# КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА

УДК 801.82:004.9

*В. А. Баранов*, доктор филологических наук, профессор;
*С. В. Дубовцев*, программист
Ижевский государственный технический университет

## ЭЛЕКТРОННОЕ КРИТИЧЕСКОЕ ИЗДАНИЕ СРЕДНЕВЕКОВОГО СЛАВЯНСКОГО ТЕКСТА: МОДЕЛЬ ДАННЫХ И ВИЗУАЛИЗАЦИЯ ЛИНГВИСТИЧЕСКИХ ЕДИНИЦ

*Статья посвящена модулю электронного критического издания информационно-аналитической системы «Манускрипт», который предназначен для подготовки и демонстрации в Интернете текстологических и лингвистических различий между списками одного и того же текста. Основное внимание уделено описанию модели лингвистических данных, а также параметрам запроса и способам визуализации выборки.*

**Ключевые слова:** полнотекстовая база данных, исторический корпус, критическое издание

В настоящее время компьютерные технологии все активнее используются для решения задач классического языкознания. Этому способствует как появление новых инструментов обработки данных и разработка более совершенных средств визуализации нестандартных объектов, так и расширение круга интересов компьютерной лингвистики, в частности, постановка и решение вопросов, связанных с электронным хранением, анализом и научной публикацией древнейших и средневековых текстов.

В 2008 г. в рамках проекта «Манускрипт» (портал проекта – http://manuscripts.ru) были начаты работы по созданию специализированного *модуля электронного критического издания* (далее – МЭКИ), предназначенного для демонстрации в Интернете списков одного текста с целью предоставления читателю-пользователю возможности сравнения их состава, структуры и лингвистических компонентов.

1. Как известно, традиционное печатное критическое издание средневекового памятника письменности представляет собой научную публикацию, предназначенную для сопоставления списков одного текста. Издание структурировано таким образом, что позволяет увидеть различия между рукописями, возникшие в результате правки, редактирования, вставок, утрат, во время переписывания, в конечном счете – исследовать историю текста или языковые особенности списков на основе текстологических, лингвистических и иных расхождений между рукописями, а при наличии текста на разных языках – установить значимое для историка, богослова, лингвиста, культуролога, литературоведа соотношение перевода и оригинала.

Существующие критические издания, несмотря на значительную вариативность в способах подачи материала, устойчивы в основных своих чертах: наличие одного основного текста, подведение разночтений по отобранным для этого спискам, включение критического аппарата – приложений в виде справочников, указателей, комментариев, а также, в случае «переводности» текста, оригинала на другом языке.

2. Требования к МЭКИ, сформулированные в [1], опираются как на традиции печатных изданий, так и на возможности современных компьютерных технологий, использование которых при создании модуля изложено в [2].

В основе электронного критического издания проекта «Манускрипт» лежит полнотекстовая база данных, содержащая транскрипции нескольких разновремен-

ных древнерусских списков и отрывков майской Минеи, греческий текст Минеи (см. список источников), а также необходимые для работы справочники – словарь фрагментов (малых и больших песнопений – канонов, стихир, кондаков и др.) и пополняемый перечень словоформ, являющиеся, по сути, инвариантами фрагментов и текстовых прецедентов в рукописях (подробнее о информационно-аналитической системе «Манускрипт», модели ее базы данных, веб-модулях обработки текстов, о веб-формах запросов см., например, [3, 4, 5, 6, 7, 8, 9]).

Созданный МЭКИ (URL: http://manuscripts.ru/mns/portal.main?p1=26, раздел «Критическое издание») является инструментом для демонстрации в Интернете: (1) различий в составе и структуре славянских списков и греческого текста; (2) соответствий и разночтений на уровне лексики, морфологии, синтаксиса, словообразования, семантики, графики и орфографии; (3) языковых отношений греческого оригинала и славянского перевода.

В ходе подготовки критического издания было решено и решается несколько конкретных задач: (1) спроектирована и наполняется база данных, содержащая инварианты фрагментов и лингвистических единиц и их связи с текстовыми прецедентами; (2) осуществляется текстологический и лингвистический анализ списков с целью выявления соотношений на уровне текстовых фрагментов и лингвистических единиц; (3) разработаны и созданы процедуры и веб-интерфейсы доступа к данным, обеспечивающие поиск, упорядочение и визуализацию выборки; (4) существенно доработан специализированный редактор OldEd, с помощью которого осуществляется создание и редактирование справочников инвариантов и установление связи между ними и текстовыми прецедентами.

3. Основой лингвистической базы данных электронного критического издания являются связи соответствующих друг другу лингвистических объектов разных списков с единицами справочника-прототекста, которые являются инвариантами текстовых прецедентов.

Структура справочника, включающая листы, страницы, слои, строки и подчиненные строке словоформы и/или синтаксические фрагментов и знаки, представляет собой аналог документа и позволяет автору создать архетип реконструируемого текста.

Приведем пример связи единиц текстовых прецедентов с единицами прототекста.

*Пример 1*: *яви ся* Р1 = *яви си* Р2 = *виденъ* Р3

```
ПТ
лист_ПТ
        страница_ПТ
                слой_ПТ
                        строка_ПТ
                                ССФ_ПТ                  яви ся
                                        ССФ_Р1                  ави ся
                                        ССФ_Р2                  яви си
                                        СФ_Р3                   виденъ
                                СФ_ПТ
                                        КССФ1_ПТ                яви
                                                КССФ_Р1                 ави
                                                КССФ_Р2                 яви
                                        КССФ2_ПТ                ся
                                                КССФ_Р1                 ся
                                                КССФ_Р2                 си,
```

где ПТ – прототекст, лист_ПТ – лист прототекста, страница_ПТ – страница прототекста, слой_ПТ – слой прототекста, строка_ПТ – строка прототекста;

Р1, РN – рукопись 1, рукопись N;

ССФ_ПТ – сложная словоформа прототекста;

ССФ_Р1, ССФ_РN – сложные словоформы рукописей;

СФ_ПТ – словоформа прототекста;

КССФ1_ПТ, КССФN_ПТ – компоненты сложных словоформ прототекста;

КССФ_Р1, КССФ_РN – компоненты сложных словоформ рукописей.

Возможность ввода и редактирования графико-орфографической формы единиц прототекста, изменения их значений, комментирования позволяют автору подготовить подробный справочный аппарат издания.

Для установления соответствий между рукописями на уровне фрагментов используется справочник фрагментов, между единицами и текстовыми фрагментами которого также устанавливаются связи. Единицы справочника могут иметь значения, которые наследуются единицами рукописей. В настоящее время тестируется модуль автоматического нахождения отрывков, максимально совпадающих с частями фрагментированного вручную документа.

Результатом описанной работы является база данных, содержащая информацию о связях между справочником фрагментов и фрагментами рукописей, а также прототекстом и лингвистическими единицами списков.

4. Веб-интерфейс МЭКИ является инструментом для создания запросов, указания состава и формы вывода данных и визуализации выборок. Выборка может быть представлена в одном из предусмотренных вариантов демонстрации фрагментов рукописей, соответствий и разночтений лингвистических единиц.

4.1. В настоящее время в издании предусмотрено четыре формы визуализации соответствий и разночтений лингвистических единиц. Различия между формами заключаются: (1) в типе демонстрируемых отношений – визуализируются или все соответствующие единицы, или только те, между которыми существуют различия; (2) в расположении единиц – текстовые прецеденты расположены вертикально или построчно; (3) в наличии дополнительной информации – форма может включать сведения о типе соответствий; (4) в наличии или отсутствии сводного списка разночтений выборки.

4.2. Важной особенностью МЭКИ является построение перечней соответствий и разночтений выборки между списками на уровне лингвистических единиц, сгруппированных по типам в зависимости от близости или отдаленности соответствующих друг другу компонентов аналогичных контекстов. В связи со значительной степенью графической и орфографической вариативности словоформ для установления степени их близости используется перечень правил, описанных в [8], для установления принадлежности соответствующих друг другу словоформ одной или разным леммам используется автоматический морфологический лемматизатор, с помощью которого текстовые прецеденты приводятся к леммам (см., например, [8]).

4.3. Приведем примеры визуализации двух тропарей из канона святому пророку Иеремии на 1 мая. Режим «На уровне лингвистических единиц (форма и значение разночтений)» соответствует структуре показа разночтений в печатном критическом издании.

*Пример 1.*

Море мнѷное · въꙁдвижемо напастьноѭ боуреѭ · вндѣвъ послѣдьнѩѧ · дворы въꙁлюбилъ ксн · н҃ ѵстока рѣчьно · тоучашта слъꙁы людни своихъ · противьныихъ горько сѧ плака ·~ РНБ, Соф. 202, 2.2.1.6-2.2.1.9

| | |
|---|---|
| 2.2.1.6 **Море** : | РНБ, Соф. 203 3.2.1.21 **Мора** |
| 2.2.1.6 **мнѷное** : | РНБ, Соф. 203 3.2.1.21 **мнрьнаго** |
| 2.2.1.6 **въꙁдвижемо** : | РНБ, Соф. 203 |
| 3.2.1.21 **въздвижема** | |
| 2.2.1.6 **напастьноѭ** : | РНБ, Соф. 203 3.2.1.22 **напастьною** |
| 2.2.1.6 **боуреѭ** : | ГИМ, Син. 166 3.1.1.1 **боурею** |
| | РНБ, Соф. 203 3.2.1.22 **боурею** |
| 2.2.1.7 **послѣдьнѩѧ** : | ГИМ, Син. 166 3.1.1.1 **послѣдьнѩѧ** |
| | РНБ, Соф. 203 3.2.1.23 **послѣдьнаа** |
| 2.2.1.7 **дворы** : | ГИМ, Син. 166 3.1.1.1 **ставнла** |
| | РНБ, Соф. 203 3.2.1.23 **дбьрн** |
| 2.2.1.7 **ѵстока** : | ГИМ, Син. 166 3.1.1.2 **нстоуьнкъ** |
| 2.2.1.8 **рѣчьно** : | ГИМ, Син. 166 3.1.1.2 **рѣкою** |
| | РНБ, Соф. 203 4.1.1.1 **рѣчьна** |
| 2.2.1.8 **тоучашта** : | ГИМ, Син. 166 3.1.1.2 **тоцаша** |
| | РНБ, Соф. 203 4.1.1.2 **тоучаща** |
| 2.2.1.8 **слъꙁы** : | ГИМ, Син. 166 3.1.1.3 **сльꙁы** |
| | РНБ, Соф. 203 4.1.1.2 **сльꙁы** |
| 2.2.1.8 **своихъ** : | РНБ, Соф. 203 4.1.1.2 **твонхъ** |
| 2.2.1.8 **противьныихъ** : | РНБ, Соф. 203 4.1.1.3 **противьнъінхъ** |
| 2.2.1.9 **сѧ** : | ГИМ, Син. 166 3.1.1.3 **сѩ** |
| | РНБ, Соф. 203 4.1.1.4 **съı** |

Как видно, эта форма содержит только разночтения, их тип – графические, орфографические, лексические, словообразовательные – устанавливается самим читателем-пользователем.

Режим «На уровне лингвистических единиц (форма и значение соответствий)» показывает не только разночтения, но и соответствия, он позволяет увидеть степень близости списков, а при необходимости и их соотношение с греческим текстом.

*Пример 2.*

| № | РНБ, Соф. 202 | ГИМ, Син. 166 | РНБ, Соф. 203 | Греч. | РНБ, Соф. 204 |
|---|---|---|---|---|---|
| | – | – | – | О | – |
| 1 | Вс҃ѣхъ | Вс҃ѣхъ | Вьс҃ѣхъ | *πάντων* | – |
| 2 | разумнѣ | разумнѣ | разумне | *τήν γνώσιν* | – |
| 3 | пр҃ѣжде прінімъ | пр҃ѣже прнимъ | пр҃ѣже прнимъін | *προειληφώς* | – |
| 4 | помышленнїа | помышленнїа | помъшленна | *σού* | – |
| 5 | тн | твоего | тн | *τής διανοίας* | – |
| 6 | двнженнїа | двнженна | дв женна | *τάς κινήσεις* | – |
| 7 | пр҃ѣдъзьра | пр҃ѣзра | пр҃ѣзьра | *προθεωρών* | – |
| 8 | о҆ | – | о | *ώ* | – |
| 9 | нєремнѣ | нєремнѣ | неремне | Ιερεμία | – |
| 10 | асе | богоѧвлене | б гласе | *θεοφάντορ* | – |
| 11 | наставьнка | наставьнка | наставьнка | *καθηγητήν* | – |
| | – | тѧ | – | *σε* | – |
| 12 | людьмъ | людьмъ | людьм | *λαού* | – |
| 13 | поставлѧєть | поставлѧєть | пропов҃дает҃ | *προχειρίζεται* | – |
| | 1.1.1.6-1.1.1.8 | 2.1.1.18-2.2.1.3 | 2.1.1.16-2.1.1.20 | 4.1.1.8-4.1.1.10 | Отсутствует |

Этот режим дополняется систематизированным списком соответствий и разночтений. Дадим фрагмент такого перечня для приведенного выше тропаря.

Соотношение РНБ, Соф. 202 (МП) – РНБ, Соф. 203:

Идентичные леммы, идентичные словоформы, идентичный графико-орфографический вариант:

Словоформа – Словоформа (3): **тн – тн**, **о҆ – о**, **наставьнка – наставьнка**.

Идентичные леммы, идентичные словоформы, разные графико-орфографические варианты:

Словоформа – Словоформа (6): **Вс҃ѣхъ – Вьс҃ѣхъ**, **разумнѣ – разумне**, **помышленнїа – помъшленна**, **двнженнїа – дв женна**, **нєремнѣ – неремне**, **людьмъ – людьм** .

Идентичные леммы, идентичные словоформы, разные грамматические формы:

Аналитическая форма – аналитическая форма (1): **пр҃ѣжде прнимъ – пр҃ѣже прнимъін**.

Разные леммы:

Словоформа – Словоформа (3): **прѣдъзьра – прѣзьра, асѥ – б глась, поставлѭѥть – проповѣдаѥть.**

Построение подобных перечней, позволяющих выявить степень близости списков, осуществляется специальной процедурой, использующей сведения о принадлежности соответствующих друг другу словоформ одной или разным леммам, о наличии у словоформ идентичных или различных грамматических значений и сведений о количестве совпадений в графической форме. При этом учитываются следующие признаки лингвистических единиц: (1) форма единиц: совпадение / несовпадение, (2) тип единиц: словоформа, синтаксическая единица, (3) лемма словоформ: идентичные / разные, (4) количество компонентов единиц: одна / несколько, (5) грамматические значения словоформ: совпадение / несовпадение, (6) значение синтаксических единиц: предложно-падежная форма, аналитическая грамматическая форма и некоторые другие.

5. Таким образом, модуль электронного критического издания и критическое издание майской служебной Минеи, подготовленное с помощью модуля, предоставляют читателю возможность:

– познакомиться со списками славянской майской служебной Минеи и греческим текстом Минеи;

– познакомиться с их структурой и составом;

– получить информацию о соответствиях между структурными и лингвистическими единицами рукописей;

– получить сведения о разночтениях между рукописями;

– получить материал для анализа соотношений греческого и славянского текста, т.е. обо всем, что представлено в печатном издании.

Вместе с тем электронное критическое издание, в отличие от печатного, позволяет пользователю самостоятельно формировать страницу электронного критического издания, выбирая необходимые фрагменты текстов или страницы рукописей, располагая фрагменты или словоформы в нужном порядке, выбирая необходимую форму визуализации данных.

В конечном счете, пользователь имеет возможность так настроить представление отобранных для сравнения материалов, что это позволяет ему решать широкий круг задач, связанных с историей текста и с лингвистическими особенностями его списков.

## Благодарности

Создание процедур и веб-интерфейса модуля электронного критического издания информационно-аналитической системы «Манускрипт» выполняется в рамках аналитической ведомственной целевой программы Минобрнауки России «Развитие научного потенциала высшей школы (2009–2010 годы)», регистрационный номер темы 2.1.3/2987; подготовка полнотекстовой базы данных списков майской служебной Минеи осуществляется при поддержке Российского фонда фундаментальных исследований, проект № 09-06-00298.

## Источники

Menaia tou olou eniautou. Akolouphiai Maiou kai Iouniou. En Rome, 1899. T. E´. P. 3-208.

РНБ, Соф. 202, XI в., 135 л.
РНБ, Соф. 203, XII в., 136 л.
ГИМ, Син. 166, XII в., 176 л.
РНБ, Соф. 204, XIII в., 133 л.
БАН, Алекс.-Свирск. 37, пер. пол. XIII в., отрывок, 2 л.
РНБ, F. п. I. 25, кон. XII – нач. XIII в., отрывок, 1 л.
РНБ, ОЛДП, Q. 180, XIV в., отрывок, 7 л.

**Список литературы**

1. *Баранов В. А., Гнутиков Р. М.* Электронное критическое издание средневекового текста: постановка задачи, основные требования и инструментальная подготовка // 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 Каз. гос. ун-та, 2008. – С. 36–44.

2. *Дубовцев С. В.* Электронное критическое издания средневекового текста: инструментальные средства визуализации соответствий и разночтений // Письменное наследие и современные информационные технологии : материалы конкурса науч. работ слушателей междунар. шк. для молодежи (Ижевск, 12–15 окт. 2009 г.) / отв. ред. В. А. Баранов. – Ижевск, 2009. – С. 40–46. URL: http://textualheritage.org/content/view/275/113/lang,russian/ (дата обращения: 18.05.2010).

3. *Баранов В. А., Гнутиков Р. М.* Редактор OldEd как специализированный инструмент для редактирования документов в базе данных «Манускрипт» // Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам : материалы междунар. науч. конф. (Ижевск, 13–17 июля 2006 г.) / отв. ред. В. А. Баранов. – Ижевск : Изд-во ИжГТУ, 2006. – С. 43–46.

4. *Baranov*, *V. A.* The ideology and technology of creating online full-text digital collections of ancient and medieval Slavonic manuscripts // International Conference on Applied Natural Sciences, Trnava, Nov. 7-9, 2007. – Trnava : Univ. sv. Cyrila a Metoda, 2007. – P. 199-207. – ISBN 978-80-89220-91-5.

5. *Baranov, V., Gnutikov, R.* Up-to-date means of access to full-text databases // Digital Humanities 2007 : Conf. Abstr. // The 19th Joint Int. Conf. of the Assoc. for Computers and the Humanities, a. the Assoc. for Literary and Ling. Computing, at the Univ. of Illinois. Urbana-Champaign, USA, June 4 – June 8, 2007. Urbana-Champaign : Graduate School of Library and Inform. Science; Univ. of Illinois, 2007. – P. 74–76. – ISBN: 0-87845-125-0.

6. *Баранов В. А.* Полнотекстовые базы данных как основа для электронных изданий средневековых рукописей в Интернете: требования, реализация, перспективы // Scripta & e-Scripta : The J. of Interdisciplinary Mediaeval Studies. Vol. 6. – Sofia : “Boyan Penev” Publ. Center ; Inst. of Lit., BAS, 2008. – С. 47–64, 422. – ISSN 1312-238X.

7. *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 Каз. гос. ун-та, 2008. – С. 32–36.

8.. Development of the Processing and Visualization Technologies for the Linguistic Information in the Manuscript System: Lemmatization / Victor A. Baranov, Aleksey N. Mironov, Aleksey N. Lapin et al. // Actes des 9es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2008), Lyon, 12-14 mars 2008 : proceedings of 9th International Conference on Textual Data statistical Analysis, Lyon, March 12-14, 2008 / Sci. ed.: S. Heiden, B. Pincemin. – Lyon : Presses Universitaires de Lyon. – 2 vol. – P. 137-145. – ISBN 978-2-7297-0810-8 (pb.). URL: http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2008/pdf/baranov-mironov-lapin-melnikova-sokolova.pdf (дата обращения: 18.05.2010).

9. Интернет-средства поиска и визуализации данных для лингвистического анализа информационно-аналитической системы «Манускрипт» / В. А. Баранов, А. А. Вотинцев,

П. А. Вотинцев и др. // 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 Каз. гос. ун-та, 2008. – С. 64–68.

* * *

*V. A. Baranov*, Doctor of Philology, Professor, Izhevsk State Technical University
*S. V. Dubovtsev*, Programmer, Izhevsk State Technical University

**Electronic critical editions of medieval Slavonic texts: data model and visualization of linguistic units**

*The article is devoted to an Electronic Critical Edition of the Information-Analytical System "Manuscript", which is designed for demonstration of online textual and linguistic differences between the manuscripts of the same text. The focus is on the description of the model of linguistic data, as well as request parameters and visualization of a sample.*

**Keywords:** full-text database, historical corpus, critical edition

Получено 13.05.10

УДК 801.82:004.9

*А.М. Лаврентьев*, кандидат филологических наук, доктор Лионского университета,
научный сотрудник лаборатории ICAR
Национальный центр научных исследований Франции (CNRS)

# ТЕКСТОМЕТРИЧЕСКИЙ ИНСТРУМЕНТАРИЙ В ИССЛЕДОВАНИИ СРЕДНЕВЕКОВОЙ ПУНКТУАЦИИ

*Рассмотрены возможности использования поисково-аналитической машины Weblex для проведения лингвистического исследования на материале корпуса транскрипций средневековых рукописей, содержащих аналитическую и лингвистическую разметку на основе стандарта XML TEI <http://www.tei-c.org>. Объектом исследования была эволюция пунктуации в рукописях французских прозаических текстов XIII–XV веков.*

**Ключевые слова:** текстометрия, корпусная лингвистика, транскрипция рукописи, TEI

Использованный в настоящем исследовании корпус включает 28 «многоуровневых транскрипций» фрагментов рукописей объемом от 550 до 2 250 текстоформ. Общий объем корпуса составляет около 28100 текстоформ. XIII век представлен шестью рукописями, а XIV век – пятью. XV веком датируются 13 рукописей и 2 инкунабулы. Кроме того, в корпус включены две печатные книги первой половины XVI в.

Все тексты корпуса детально описаны в соответствии с принятой в Базе средневекового французского <http://bfm.ens-lsh.fr> системой типологической классификации. Важнейшим параметром типологического описания текста является его принадлежность к определенной функциональной сфере. В использованном корпусе по девять текстов относятся к литературной и к научно-дидактической сферам, шесть текстов – к исторической и два – к религиозной. По одному тексту представляют юридическую и политическую сферу.