

УДК 004.738.5

В. А. Баранов, доктор филологических наук, профессор  
Р. А. Верняева, кандидат филологических наук  
ИжГТУ имени М. Т. Калашникова

## КОРПУС М. В. ЛОМОНОСОВА В ИНТЕРНЕТЕ: НОВЫЕ ВОЗМОЖНОСТИ<sup>1</sup>

В статье представлен обзор новых возможностей корпуса текстов Ломоносова (*lomonosov.pro*). Усовершенствования коснулись аналитической и лингвистической разметки текстов, параметров запроса и форм демонстрации выборок. Особое внимание уделено результатам работ по созданию указателей личных и географических имен, словаря лингвистических терминов, третьей версии модулей *n-грамм* и статистики и некоторых других средств доступа к данным.

**Ключевые слова:** исторический корпус, полнотекстовая база данных, Ломоносов, лингвистическая разметка, аналитическая разметка, модули доступа, XML TEI.

### 1. История создания корпуса

В ноябре 2011 г. в Интернете по адресу *lomonosov.pro* был официально открыт корпус языка М. В. Ломоносова [1], содержащий тексты, включенные в 11 томов полного собрания сочинений [2].

Работы над будущим корпусом начались более 25 лет назад. Первые электронные транскрипции писем и поэтических произведений были подготовлены в конце 80-х – начале 90-х годов. Это позволило получить первый вариант конкорданса и начать работу над статьями словаря языка Ломоносова, который создавался сотрудниками кафедры русского языка Удмуртского государственного университета под руководством профессора В. М. Маркова. В последующие годы были созданы электронные машиночитаемые копии других томов, а в 2007 г. коллекция произведений М. В. Ломоносова была открыта на портале «Манускрипт: Славянское письменное наследие» (*manuscripts.ru*). Большую роль в работе над корпусом сыграло сотрудничество в 2007–2009 гг. с коллективом под руководством В. Д. Соловьева (Казанский государственный университет), которое дало возможность завершить подготовку транскрипций. К моменту переноса корпуса на отдельный домен тексты имели минимальную аннотацию, была разработана форма запроса с расширенными возможностями, осуществлена первичная автоматическая лемматизация, доработаны формы демонстрации данных.

В 2014–2016 гг. в рамках проекта, поддержанного Российским гуманитарным научным фондом (проект № 14–04–00585), продолжалась работа по расширению возможностей корпуса, в частности, осуществлено усовершенствование: 1) лингвистической и аналитической разметки; 2) программно-инструментальных средств обработки и поиска данных; 3) форм и способов визуализации данных корпуса.

### 2. Разметка

#### 2.1. Аналитическая разметка

Ранее в работах [3–5] были описаны функциональные возможности модулей и особенности метаинформации, обеспечивающие доступ к корпусу. Помимо стандартных – *автор, язык, время и место создания, жанр, тема, ключевые слова* и некоторые

др., – перечень характеристик текстов включал свойство *субъекты текста*, заполнение которого позволяло сформировать подкорпус на основе имен лиц, упомянутых в текстах. Но подобное вынесение имен собственных на уровень метаразметки не обеспечивало их нахождения в самом тексте и тем более не предполагало наличие дополнительной информации о людях. Другим решением является указание имен собственных в самих текстах и внесение в корпус общей информации о лицах.

В настоящее время корпус дополнен разметкой имен собственных двух категорий – личных и географических. Разметка осуществлена с помощью двух словарей – словаря персоналий и словаря локализаций. Единицы словаря содержат информацию о конкретном лице или географическом объекте, связаны с именами в текстах; значения единиц наследуются текстовыми единицами.

Единицы словаря персоналий могут содержать информацию о фамилии, имени, отчестве, прозвище, псевдониме, фамилии, имени, отчестве и других именах на родном языке, о именах в крещении, в монашестве, в лике святых, о социальном статусе, должности, роде занятий, родном языке, датах рождения и смерти, роли (автор, писец, редактор, субъект текста и др.), а также комментарий, помету *мифоним* для имен мифологических персонажей и др.

Единицы словаря локализаций включают информацию о типе объекта в тексте, наименовании в настоящее время, названии на языке-источнике, названии на языке-источнике в настоящее время, о типе объекта в настоящее время, роли (место создания, место изготовления, место события), а также комментарий и помету *мифоним*.

Каждая единица словаря отнесена к классу или разряду онима.

Особое значение имеет свойство *роль*, которое позволяет указать отношения *текст – лицо* (автор / субъект текста), *текст – географический объект* (место создания / место события) и использовать эту информацию для идентификации и текстов, и их содержания.

Визуализация может быть осуществлена с помощью запроса на этапе формирования подкорпуса,

выбора вида разметки или соответствующего указателя на этапе вывода текстовых данных.

Указатель личных имен представляет собой перечень, включающий имя лица на русском языке (имя на родном языке), даты рождения и смерти, сведения справочного характера, адрес упоминания лица в тексте корпуса, текстовая форма имени:

**Котельников, Семен Кириллович**, род. 1723, ум. 1806, профессор высшей математики, инспектор Академической гимназии. – Котельникова *10, 44*: 508–2, Котельникову *10, 44*: 508–7.

**Толстой, Петр Андреевич**, род. 1645, ум. 1729, государственный деятель и дипломат. – Pierre Tolstoj *6, 7*: 107–13.

**Рихман, Георг-Вильгельм (Richmann, Georg Wilhelm)**, род. 1711, ум. 1753, физик, соратник и друг М. В. Ломоносова. – Рихман *10, 30*: 484–15, 485–10, Рихмана *10, 30*: 484–5, 484–22, *10, 33*: 488–23, Рихманом *10, 29*: 482–15, 482–28.

Статья указателя географических наименований включает название географического объекта (название на языке-источнике), класс топонима, тип объекта, сведения справочного характера, адрес и наименования в тексте:

**Сибирский океан**, *гидр.*, океан [в наст. время – Северный Ледовитый океан]. – Сибирским океаном *10, 88*: 570–2.

**Воскресенский монастырь**, *хорон.*, монастырь (монастырь, основанный патриархом Никоном) [в наст. время – Воскресенский Ново-Иерусалимский монастырь]. – Resurrection *6, 7*: 130–13.

**Ост-Индия**, *хорон.*, географический регион (устаревшее название Индии и некоторых стран Южной и Юго-Восточной Азии). – Ost-Indien *10, 88*: 570–3.

Таким образом, если ранее корпус содержал лишь метаинформацию в виде имен лиц, упомянутых в текстах, мест, где созданы тексты, то сейчас в нем имеется справочная информация о лицах и географических объектах с привязкой к контексту.

## 2.2. Грамматический словарь языка М. В. Ломоносова

Лемматизация текстов осуществляется с помощью морфологического анализатора, основой которого является электронный грамматический словарь языка М. В. Ломоносова. Последний представляет собой базу данных (псевдо) основ и (псевдо) окончаний и создан на основе данных электронного грамматического словаря современного русского языка Зализняка – Сокирко [6, 7].

Интерфейс словаря позволяет редактировать, добавлять и удалять единицы и их значения, устанавливать и удалять связи единиц [8].

### 2.2.1. Редактирование грамматического словаря языка Ломоносова

Первичная автоматическая лемматизация на основе электронного словаря СРЛЯ показала большое количество нелеммутизируемых словоформ. Их анализ позволил установить, что причинами являются: «1) отсутствие в словаре необходимой леммы, а соответственно, и ее словоформ; 2) отсутствие у имеющейся в словаре леммы искомой словоформы;

3) иной, чем в словаре, графический и/или орфографический состав текстовой словоформы» [9].

В базу данных грамматического словаря добавлены и добавляются (псевдо) основы слов, отсутствующих в СРЛЯ, а в существующие парадигмы – необходимые (псевдо) окончания (сущ.  $\emptyset$ /-ов/-ев, -и/-ы/-а/-я и др., прил. -аго/-яго, -ья/-ия и др., глаг. -ть/-ти) [10]. Лемматизация с использованием парадигм XVIII в. позволяет увеличить количество леммутизированных словоформ в тексте до 85–92 %.

### 2.2.2. Доработка лемматизатора

Большая часть нелеммутизированных форм связана с использованием в транскрипции букв, которые отсутствуют в современном алфавите.

В пятую версию морфологического анализатора добавлены правила унификации графем и их сочетаний как без учета позиции, так и с ее учетом ( $io = jo = j^o = i^o = \ddot{e}$ ;  $i = \text{и}$ ;  $\text{ѣ} = \text{е}$  и др.) (о морфологическом анализаторе см. работы [11–13]), что существенно снижает процент нелеммутизированных форм в текстах, в которых такие написания часты.

### 2.2.3. Электронный словарь лингвистической терминологии

Одной из задач проекта является создание терминологического словаря. Ее решение потребовало анализа терминов и терминологических сочетаний в области филологии и лингвистики. Описаны основные характеристики терминов и терминологических сочетаний, среди них: 1) отношение к той или иной тематической области; 2) способы толкования; 3) иерархические и синонимическое связи, определяющие целостность метаязыка [14, 15].

В ходе работы выявлен состав терминов, их парадигмы; осуществлена типология, определены параметры словаря; в базе данных созданы необходимые поля, обеспечивающие ввод характеристик терминологических единиц.

Конечным результатом исследования стало создание электронного словаря лингвистических (филологических) терминов М. В. Ломоносова, структура которого включает следующие параметры: терминологическая область и подобласть, толкование, синонимическое определение, развернутое описание, этимологическое объяснение и некоторые другие. Модель обеспечивает хранение одно- и многокомпонентных терминов, терминов-омонимов, возможность дифференцированного ввода определений Ломоносова (авторское), словарных определений (внешнее), определений авторов словаря, грамматическую часть и адреса текстовых прецедентов. Например:

СИЛЛОГИЗМ (29) – сущ. муж (парадигма 33979008)

Терминологическая область: филология

Терминологическая подобласть: ритороведение  
Значение

словарное: Вид дедуктивного доказательства.

авторское: Доводы состоят иногда из одного, иногда из многих силлогизмов, между собою сопряженных. Силлогизм составляется из трех рассуждений, из которых две первые называются посылками,

а третье – следствием, которое из посылок выводят (ЛМН, ПСС, т. 7, стр. 154).

Развернутое описание

авторское: Доводы предлагаются: 1) силлогизмом; 2) энтимемой; 3) дилеммой <...> (ЛМН, ПСС, т. 7, стр. 68). пример условного силлогизма <...>. Пример разделительного силлогизма <...> (ЛМН, ПСС, т. 7, стр. 156). Силлогизм составляют два предыдущих предложения и следствие (ЛМН, ПСС, т. 7, стр. 68).

Этимологическое объяснение

словарное: Греч.; лат. *sylogismos*. Умозаключение, в котором из двух данных суждений (посылок) получается третье (вывод) (Большой толковый словарь русского языка / гл. ред. С. А. Кузнецов. СПб.: Норинт, 2004. С. 1185).

& *силлогизм* (5) *ед. муж. вин.* | *ед. муж. им.* ЛМН. ПСС, 7, 2 68.1.1–5.14; ЛМН. ПСС, 7, 4 (4) 154.1.1–20.34, 157.1.1–8.19, 312.1.1–3.27, 319.1.1–3.53 & *силлогизма* (7) *ед. муж. род.* ЛМН. ПСС, 7, 2 68.1.1–9.23; ЛМН. ПСС, 7, 4 (6) 155.1.1–17.36, 156.1.1–2.57, 156.1.1–21.21, 313.1.1–21.25, 328.1.1–2.35, 331.1.1–5.26 & *силлогизмах* *мн. муж. предложн.* (2) ЛМН. ПСС, 7, 4 (2) 319.1.1–2.16, 326.1.1–4.22 & *силлогизме* *ед. муж. предложн.* (3) ЛМН. ПСС, 7, 4 (3) 155.1.1–19.50, 313.1.1–26.16, 328.1.1–5.16 & *силлогизмов* *мн. муж. род.* (2) ЛМН. ПСС, 7, 4 (2) 154.1.1–19.54, 157.1.1–7.46 & *силлогизмом* *ед. муж. твор.* (2) ЛМН. ПСС, 7, 2 68.1.1–4.28; ЛМН. ПСС, 7, 4 197.1.1–25.57 & *силлогизму* *ед. муж. дат.* (4) ЛМН. ПСС, 7, 4 (4) 311.1.1–13.29, 311.1.1–15.8, 311.1.1–17.10, 326.1.1–13.37 & *силлогизмы* (4) *мн. муж. вин.* | *мн. муж. им.* ЛМН. ПСС, 7, 4 (4) 156.1.1–12.29, 156.1.1–15.22, 156.1.1–18.10, 157.1.1–4.24.

3. Инструментарий

3.1. Формы запроса и вывода

Сегодня запросная форма представляет собой инструмент, позволяющий осуществить выборку данных с помощью точной и неточной маски, применения образца в поиске к словоформе или лемме, поиска на основе регулярных выражений, на основе грамматических признаков, на основе аналитической разметки и др.

На предыдущих этапах работы с текстами М. В. Ломоносова пользователю была предоставлена возможность просмотра результатов запросов в виде текстов (оригинальный / преобразованный), указателей словоформ (прямой, инверсированный, количественный), указателей начальных форм (прямой, инверсированный, количественный, полный грамматический указатель) [16, 17].

В настоящее время формы вывода корпуса доработаны и включают также конкордансы (прямой, инверсированный), дополнительные указатели (указатели нелематизированных словоформ прямой и инверсированный), указатели собственных имен (географических наименований, личных имен), словарь терминов. Наличие разметки имен собственных позволяет визуализировать их в тексте, а при наведении курсора – показать сведения о них.

3.2. Модуль *n*-грамм

В работе [18] была представлена первая версия специализированного модуля, который позволяет получить сортированные по частоте встречаемости перечни сочетаний из нескольких компонентов-словоформ, а в [19, 20] продемонстрированы возможности третьей версии этого модуля ([http://manuscripts.ru/mns/cred\\_ngr.stat?p\\_collect=50584966](http://manuscripts.ru/mns/cred_ngr.stat?p_collect=50584966)), которая дает возможность на основе формальных, лингвистических, количественных и статистических параметров осуществить анализ сочетаний с точки зрения степени их грамматической и/или семантической устойчивости.

Особенностью модуля является предоставление самому пользователю возможности сформировать запрос на основе характеристик компонентов, их соотношения. Параметры предусматривают: а) указание количества компонентов; б) расстояния между ними; в) использование образца для каждого (в том числе на основе регулярных выражений); г) наложение морфологического фильтра; д) выбор вида сортировки (в том числе на основе мер ассоциации) и др. Параметры запроса позволяют задать и другие характеристики сочетания, например нефиксированное положение компонентов относительно друг друга, включение в выборку только симметричных сочетаний (компоненты встречаются только вместе), игнорирование служебных слов, учет границ предложений и др. При выводе сочетаний и их количественных и статистических характеристик приводятся адреса-гиперссылки в подкорпусе, что позволяет просмотреть контекст.

3.3. Модуль статистики

Модуль статистики представляет собой инструмент, предназначенный для получения количественных и статистических данных о лингвистических объектах полнотекстовой базы данных.

В работе [21] описаны функциональные возможности первой версии модуля и принципы и особенности демонстрации данных. В 2015 г. была представлена вторая версия модуля, в 2016 г. итоговая – третья ([http://manuscripts.ru/mns/cred.stat?p\\_collect=50584966](http://manuscripts.ru/mns/cred.stat?p_collect=50584966)), которую отличает от предыдущих возможность, например, указать текст как единицу шага, время создания – как параметр дискретности. Модуль позволяет использовать в качестве запроса и другие метаданные текстов, в частности, можно указать в качестве группировки жанр и вывести на экран диаграмму частоты встречаемости искомой единицы в текстах, различающихся значением этого параметра.

3.4. Формат XML TEI

Известно, что формат кодирования текстов TEI направлен на обеспечение обмена информацией, хранимой в электронной форме ([tei-c.org](http://tei-c.org)).

Одной из задач при создании электронного корпуса являлось обеспечение конвертирования текстов М. В. Ломоносова из базы данных информационно-аналитической системы «Манускрипт» в формат XML TEI, а для этого – создание англоязычного варианта предусмотренных справочниками мета-, аналитических и лингвистических значений.

Для реализации поставленной задачи были: 1) выделены композиционно-структурные особенности научной и эпистолярной прозы М. В. Ломоносова и драматических произведений; 2) доработана модель базы данных «Манускрипт» с учетом параметров, предусмотренных в формате XML TEI; 3) доработан конвертор из БД «Манускрипт» в формат XML TEI с целью выгрузки фрагментов.

Для представления корпуса в формате TEI установлены соответствия свойств и значений для единиц Текст и Рукопись; найдены соответствующие значения русских и английских параметров текстов; фрагментов и лингвистических единиц; для удобной настройки выгрузки в формат TEI добавлены таблицы групп и характеристик свойств единиц, что в свою очередь позволяет настраивать структуру выгрузки без изменений исходного кода процедур; доработана форма справочника свойств единиц; корпус выгружен в формат XML TEI, что позволяет использовать его на других технологических платформах.

Таким образом, корпус языка М. В. Ломоносова сегодня – это полнотекстовая база данных, имеющая аналитическую и лингвистическую разметку и разнообразные формы поиска и демонстрации данных, позволяющие знакомиться с наследием великого русского поэта, ученого, просветителя и осуществлять широкий круг научных исследований в области истории науки и культуры, литературоведения и лингвистики.

#### Библиографические ссылки

1. Корпус М. В. Ломоносова [Электронный ресурс] / Ижевский государственный технический университет, кафедра лингвистики, Центр теоретической и прикладной лингвистики, 2005–2016; Казанский (Приволжский) федеральный университет, 2007–2009; Удмуртский государственный университет, лаборатория по автоматизации филологических работ, 1989–2013; рук. проекта В. М. Марков, 1989–2010; срук. и рук. В. А. Баранов, 1989–2016. – URL: [lomonosov.pro](http://lomonosov.pro) (дата обращения: 18.09.2016).
2. Ломоносов М. В. Полное собрание сочинений : в 10 т. Т. 11 доп. / М. В. Ломоносов. – М. ; Л. : Изд-во АН СССР, 1950–1983. – Т. 1–11.
3. Баранов В. А., Аникина Р. А., Кокорина Т. В., Ощепков С. В., Соколова А. А. Метаинформация в коллекции М. В. Ломоносова на портале «Манускрипт: Славянское письменное наследие» // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам: материалы Междунар. науч. конф. (Казань, 26–30 августа 2008 г.) / отв. ред. В. А. Баранов, В. Д. Соловьев. – Казань : Изд-во КГУ, 2008. – С. 23–27.
4. Баранов В. А. Полное собрание сочинений М. В. Ломоносова в интернете: подготовка электронной коллекции и функциональные возможности модулей корпуса // Уч. зап. Казанского ун-та. Серия: Гуманитарные науки. – Т. 152. Вып. 6. – 2010. – С. 223–234.
5. Баранов В. А. Корпус языка М. В. Ломоносова // Русский язык: функционирование и развитие (к 85-летию со дня рождения заслуженного деятеля науки Российской Федерации профессора Виталия Михайловича Маркова) : материалы Международной научной конференции (Казань, 18–21 апреля 2012 г.) / Казан. ун-т; Ин-т филологии и искусств; Каф. ист. рус. яз. и слав. языкозн.; под общ. ред. Л. П. Абдулхаковой, Д. Р. Копосова. – Казань : Казан. ун-т, 2012. – Т. 1. – С. 58–63.
6. Зализняк А. А. Грамматический словарь русского языка: Словоизменение. – 3-е изд., стер. – М. : Рус. яз., 1987. – 880 с.
7. Сокирко А. В. Русский морфологический словарь [Электронный ресурс] // Автоматическая обработка текста. – URL: <http://aot.ru/docs/rusmorph.html> (дата обращения: 25.12.2015).
8. Баранов В. А., Гулина О. В., Миронов А. Н. Морфологическая парадигма и ее составляющие в системе «Манускрипт» // Информационные технологии и письменное наследие: материалы междунар. науч. конф. (Уфа, 28–31 октября 2010 г.) / отв. ред. В. А. Баранов. – Уфа ; Ижевск : Вагант, 2010. – С. 28–31.
9. Баранов В. А. Нестандартная морфология, орфография, графика текстов М. В. Ломоносова: к проблеме автоматической лемматизации исторического корпуса // Письменное наследие и современные информационные технологии: материалы Междунар. науч. конф. (Вильнюс, 22–28 августа 2016 г.) / отв. ред. В. А. Баранов, Т. Тимченко (в печати).
10. Там же.
11. Автоматический морфологический анализатор древнерусского языка: лингвистические и технологические решения [Электронная публикация] / В. А. Баранов, А. Н. Миронов, А. Н. Лапин, И. С. Мельникова, А. А. Соколова, Е. А. Корепанова // 10-я юбилейная международная конференция «EVA 2007 Москва». – М., 2007. – URL: [http://conf.evarussia.ru/eva2007/rus/reports/report\\_1130.html](http://conf.evarussia.ru/eva2007/rus/reports/report_1130.html).
12. Баранов В. А. Полное собрание сочинений М. В. Ломоносова в Интернете: подготовка электронной коллекции и функциональные возможности модулей корпуса // Уч. зап. Казанского ун-та. Серия: Гуманитарные науки. – Т. 152. Вып. 6. – 2010. – С. 223–234.
13. Баранов В. А. Нестандартная морфология, орфография, графика текстов М. В. Ломоносова: к проблеме автоматической лемматизации исторического корпуса // Письменное наследие и современные информационные технологии: материалы Междунар. науч. конф. (Вильнюс, 22–28 августа 2016 г.) / отв. ред. В. А. Баранов, Т. Тимченко (в печати).
14. Зливко С. Д. О создании электронного словаря лингвистических терминов (на материале работ М. В. Ломоносова) // В мире научных открытий. – 2014. – № 11. – С. 993–1000.
15. Баранов В. А., Гнутиков Р. М., Зливко С. Д. Авторский электронный словарь-справочник лингвистической терминологии М. В. Ломоносова // Интеллектуальные системы в производстве. – 2015. – № 3 (27). – С. 88–92.
16. Баранов В. А. Полное собрание сочинений М. В. Ломоносова в интернете: подготовка электронной коллекции и функциональные возможности модулей корпуса // Уч. зап. Казанского ун-та. Серия: Гуманитарные науки. – 2010. – Т. 152. Вып. 6. – С. 223–234.
17. Баранов В. А. Корпус языка М. В. Ломоносова // Русский язык: функционирование и развитие (к 85-летию со дня рождения заслуженного деятеля науки Российской Федерации профессора Виталия Михайловича Маркова) : материалы Международной научной конференции (Казань, 18–21 апреля 2012 г.) / Казан. ун-т; Ин-т филологии и искусств; каф. ист. рус. яз. и слав. языкозн. ; под общ. ред. Л. П. Абдулхаковой, Д. Р. Копосова. – Казань : Казан. ун-т, 2012. – Т. 1. – С. 58–63.
18. Баранов В. А. Организация поиска и демонстрации коллокаций в корпусе «Манускрипт» // Проблемы истории, филологии, культуры. – 2014. – № 3 (45). – С. 275–277.

19. Баранов В. А. Модуль  $n$ -грамм исторического корпуса «Манускрипт»: структурные и лингвистические параметры // Научное наследие В. А. Богородицкого и современный вектор исследований Казанской лингвистической школы: тр. и матер. междунар. конф. (Казань, 31 окт. – 3 нояб. 2016 г.): в 2 т. / под общ. ред. К. Р. Галиуллина, Е. А. Горобец, Г. А. Николаева. – Казань : Изд-во Казан. ун-та, 2016. – Т. 1. – С. 50–61.

20. Баранов В. А. Опыт создания модуля  $n$ -грамм системы «Манускрипт» и оценки эффективности его исполь-

зования для поиска коллокаций в корпусе М. В. Ломоносова // Интеллектуальные системы в производстве. – 2016. (в печати).

21. Баранов В. А., Дубовцев С. В. Модуль статистики информационно-аналитической системы «Манускрипт»: функции и демонстрация данных // Информационные технологии и письменное наследие : материалы IV Междунар. науч. конф. (Петрозаводск, 3–8 сентября 2012 г.) / отв. ред. В. А. Баранов, А. Г. Варфоломеев. – Петрозаводск ; Ижевск, 2012. – С. 23–26.

\*\*\*

V. A. Baranov, Doctor of Philology, Professor, Kalashnikov ISTU  
R. A. Vernyaeva, PhD in Philology, Kalashnikov ISTU

#### **Corpus of M. V. Lomonosov in the Internet: New Opportunities**

*The article provides an overview of the new features of the Lomonosov corpus (lomonosov.pro). Improvements touched the analytical and linguistic markup of texts, request parameters and forms of demonstration samples. Particular attention is paid to the results of creating index of personal names and geographical names, the dictionary of linguistic terms, the development of the third version of modules for  $n$ -grams and statistics and some other means of access to the data.*

**Keywords:** historical corpus, full-text database, Lomonosov, linguistic marking, analytical marking, access modules, XML TEI.

Получено: 18.11.16