МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА НА РЕПУБЛИКА БЪЛГАРИЯ
КИРИЛО-МЕТОДИЕВСКИ НАУЧЕН ЦЕНТЪР ПРИ БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМ. М. Т. КАЛАШНИКОВА
НАУЧНОЕ СООБЩЕСТВО “ПИСЬМЕННОЕ НАСЛЕДИЕ”
DIGITAL MEDIEVALIST SCHOLARLY COMMUNITY
ФОНДАЦИЯ „УСТОЙЧИВО РАЗВИТИЕ НА БЪЛГАРИЯ “

Писменото наследство и информационните технологии

El’Manuscript–2014

Материали от V международна научна конференция
Варна, 15–20 септември 2014 г.

София · Ижевск
2014

Сборникът е издаден с финансовата подкрепа на Министерството на образованието и науката на Република България по процедура за подкрепа на международни научни форуми.

Отговорни редактори: проф. дфн В. А. Баранов
доц. д-р В. Желязкова
д-р А. М. Лаврентьев

Редактори: Нели Ганчева, Веселка Желязкова (български текст)
О. В. Зуга, В. А. Баранов (руски текст)
Кевин Хокинс (Kevin Hawkins) (английски текст)

Писменото наследство и информационните технологии [Текст] : материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.) / отг. ред. В. А. Баранов, В. Желязкова, А. М. Лаврентьев. — София ; Ижевск, 2014. — 448 с.

Сборникът съдържа материали от конференция, посветена на разработването и създаването на съвременни средства за съхраняване, описване, обработка, анализ и публикуване на ръкописни и старопечатни книжовни паметници и исторически извори, а също и на въпросите за подготвянето на електронни ресурси в областта на хуманитаристиката и тяхното използване в научните изследвания и преподаването.

ISBN 978–954–9787–25–2

Модель разметки транскрипций рукописей системы “Манускрипт”: гаплография, диттография, вставки, пропуски и комментарии[1]

В. А. Баранов, Р. А. Аникина, Р. М. Гнутиков

Славянские рукописи, транскрипция, разметка, гаплография, диттография

A Model of Transcription Markup for Manuscripts in the “Manuscript” system: Haplography, Dittography, Inserts, Omissions and Comments

Victor Baranov, Regina Anikina, Roman Gnutikov

This paper is devoted to the means of transmission of haplography, dittography and some graphic-phonetic features in preparation of electronic transcription in the information-analytical system “Manuscript”. The paper describes the ways of differentiating features and their visualization in contexts and in linguistic indexes in a corpora of texts by Mikhail Lomonosov (lomonosov.pro) and in a historical corpus of medieval Slavic manuscripts (manuscripts.ru).

В работе [Баранов, Аникина 2013] рассмотрены типы утрат, правки и добавлений в средневековом тексте и методика и приемы его разметки при подготовке электронной машиночитаемый транскрипции[2] в информационно-аналитической системе “Манускрипт” (портал системы — manuscripts.ru; о ИАС “Манускрипт”, см., например, [Баранов 2012а; Баранов 2012б]). Внесение дополнительной информации обусловлено необходимостью сохранения максимально полных сведений о корректуре и редактуре конкретного документа с целью дифференциации первоначально написанного и внесенных изменений для осуществления выборки, обработки, упорядочения и визуализации данных в автоматическом режиме с учетом или без учета утрат и правки текста [Там же, 182].

[1] Работа выполнена при финансовой поддержке Российского гуманитарного научного фонда / Russian Foundation for Humanities (РГНФ / RFH), проект № 14–04–00585 “Корпус языка М. В. Ломоносова: аналитическая и лингвистическая разметка”.

[2] Машиночитаемая транскрипция — форма передачи текста, первично существующего в виде звучания, фиксации на бумаге, пергамене или другом материальном носителе, которая обеспечивает воспроизведение буквенного и символьного состава текста в виде последовательности электронных двоичных кодов, соответствующих минимальным единицам первичного текста и соотнесенных с соответствующими изображениями на материальном носителе или экране, которая (последовательность) предназначена для автоматизированной и автоматической обработки текста-транскрипции с помощью ЭВМ.
Разбиение транскрипции на цепочки, соответствующие лингвистическим единицам и текстовым блокам, и указание границ с помощью предусмотренных системой средств называется разметкой — лингвистической и аналитической.

Основными характеристиками, положенными в основу предложений, являются: а) способ и степень удаления части текста, б) достоверность прочтения утраченного текста, в) способ восполнения, правки, добавления, г) авторство и время корректуры, редактуры, конъектуры, д) средство передачи утраты в транскрипции и некоторые другие.

Помимо необходимости сохранить в транскрипции сведения об удаленных компонентах написанного, о изменении частей текста в ходе создания и бытования документа, перед создателем транскрипции стоит также задача отметить и прокомментировать в тексте особенности, которые представляют собой нестандартные способы передачи буквенных сочетаний, словоформ и их сочетаний. Это различного рода описки, которые не были исправлены ни писцом, ни редакторами, ни читателями документа, а также особые способы написания. Такого рода случаи могут быть как нетиповыми, так и типовыми. К типовым относятся гаплография, диттография, графическая ассимиляция, вставка в словоформу неэтимологических буквенных компонентов, лигатурные написания. Разметка таких случаев дает возможность сопоставления написаний со стандартными, организации их поиска и демонстрации как в оригинальной форме, так и в стандартизованной.

Приведем примеры решения задач разметки и визуализации такого рода случаев в информационно-аналитической системе “Манускрипт”.

Гаплография и графическая ассимилиция

Гаплография — пропуск одной или нескольких рядом находящихся/находившихся идентичных букв одной словоформы или двух словоформ: *безла* вм. *беззла*, *естьзя* вм. *естьстьзя*.

Графическая ассимиляция — написание одного символа в соответствии с двумя разными: *ищисла* вм. *изчисла*.

Модель базы данных ИАС “Манускрипт” позволяет указать связь одного символа с двумя словоформами, что обеспечивает хранение в базе сведений о стандартной форме компонентов гаплографии и графической ассимиляции и нахождение таких примеров в большом массиве текстов с помощью стандартного запроса. А для визуализации на сайте используются дополнительный символ “-” и раздельное написание:

- оригинальный текст: *безла*, *естьзя*, *ищисла*;
- преобразованный текст: *бе-зла*, *е-стьзя*, *и-щисла*;
- указатели: бе-з, з-ла; е-сть, сть-зя; и-щ, щ-исла.

Лигатуры на стыке слов

Чаще всего лигатуры (связное написание буквенных символов) используются внутри словоформ — ***тв****орити*. Лигатуры на стыке слов встречаются крайне редко — *н****ау****гъ*. И те и другие раскрываются автоматически на основе таблиц соответствий. При вхождении компонентов лигатуры в две словоформы устанавли-

вается ее связь с обеими, при поиске и демонстрации используются компоненты раскрытой лигатуры:

- оригинальный текст: ***тв****орити*, *н(т)****ау****женочь*, *н****ау****гъ*;
- преобразованный текст: *т+ворити*, *на+т+у же ночь*, *на+угъ*;
- указатели: т+ворити, на+, т+у; на+, +угъ.

Вставка неэтимологических компонентов

Неэтимологический компонент — не соответствующая этимологии морфем буква внутри словоформы или на стыке словоформ — *ндравъ*, *издрая*.

При разметке неэтимологический знак между словоформами связывается только с первой. Поиск и демонстрация могут осуществляться как с учетом, так и без учета этого компонента:

- оригинальный текст: *ндравъ, издрая;*
- преобразованный текст: *н-д-равъ, из-д-рая;*
- указатели: н-д-равъ, из-д-, -рая.

Диттография

Диттография — случайное повторение букв(ы) в словоформе, слов в тексте — *велеликъ*; *великъ день день*.

При повторении букв автор транскрипции указывает лишние, в комментарии словоформы отмечает *Так!*, в комментариях букв — *Лишнее*:

- оригинальный текст: *велеликъ*;
- преобразованный текст: *вел^ел^икъ*;
- указатели: *вел^ел^икъ*.

При необусловленном контекстом повторении словоформы вторая отмечается как лишняя, в ее комментарии указывается *Так! Лишнее*:

- оригинальный текст: *великъденьдень*;
- преобразованный текст: *великъ день ^день^*;
- указатели: *день*, *^день^*.

Комментарии

Нетиповые случаи описок могут быть отмечены неформализованными способами, например, с помощью текстового комментария.

Все единицы базы данных имеют свойство *Комментарий*, в поле которого автором транскрипции может быть помещена любая текстовая информация.

Особым родом комментария является возможность указать вероятность для каждого из вводимых значений единицы. Вероятность устанавливается исходя из шкалы в сто баллов. В частности, с помощью вероятности может быть указана степень видимости символа: при плохой его видимости в оригинале для аналогичной его визуализации на сайте устанавливается значение свойства *Видимое представление* меньше, чем 100 %. На сайте в режиме *Текст оригинальный* символ отображается с меньшей интенсивностью (с большей прозрачностью), в ре-

жиме *Текст преобразованный* такой символ отмечается квадратными скобками. Значение 0 % свойства *Вероятность* присваивается символу, который в оригинале не виден, но может быть восстановлен. В режиме *Текст оригинальный* такой символ практически не отображается, при поиске словоформ и включении их в указатели он будет учитываться и показываться.

Комментирование символов, словоформ, в которых отсутствует правка

Не исправленные писцом, справщиком, редактором, корректором описки (гаплография, диттография, неэтимологические компоненты словоформ и т. п.) передаются в транскрипции в той форме, какую они имеют в оригинале, а словоформа(ы) отмечаются с помощью значения свойства *Вероятность*.

При несоответствии формы контексту (буквенного состава словоформы, лишняя словоформа и под.), что отмечается в печатных издания *Sic! Так! Так в рукописи! Так в подлиннике!* и под., устанавливается значение поля *Вероятность* более 100 %. Предлагаемый автором транскрипции правильный вариант указывается в свойстве *Словоформа преобразованная* с дополнительными знаками, которые позволяют различить исходное и восстановленное написания. В свойстве *Комментарий* в соответствии с комментариями печатных изданий *Следует читать... Нужно читать… Читать... Лишнее слово* и под. может быть указано, какую словоформу и почему следует считать правильной.

Сводная таблица передачи правки, добавлений и отсутствия текста

Приведем фрагмент сводной таблицы, содержащий параметры используемой в ИАС “Манускрипт” разметки указанных особенностей исполнения текста.

Тип данных	*Оригинал*	*Преобразование*	*Указатели*	*Вероятность*
Гаплография и графическая ассимиляция	*естьзя; ищисла*	*е-стьзя; и-щисла*	*е-сть, сть-зя; и-щ, щ-исла*	100 %
Лигатуры	***тв**орити; н**ау**гъ*	*т+ворити; на+угъ*	*т+ворити; на+, +угъ*	100 %
Неэтимологические компоненты словоформы	*ндравъ; издрая*	*н-д-равъ; из-д-рая*	*н-д-равъ; из-д-, -рая*	100 %
Диттография	*велеликъ; великъденьдень*	*вел^ел^икъ; великъ день ^день^*	*вел^ел^икъ; великъ, день, ^день^*	>100 %

Предложенные здесь и в [Баранов, Аникина 2013] способы разметки транскрипции в ИАС “Манускрипт” используются также для подготовки текстов в корпусе языка М. В. Ломоносова (lomonosov.pro). Доработка транскрипции и ее разметка, передающая особенности рукописных оригиналов (см. рис. 1, 2), осу-

ществляются на основе примечаний Полного собрания сочинений М. В. Ломоносова в 11 томах (М.; Л., 1950–1959, 1984).

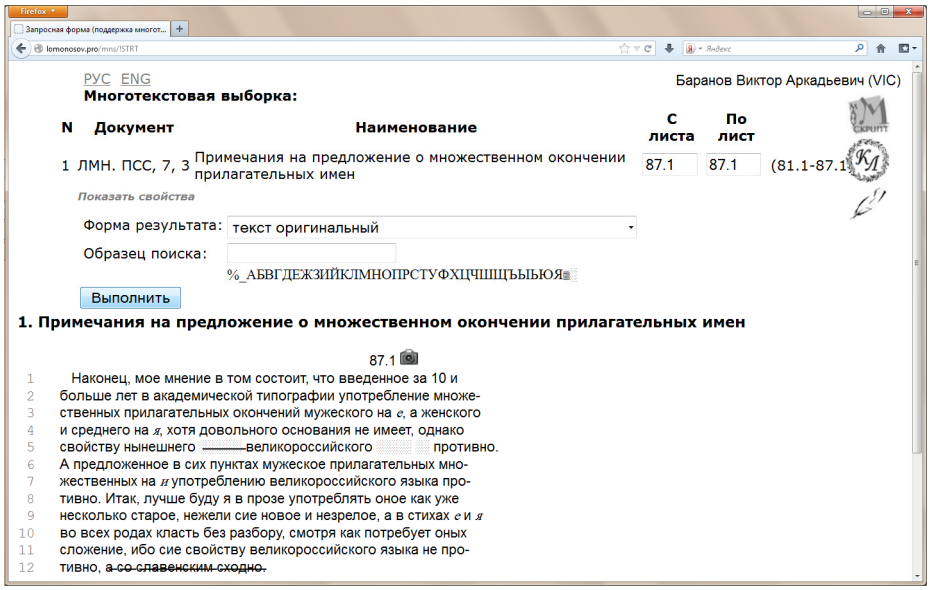

Рис. 1. Исходная транскрипция в корпусе М. В. Ломоносова

Используемые в историческом корпусе славянских рукописей и в корпусе языка М. В. Ломоносова типология и средства разметки утрат, восполнений, дополнений и нестандартных способов исполнения текста позволяют организовать поиск с учетом и без учета таких случаев и понятным пользователю способом визуализировать их при демонстрации текстов и указателей. Кроме того, целью систематизации таких особенностей является и необходимость подготовки средств выгрузки транскрипций во внешний формат с сохранением уже внесенной в документы разметки.

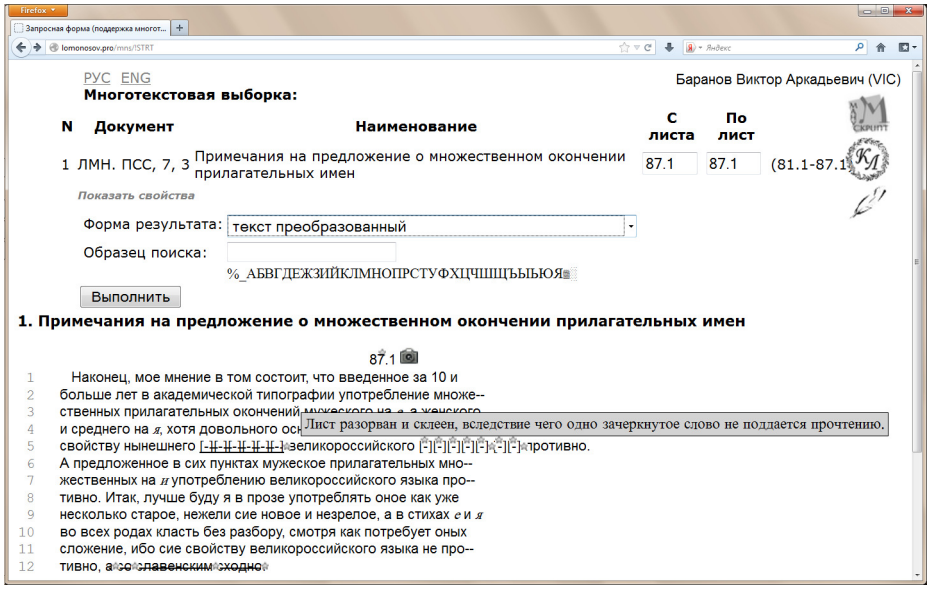

Рис. 2. Визуализация разметки и комментариев в корпусе М. В. Ломоносова

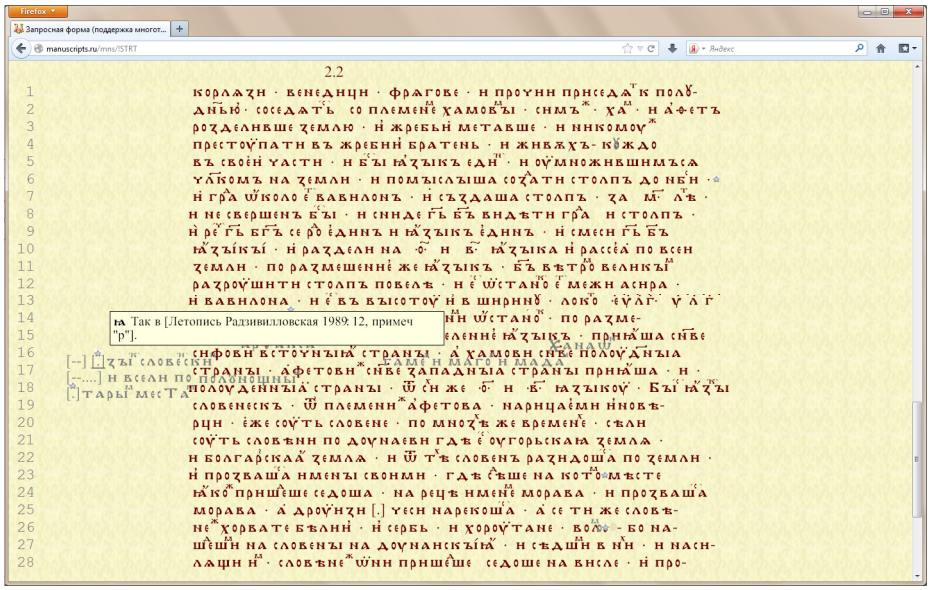

Рис. 3. Визуализация комментариев в историческом корпусе

Литература

Баранов, Аникина 2013 — *Баранов В. А., Аникина Р. А.* Модель текстологической разметки средневековых рукописей: правка и утраты // Труды междунар. конф. “Корпусная лингвистика–2013” (Санкт-Петербург, 25–27 июня 2013 г.) / отв. ред. В. П. Захаров, О. А. Митрофанова, М. В. Хохлова. СПб., 2013. С. 182–192.

Баранов 2012а — *Баранов В. А.* Электронные коллекции древнейших и средневековых славянских рукописей на портале “Манускрипт”: функциональные возможности // Синайский кодекс и памятники древней христианской письменности: традиции и инновации в современных исследованиях. Труды Междунар. научн. конф. “Синайский кодекс. Рукопись в современном информационном пространстве” (Пятые Загребинские чтения) (Санкт-Петербург, 12–13 ноября 2009 г.). СПб., 2012. С. 169–182.

Баранов 2012б — *Баранов В. А.* Лингвистические, методические и технологические вопросы создания и использования корпуса средневековых славянских текстов // Русистика: язык, культура, перевод: сб. докладов юбилейной междунар. научн. конф. (София, 23–25 ноября 2011 г.). София : Изток-Запад, 2012. С. 404–414.