УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО УНИВЕРСИТЕТА
Том 152, кн. 6 Гуманитарные науки *2010*

ТЕКСТ

УДК 801.82(04)

ПОЛНОЕ СОБРАНИЕ СОЧИНЕНИЙ М.В. ЛОМОНОСОВА В ИНТЕРНЕТЕ: ПОДГОТОВКА ЭЛЕКТРОННОЙ КОЛЛЕКЦИИ И ФУНКЦИОНАЛЬНЫЕ ВОЗМОЖНОСТИ МОДУЛЕЙ КОРПУСА

В.А. Баранов

Аннотация

В статье рассматриваются теоретические и прикладные вопросы разработки, создания и использования в лингвистических исторических исследованиях электронного корпуса текстов М.В. Ломоносова. Основное внимание уделено средствам и приемам автоматизированной обработки и анализа данных с помощью специализированных модулей – лемматизации и выборке данных.

Ключевые слова: корпусная лингвистика, компьютерная обработка лингвистической информации, полнотекстовые базы данных, лингвистические интернет-ресурсы, язык Ломоносова.

Анализ ситуации в области подготовки с помощью информационных технологий текстового материала для лингвистических исследований показывает, что на смену технически достаточно простым формам демонстрации документов в виде сканированных копий или гипертекста приходят машиночитаемые транскрипции со структурной и лингвистической разметкой, а также с метаописанием и аналитическим описанием документа и его частей.

Большая работа в этом направлении осуществляется в рамках проектов по созданию лингвистических ресурсов на основе современного русского языка, в частности при подготовке исследовательских корпусов. Наиболее значимые результаты хорошо известны: это Национальный корпус русского языка, корпусы Лаборатории общей и компьютерной лексикологии и лексикографии Московского государственного университета, Хельсинский аннотированный корпус, Уппсальский и Тюбингенский корпусы.

Достаточно давно в России ведутся работы и по созданию машиночитаемых ресурсов на основе текстов донационального периода. Например, в рамках подготовки Машинного фонда русского языка были запланированы и начаты работы по накоплению электронных копий произведений XVIII века, в настоящее время произведения XVIII века включены в Национальный корпус русского языка [1].

Тексты более раннего периода также уже довольно широко распространены в Интернете в машиночитаемой форме. Они представляют собой как переводы

или упрощенные транскрипции исторических источников, так и документы, которые передаются в близком к оригиналам виде. Последние составляют основу, например, электронной публикации древнерусских берестяных грамот, средневековых коллекций проекта TITUS, Хельсинского корпуса старославянских памятников, нескольких электронных коллекций русских летописей и некоторых других интернет-коллекций.

Как известно, ценность корпуса заключается в его направленности на фиксацию и демонстрацию языковых единиц в таком объеме документов, который обеспечивает репрезентацию достаточного их количества для исследования текстовых и языковых свойств, в первую очередь – вариативности. В связи с тем что при корпусных исследованиях для анализа лингвистических явлений требуется определенный уровень частотности единиц, корпусы современных текстов создаются объемом в миллион и более словоупотреблений. Понятно, что для современных корпусов практически отсутствуют ограничения в текстовом материале, так как количество документов, которое может быть привлечено для создания такого корпуса, практически безгранично.

Создание достаточного для лингвистических исследований по объему, полного и сбалансированного современного корпуса – задача во многом техническая и технологическая. Не случайно, одним из видов корпуса является интернет-корпус, материалом для которого служат документы, хранящиеся в Интернете.

Используемая при анализе корпуса разметка позволяет выделить подкорпусы на основе различных текстовых параметров, например, на основе жанровых, функционально-стилевых, тематических, содержательных и иных характеристик. Чрезвычайно важным при исследовании существенных речевых и языковых характеристик является и параметр времени. Понятно, что в рамках современного корпуса временная дифференциация документов сильно ограничена. Это ограничение устраняется в диахронических корпусах. В них временной параметр является одним из важнейших и, как следствие, позволяет на основе наблюдений над изменениями соотношения языковых вариантов или над сменой одной единицы другой в подкорпусах, различающихся временем создания документов, выявить общие и частные характеристики некоторого речевого и/или языкового явления в их изменении.

В то же время есть объективные причины, по которым создание сопоставимых с современными по объему, полноте и сбалансированности диахронических корпусов, демонстрирующих документы продолжительного временного диапазона, вряд ли возможно (мы различаем исторический корпус, в котором представлены документы, созданные в некотором временном диапазоне, предшествующем XIX веку, и диахронический корпус, в котором наряду с современными представлены и тексты донационального периода). Очевидно, что объем и полнота исторического корпуса ограничены количеством сохранившихся документов того или иного времени, этим же ограничены возможности достижения сбалансированности корпуса.

Есть еще несколько причин, которые в настоящее время существенно тормозят создание и использование диахронических корпусов: очень незначительное количество электронных машиночитаемых копий документов предшествующих периодов, большие трудозатраты на их подготовку, отсутствие исторических

корпусов, сопоставимых с современными по важнейшим для корпуса параметрам, в частности по объему.

Особым видом корпуса является авторский корпус, основные характеристики которого также определяются объемом дошедших до настоящего времени произведений автора, их тематикой и жанрово-стилистическими характеристиками.

Несмотря на то что при создании исторического и авторского корпусов часто невозможно достичь такого количества единиц (например, слов или синтаксических конструкций), которое необходимо для получения статистически значимых величин, представленность в таких корпусах всех произведений конкретного автора или всех текстов определенного времени позволяет считать их не менее важными для исследования языка, чем современные корпусы большого объема: авторский корпус репрезентирует подъязык автора, являющийся частью языка конкретной эпохи; исторический корпус в определенном временном диапазоне дает факты для описания языковой системы, сопоставимой с системами предшествующего и последующего периодов.

Таким образом, учет свойств вхождения и сопоставимости позволяет снять ограничения, связанные с недостаточной представленностью в текстах языковых явлений, и даже единичные, статистически непоказательные случаи вариативности рассматривать как достаточно надежные на более широком фоне языка определенного времени или конкретного автора.

Существенные отличия авторского и исторического корпусов от современных, определяющие использование особых методов анализа материала, требуют применения не только уже утвердившихся при исследовании корпусов приемов демонстрации данных, но и специализированного, направленного на анализ исторических и авторских речевых и языковых явлений инструментария, обеспечивающего визуализацию данных в соответствии с целями и задачами работы исследователя.

В настоящее время в нескольких организациях России активно ведутся работы по подготовке электронных словарей, конкордансов, электронных коллекций произведений М.В. Ломоносова. Изучение научного и художественного наследия Ломоносова, начавшееся в XIX веке, с разной степенью активности продолжалось на протяжении всего XX века. В середине прошлого столетия было издано академическое полное собрание сочинений в 11 томах (ПСС), которое позволяет знакомиться с наследием Ломоносова и изучать его. В то же время следует констатировать, что некоторые задачи, поставленные еще Императорской Академией наук, не выполнены до сих пор, хотя попытки неоднократно предпринимались. К таким задачам относится и создание словаря языка Ломоносова.

В 60-е годы XX века в Казанском государственном университете под руководством профессора В.М. Маркова начались работы по созданию словаря языка М.В. Ломоносова: была подготовлена картотека Полного собрания сочинений, выполнены работы по исследованию лексики, грамматики, стиля автора. Полученные результаты, а также вопросы, связанные с изучением языка XVIII века, обсуждались на нескольких конференциях серии «Ломоносовские чтения», прошедших в Казанском государственном университете в 60–70-х годах, и изложены в большом количестве статей, опубликованных их участниками [2–6].

В первой половине 90-х годов попытка продолжить начатое в Казани была предпринята в Удмуртском государственном университете. Именно тогда на кафедре русского языка, возглавляемой В.М. Марковым, были сформулированы принципы словаря [7] и создан конкорданс к письмам и поэтическим произведениям. Подготовленные в то время машиночитаемые копии нескольких томов ПСС явились первым шагом к созданию полной электронной коллекции текстов Ломоносова.

В 2007 году работы по созданию этой коллекции были возобновлены в Ижевске в рамках проекта «Большой корпус русского языка XVIII в.» (рук. В.Д. Соловьев (Казанский государственный университет), группа в Ижевске – В.А. Баранов, Р.А. Аникина, А.А. Вотинцев, Р.М. Гнутиков, Т.В. Кокорина, В.А. Романенко, И.С. Соломенников).

Технологической основой для создания коллекции и инструментов работы с ней стала информационно-аналитическая система «Манускрипт» (далее – Система), созданная и развиваемая усилиями лингвистов и программистов Ижевского государственного технического и Удмуртского государственного университетов (адрес портала «Манускрипт: славянское письменное наследие»: http://manuscripts.ru/) [8–10]. Универсальная модель полнотекстовой базы данных Системы, интернет-сервисы, предназначенные для работы со средневековыми славянскими рукописями, позволили поставить и решить задачу создания корпуса языка Ломоносова.

Целями работы явились демонстрация в Интернете письменного наследия М.В. Ломоносова, предоставление пользователям специализированных веб-сервисов работы с корпусом – форм для создания запросов, интерфейсов для визуализации выборок и упорядочения их единиц, средств для просмотра контекстов и др.

В ходе выполнения проекта было решено большое количество лингвистических и технологических задач, которые могут быть разделены на несколько групп: создание полнотекстовой базы данных машиночитаемой коллекции произведений М.В. Ломоносова, разработка и создание инструментов для ввода, распределенного редактирования, аннотирования и лемматизации корпуса текстов XVIII века, организация и обеспечение доступа к коллекции через Интернет, создание веб-сервисов первичного анализа текстового материала корпуса.

Материалом для электронной коллекции (URL: http://manuscripts.ru/mns/portal.main?p1=31) стало Полное собрание сочинений М.В. Ломоносова в 11 томах.

На первом этапе подготовки коллекции была проведена большая работа по переводу ПСС в машиночитаемую форму, по сверке набора с печатным изданием. Итогом этого этапа явилось создание электронной транскрипции ПСС, в которой с помощью разметки сохранены все структурно-содержательные особенности печатного издания: заголовки, абзацы, выделения, сноски, таблицы, ссылки на нетекстовые элементы и т. п. Размеченный текст был подготовлен для конвертирования в формат mnsXML, используемый для загрузки в базу данных «Манускрипт» транскрипций, созданных с помощью внешних программно-инструментальных средств.

Загрузка mnsXML-документа в базу данных Системы была выполнена при помощи универсальной программы-конвертора, представляющей собой веб-при-

ложение, пользовательский интерфейс которого реализован с использованием Oracle Applications Express 3.0.1, а логика конвертации – на основе хранимых процедур на PL/SQL. Конвертор использует ту же систему определения прав доступа пользователей к текстам, что и Система (см. подробнее [11, 12]).

С целью включения текстов в состав Национального корпуса русского языка подготовленная ижевской группой проекта машиночитаемая копия ПСС была размечена по правилам корпуса членами казанской группы участников проекта.

Одновременно с загрузкой текстов в базу данных разрабатывалась модель для мета- и аналитической разметки документов, велись работы по созданию веб-страниц запросов и вывода результатов на портале «Манускрипт: славянское письменное наследие», а также работы по созданию морфологического анализатора для лемматизации корпуса XVIII века.

Мета- и аналитическая информация о документах корпуса представляет собой набор свойств и их значений, которые обеспечивают описание документа (единицы печатного издания или произведения) и его фрагментов с точки зрения их основных характеристик. Минимальной единицей для разметки является знак, но фактически в качестве минимального объекта устанавливается фрагмент текста или тома ПСС. Набор характеристик фрагмента определяется по типу фрагмента в соответствии с его структурно-функциональными свойствами.

На основе анализа произведений и их фрагментов была разработана модель данных документов корпуса, выявлены необходимые и достаточные для поиска и отбора материала мета- и аналитические параметры и значения.

Требования к веб-формам запроса и представления выборки были сформулированы следующим образом:

– выбор атрибутов поиска с применением логических условий;

– выбор формальных и грамматических характеристик лингвистических единиц;

– представление указателей на основе параметров запроса;

– представление указателей на основе нескольких произведений;

– представление прямых, реверсивных, количественных указателей;

– переход от указателей к контекстам.

Веб-интерфейс поиска текстов и их фрагментов в корпусе по мета- и аналитическим описаниям разработан и создан в двух режимах – простом и расширенном. В первом режиме использован механизм уточнения поиска, дающий возможность детализировать выборку до нужного уровня (URL: http://manuscripts.ru/mns/srch.simple?p_ed_id=50584966). Во втором режиме создание фильтра возможно одновременно на основе нескольких основных объектов коллекции – издания, произведения и фрагментов текстов, объединенных логическими условиями, и на основе параметров объектов (URL: http://manuscripts.ru/mns/srch.complex?p_lang=RU&p_ed_id=50584966).

Полученный результат нужного уровня детализации может быть передан на следующий шаг, который дает возможность определить дополнительные условия выборки для ее уточнения и визуализировать результат поиска.

Веб-форма, предназначенная для визуализации результатов выборки документов и их фрагментов, позволяет выбрать форму и параметры представления данных – тексты или указатели, а также их объем, состав и форму представления

с помощью указания диапазона страниц издания, маски искомых словоформ и формы выдачи результата.

В связи с необходимостью предоставить пользователю возможность отобрать нужный для исследования лингвистический материал и получить его в удобном для анализа виде, предусмотрено два вида визуализации выборки: визуализация текстов и их фрагментов и визуализация перечней лингвистических единиц (словоформ и начальных форм), второй вид может быть представлен в двух формах – табличной и списочной. Табличные формы являются основными для инверсированных и количественных указателей. Отбор материала может быть также осуществлен указанием маски искомой словоформы или слова. Переход к контексту осуществляется по адресу словоформы.

Система позволяет лемматизировать корпус, а веб-интерфейсы – получить данные на основе морфологической информации о слове и словоформе. Для лемматизации корпуса был создан автоматизированный лемматизатор на базе грамматического словаря современного русского языка Системы (URL: http://manuscripts.ru/apex/f?p=104:1, требуется авторизация).

Лемматизатор имеет веб-интерфейс, позволяющий в территориально распределенном режиме лемматизировать тексты Ломоносова, осуществлять контроль за ходом лемматизации и редактировать ее результаты. Результаты лемматизации сохраняются в базе данных грамматического словаря языка Ломоносова.

Основные требования к модулю лемматизации были сформулированы следующим образом:

– модуль должен осуществлять лемматизацию произвольных фрагментов текстов Системы;

– интерфейс модуля должен предоставлять возможности редактирования связей текстовых прецедентов и единиц словаря вручную;

– модуль должен осуществлять сохранение полученных результатов в базе данных Системы с последующим их просмотром через веб-интерфейс;

– модуль должен обеспечивать лемматизацию внешних для Системы текстов, которые могут быть загружены а) путем ввода через веб-формы, б) путем загрузки через веб-формы в формате XML, в) путем передачи текста в формате XML через веб-сервис, а также возможность использования загруженного текста и результатов лемматизации в других модулях Системы (многотекстовый модуль, модуль выборок и запросов);

– модуль должен обеспечивать выгрузку результатов лемматизации в формате XML;

– модуль должен иметь интерфейс для ввода в грамматический словарь Системы новых лексем и словоформ для лемматизации нераспознанных текстовых прецедентов, для которых отсутствует форма в грамматическом словаре;

– лемматизация текстов должна осуществляться поочередно;

– в модуле должно быть реализовано управление очередью заданий (просмотр, отмена, удаление) через веб-интерфейс:

- доступ к очередям всех пользователей для администратора;
- доступ только к личной очереди для зарегистрированных пользователей;

• лемматизация текста строго ограниченного объема в режиме онлайн для незарегистрированных пользователей (в демонстрационных целях, не более определенного количества раз за сутки);

– в модуле должна быть предусмотрена возможность настройки автоматической отправки пользователю по электронной почте результатов лемматизации в формате XML.

Автоматический морфологический анализатор (лемматизатор) имеет трехуровневую архитектуру и включает следующие компоненты: клиентское приложение, сервер приложений и сервер базы данных.

Клиентское приложение – это интерфейсный компонент, предназначенный для работы конечного пользователя. Оно не имеет прямых связей с базой данных, не нагружено основной логикой системы. Этот уровень обеспечивает авторизацию пользователя, управление процессом автоматической лемматизации, просмотр результатов лемматизации.

Клиентское приложение лемматизатора имеет веб-интерфейс и полностью поддерживается всеми современными веб-браузерами. Для его создания была использована свободная среда разработки программного обеспечения на основе СУБД Oracle – Oracle Application Express.

На втором уровне архитектуры располагается сервер приложений. Здесь сосредоточена основная часть логики системы. В его функции входит создание и поддержание в актуальном состоянии кэшей словаря и текстов, управление пользовательской очередью заданий на лемматизацию, предварительная обработка данных, их подготовка к лемматизации, собственно процесс автоматического морфологического анализа словоформ, постобработка результатов анализа.

В системе реализован метод отложенной лемматизации текстов. Эксперименты показали, что процесс автоматического морфологического анализа – достаточно ресурсоемкий процесс. Архитектура системы «Манускрипт» такова, что даже при достаточно небольшом количестве пользователей (порядка десяти), выполняющих лемматизацию одновременно, производительность существенно падает и время, затрачиваемое на обработку одного не очень большого текста, может достигать нескольких часов. В связи с этим было принято решение, что процесс лемматизации не будет иметь прямой связи с командами пользователя, а будет выполняться автономно, невидимо для пользователя системы. Кроме этого, было решено отказаться от параллельного выполнения лемматизации несколькими пользователями одновременно. Так как в очереди могут находиться задания различного объема, принадлежащие различным пользователям, они выполняются не сразу полностью, а частями определенного размера. Такая методика позволила достичь оптимальной производительности и минимизировать время, необходимое на получение результатов автоматического морфологического анализа.

На третий уровень архитектуры вынесен сервер базы данных. Он обеспечивает хранение всех данных системы: нормализованных словарей, используемых при лемматизации, текстов, а также результатов лемматизации.

В ходе работы требования к лемматизатору и особенно к веб-формам модуля уточнялись, в первую очередь в направлении детализации операций и удобства использования.

Лемматизация включает в себя несколько шагов: создание задания, автоматическую лемматизацию текстов, включенных в задание, просмотр результата и дополнение грамматического словаря, повторную автоматическую лемматизацию, ручное снятие омонимии.

На этапе создания задания пользователь имеет возможность включить в него тексты, выбрав их из списка или указав страницы издания. После запуска лемматизации тексты задания лемматизируются. Об окончании лемматизации информирует изменение статуса задания на «Готово». Окно задания, помимо сведений о задании и помимо списка текстов, содержит количественные сведения об объеме текстов и о результатах лемматизации: общее количество текстовых прецедентов в задании, количество словоформ, для которых найдена одна лемма, количество словоформ, для которых предложено несколько вариантов разбора, количество словоформ, для которых лемма не найдена.

Просмотр результата позволяет выявить нелемматизированные словоформы. Перед повторной лемматизацией задания необходимо добавить в грамматический словарь Системы отсутствующие в базе данных леммы, указав тип парадигмы, или недостающие окончания парадигмы, указав их грамматические значения.

Если при автоматической лемматизации для текстового прецедента находится одна единица словаря, то эта словоформа становится претендентом на автоматическое снятие омонимии на следующем этапе. В других случаях все варианты разбора текстового прецедента сохраняются в базе данных и могут быть отредактированы в дальнейшем в режиме снятия омонимии вручную.

Снятие омонимии может быть осуществлено в нескольких режимах: в режиме просмотра, в режиме редактирования текста, в режиме редактирования повторяющихся в задании словоформ. Интерфейсы этих режимов, помимо окна редактирования, имеют окно просмотра текста, в котором подсвечивается редактируемый текстовый прецедент.

В отличие от режима просмотра, в котором пользователь должен последовательно для каждой словоформы указывать вариант разбора и заносить его в базу данных, режим редактирования позволяет указать правильные варианты для нескольких текстовых прецедентов, что существенно экономит время работы над текстом.

Необходимость предоставить пользователю детальную информацию о результатах лемматизации поставила перед разработчиками задачу типизации текстовых прецедентов с точки зрения их лемматизации и задачу их визуализации в выборках.

Как уже было указано, в Системе используются две процедуры лемматизации: лемматизация без снятия омонимии (автоматическая лемматизация) и лемматизация со снятием омонимии.

Наличие множественных связей между текстовым прецедентом и единицами словаря, установленных автоматически, называется неснятой омонимией. Установление связи текстового прецедента и единицы словаря вручную называется снятием омонимии.

При снятии омонимии может быть установлена связь с несколькими единицами словаря, при этом может быть указана вероятность связи.

В Системе выделяется четыре типа текстовых прецедентов:

а) текстовый прецедент с одной связью, подтвержденной вручную;

б) текстовый прецедент с несколькими связями, подтвержденными вручную;

в) текстовый прецедент с одной связью, неподтвержденной вручную;

г) текстовый прецедент с несколькими связями, неподтвержденными вручную.

Результатом установления нескольких связей может быть связь текстового прецедента α) с разными словоформами одной начальной формы, β) со словоформами разных начальных форм.

В Системе в указатели с неснятой омонимией включаются все прецеденты типа *в*, *г* и с помощью специальных приемов показываются различия между этими типами текстовых прецедентов. В указатели со снятой омонимией включаются прецеденты типа *а*, *б*, *в*, при этом демонстрируются различия между этими типами текстовых прецедентов.

В зависимости от вида указателя применяются различные способы визуализации типов прецедентов.

Для прямого и инверсированного указателей с неснятой омонимией (прецеденты типа *в*, *г*) способ лемматизации (автоматическая) определяется по адресу прецедента, приведенному в фигурных скобках, информация о количестве связей помечена астериском при адресе прецедента, а в случае приведения словоформы к разным леммам используется ссылка на другую лемму. Например:

– наличие одной связи – {*адрес},

– наличие нескольких связей типа α – {**адрес},

– наличие связей типа β – {**адрес см. нач._форма}.

Во всех видах словоуказателей со снятой омонимией (прецеденты типа *а*, *б*) скобки отсутствуют, кроме случаев сохранения у прецедента нескольких связей вследствие неоднозначного определения леммы:

– наличие одной связи – адрес (скобками не отмечается),

– наличие нескольких связей типа α – [**адрес],

– наличие связей типа β – [**адрес см. нач. форма].

В количественном словоуказателе в скобки заключается начальная форма (лемма).

Количественный указатель с неснятой омонимией применяется в следующих случаях:

– при наличии у начальной формы текстовых прецедентов только с одной связью – {*лемма};

– при наличии у начальной формы текстовых прецедентов с несколькими связями типа α – {**лемма};

– при наличии у начальной формы текстовых прецедентов с несколькими связями типа β – {**лемма см. лемма}.

Количественный указатель со снятой омонимией демонстрирует следующее:

– наличие у начальной формы текстовых прецедентов только с одной связью – лемма (скобками не отмечается);

– наличие у начальной формы текстовых прецедентов с несколькими связями типа α – [**лемма];

– наличие у начальной формы текстовых прецедентов с несколькими связями типа β – [**адрес см. лемма].

Особый состав и структуру имеет слово- и формоуказатель: в него включаются грамматические пометы лемм и словоформ, в пределах одной словарной статьи демонстрируются результаты автоматической лемматизации и результаты снятия омонимии. Словоформы со снятой и неснятой омонимией размещаются в разных частях словарной статьи: сначала приводятся все словоформы со снятой омонимией, затем – с неснятой омонимией. В структуре указателя предусмотрена демонстрация сложных словоформ и составных лингвистических единиц (например, предложно-падежных форм).

Компоненты слово- и формоуказателя: аббревиатура текста, аббревиатура фрагмента текста, адрес в рукописи, лемма, лингвистическая единица, компонент составной лингвистической единицы, количество лингвистических единиц, грамматические пометы (признаки основы, признаки парадигмы, признаки словоформы), признаки приоритета, номер омонима и некоторые другие.

Имеющиеся в указателях гипертекстовые переходы от компонентов статьи позволяют получить дополнительную информацию о единицах выборки. Так, переход от индекса парадигмы леммы позволяет получить список слов, имеющих идентичную лемме парадигму; переход от леммы – парадигму леммы с визуализацией словоформ, представленных в текстах Ломоносова прецедентами, от словоформы парадигмы – к текстовым прецедентам и др.

Впереди у коллектива большая работа по снятию омонимии в корпусе. Пока эта трудоемкая часть подготовки корпуса выполняется вручную. Анализ результатов автоматической лемматизации только на основе грамматического словаря современного русского языка показывает, что количество текстовых прецедентов, для которых найдено несколько омонимичных словоформ колеблется от 43% до 55% и более в разных по жанру и тематике текстах, а количество нелемматизированных текстовых прецедентов колеблется от 4% до 10%. Так, в восьмом томе из 91 654 текстовых прецедентов 46% приведены к лемме однозначно, 45% имеют несколько вариантов морфологического разбора, 9% – не имеют леммы; в десятом томе из 16 847 текстовых словоформ 46% имеют один вариант разбора, 47% – несколько вариантов, 6% – ни одного варианта.

В то же время использование для лемматизации грамматического словаря, в парадигмы которого добавлены наиболее часто встречающиеся в текстах Ломоносова морфологические показатели изменяемых частей речи, уже на первом этапе существенно снижает процент нелемматизированных текстовых прецедентов.

Применение автоматических методов снятия омонимии, основанных на количественных и вероятностных характеристиках корпусов, на синтаксическом автоматическом анализе, на контекстной сочетаемости омоформ и на некоторых других основаниях (см., например, [13]), предполагает наличие корпусов, предварительно размеченных вручную, а также доработку методов, алгоритмов и процедур в соответствии с языковым узусом XVIII века.

Коллектив благодарит Российский гуманитарный научный фонд за финансовую поддержку проекта «Большой корпус русского языка XVIII в.», которая позволила создать корпус М.В. Ломоносова в Интернете и инструменты работы с ним (проект № 07-04-12147в).

Summary

V.A. Baranov. Complete Works by M.V. Lomonosov on the Internet: Working up the Electronic Collection and the Corpus Modules Functional Capabilities.

The article deals with the theoretical and applied issues of working up and creation of electronic corpus of M.V. Lomonosov's texts and its use in linguistic and historical research. The basic attention is given to the tools and methods for automatic processing and analysis of the data by special web-modules – the lemmatization and data sampling.

Key words: corpus linguistics, linguistic information electronic processing, full-text databases, linguistic Internet resources, language of M.V. Lomonosov.

Источники

ПСС – *Ломоносов М.В.* Полное собрание сочинений: в 11 т. / Глав. ред.: С.И. Вавилов, Т.П. Кравец. – М.; Л.: Изд-во АН СССР, 1950–1983.

Литература

1. *Савчук С.О., Сичинава Д.В., Гарипов И.И.* Подкорпус текстов XVIII века в составе Национального корпуса русского языка: из опыта работ. – URL: http://fccl.ksu.ru/issue_spec/docs/Savchuk_Sichinava_Garipov.doc.
2. *Марков В.М.* Несколько слов о подготовке «Словаря языка Ломоносова» и о задачах Ломоносовских чтений // Очерки по истории русского языка и литературы XVIII века. Ломоносовские чтения. – Казань: Изд-во Казан. ун-та, 1967. – Вып. 1. – С. 3–5.
3. Очерки по истории русского языка и литературы XVIII века. Ломоносовские чтения: Материалы конф. / Науч. ред. В.М. Марков. – Казань: Изд-во Казан. ун-та, 1967. – Вып. 1. – 162 с.
4. Очерки по истории русского языка и литературы XVIII века / Отв. ред. В.М. Марков. – Казань: Изд-во Казан. ун-та, 1969. – Вып. 2–3. – 324 с.
5. Очерки грамматики и лексикологии русского языка: Сб. ст. / Науч. ред. В.М. Марков. – Казань: Изд-во Казан. ун-та, 1977. – 224 с.
6. Развитие синонимических отношений в русском литературном языке второй половины XVIII века: Сб. ст. / Науч. ред. В.М. Марков. – Казань: Изд-во Казан. ун-та, 1972. – 164 с.
7. *Заиконникова Т.П.* О словаре поэтического языка М.В. Ломоносова // Вестн. УдГУ. – 1993. – № 4. – С. 38–42.
8. *Баранов В.А.* Проект «Манускрипт»: предварительные итоги // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам: Материалы междунар. науч. конф. (Казань, 26–30 авг. 2008 г.) / Отв. ред. В.А. Баранов, В.Д. Соловьев. – Казань: Изд-во Казан. ун-та, 2008. – С. 32–36.
9. *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 авг. 2008 г.) / Отв. ред. В.А. Баранов, В.Д. Соловьев. – Казань: Изд-во Казан. ун-та, 2008. – С. 23–27.

10. *Баранов В.А., Вотинцев А.А., Вотинцев П.А., Соломенников И.С.* Интернет-средства поиска и визуализации данных для лингвистического анализа информационно-аналитической системы «Манускрипт» // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам: Материалы междунар. науч. конф. (Казань, 26–30 авг. 2008 г.) / Отв. ред. В.А. Баранов, В.Д. Соловьев. – Казань: Изд-во Казан. ун-та, 2008. – С. 64–68.
11. *Романенко В.А., Желонкин А.Н.* Разработка и использование формата mnsXML для подготовки, обработки и обмена электронными текстами // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам: Материалы междунар. науч. конф. (Казань, 26–30 авг. 2008 г.) / Отв. ред. В.А. Баранов, В.Д. Соловьев. – Казань: Изд-во Казан. ун-та, 2008. – С. 238–240.
12. *Желонкин А.Н., Романенко В.А.* Описание спецификации формата mnsXML и примеры его практического использования // Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам: Материалы междунар. науч. конф. (Казань, 26–30 авг. 2008 г.) / Отв. ред. В.А. Баранов, В.Д. Соловьев. – Казань: Изд-во Казан. ун-та, 2008. – С. 108–110.
13. *Зеленков Ю.Г., Сегалович И.В., Титов В.А.* Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов. – URL: http://www.dialog-21.ru/Archive/2005/Zelenkov%20Segalovich/Zelenkov_Segalovich.htm, свободный.

Поступила в редакцию
12.04.10

Баранов Виктор Аркадьевич – доктор филологических наук, профессор, заведующий кафедрой лингвистики Ижевского государственного технического университета, старший научный сотрудник лаборатории по автоматизации филологических работ Удмуртского государственного университета.

E-mail: *victor.a.baranov@gmail.com*