the use of several search criteria, (3) the search of text precedents in the full-text database of IAS Manuscript, (4) the construction and display on the screen of the list of lemmas, dictionary entries and paradigms and also become the base for some other operations – (1) for establishing and editing of the relationships between the units of the dictionaries and the manuscripts and other units of the database of IAS Manuscript, (2) for the lemmatization of the text fragments from the external sources of data, (3) for storage of the results of lemmatization and search in the external files.

The essential distinction from the previous versions is the output of the text precedents, their addresses and contexts as the result and also the search over the corpus with the eliminated or not homonymy. The first must be executed on the basis of the algorithms of elimination of the variance and – at present – on the basis of the analysis of the presence of the coordination between the noun parts of speech (the primary lemmatization). The second, on the basis of the primary lemmatization and manually established relationships between the corresponding dictionary units and text precedents.

Hence, the automatic morphological analyzer of the Old Russian language is based on three interdependent and correlated components – (1) the Old Russian system of inflection realized by the use in the text in their graphic-orthographic variance, (2) on the developed theoretical approaches during its restructuring and (3) on the technological solutions that are put into the basis of the model and created program and tool means.

Acknowledgements

The work has been carried out with the support from the Russian Foundation for the Humanities (RGNF) (Project No. 05-04-12408в) and the Russian Foundation for Basic Research (RFFI) (Project No. 05-07-90217в).

References

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 (July 2004)*, pages 11.1-11.8.

Baranov V. A (2006). *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). In *Proceedings of the International Conference (June 2006),* Izhevsk, Russia.

Baranov V. A (2006). 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 (December 2006).* Electronic resource: http://drops.dagstuhl.de/portals/index.php?semnr=06491.

Baranov V. A., Mironov A. N., Lapin A. N., Mel'nikova I. S., Sokolova A. A. and Korepanova E. A. (2007). *Avtomaticheskij morfologicheskij analizator drevnerusskogo yazyka:* lingvisticheskie i texnologicheskie resheniya (*The Automatic Morphological Analyzer of Old Russian Language:* Linguistic and Technological Decisions). In *Electronic Proceedings of the 10th Jubilee International Conference "EVA 2007 Moscow" (December 2007).* Electronic resource: http://conf.cpic.ru/eva2007/rus/reports/report_1130.html.