Лингвистическая обработка цифровых изданий русских текстов XVIII века

Alexei Lavrentiev, L A Kurysheva

А.М. Лаврентьев, Л.А. Курышева

Лингвистическая обработка цифровых изданий русских текстов XVIII века

Аннотация. В докладе рассмотрены проблемы лингвистической обработки русских текстов XVIII в. на материале двух цифровых корпусов: печатного издания Петровского времени «Аль Коран» и рукописной книги середины века «Повесть о Лабелле и звере». К лингвистической обработке относятся нормализация орфографии, токенизация, морфологическая разметка и лемматизация. Работа была реализована с помощью предварительной разметки в текстовом редакторе Microsoft Word, конвертации в формат TEI и последующей автоматизированной обработки на платформе TXM, включающей применение TreeTagger и построение многоуровневой транскрипции.

Ключевые слова. русский язык и литература XVIII в., нормализация орфографии, электронное издание, платформа TXM, разметка TEI XML, лемматизация.

Language Processing in Digital Editions of Russian 18th Century Texts

A.M. Lavrentiev, L.A. Kurysheva

Abstract. This paper deals with the problems of language processing of Russian 18th century texts that occurred in the work on digital editions of the printed translation of *Al'Quran* (1716) and a manuscript translation of *La Belle et la Bête* (*The Beauty and the Beast*, 1758). The linguistic processing includes spelling normalization, tokenization, morphological markup and lemmatization. The work was carried out using manual pre-markup with Microsoft Word, conversion to TEI XML format and further automatic processing on the TXM platform including annotation with TreeTagger and building multi-layer transcription. In *Al'Quaran* edition the spelling normalization is fully automated but only the simplest cases are dealt with, while in *La Belle et la Bête* manual pre-markup allows generating modern form for all words.

Keywords. Russian language and literature of the 18th century, spelling normalization, digital edition, TXM platform, TEI XML markup, lemmatization.

1. Проблемы создания исторических корпусов на примере русского языка XVIII в.

Для исторической лингвистики наличие качественно размеченных представительных корпусов различных этапов истории языка жизненно необходимо, так как исследователи не могут полагаться на собственную интуицию или языковой эксперимент. При этом распространенные инструменты цифровизации и лингвистической разметки не всегда адаптированы к орфографии, пунктуации и графической сегментации исторических текстов, что приводит к необходимости достаточно больших затрат времени и ресурсов для их «ручной» обработки. Этим объясняется относительно малый объем исторической части многих национальных корпусов. Так, в Национальном корпусе русского языка (НКРЯ) объем подкорпуса XVIII в. составляет приблизительно 6,36 млн с/у, т. е. менее 2 % от общего объема основного корпуса (по данным на март 2021 года).

С.О. Савчук и Д.В. Сичинава о подробно рассматривают принципы отбора и методику обработки текстов XVIII в. для НКРЯ [Савчук и др. 2009]. Частью этой методики является «умеренная модернизация» орфографии, принятая в большинстве академических изданий. Авторы отмечают, что практика модернизации, применяемая в различных изданиях, непостоянна и не всегда последовательна.

Поиск в НКРЯ показывает, что в нем присутствуют тексты как с модернизированной, так и с оригинальной орфографией, что создает определенные неудобства для пользователей. Модернизированная орфография облегчает поиск и лингвистическую разметку, однако оригинальная орфография источников необходима для исследований в области истории морфологии, орфографии и пунктуации.

В настоящем докладе мы представляем опыт создания лингвистически размеченных цифровых изданий памятников XVIII в., в которых проблема совмещения модернизированной и оригинальной орфографии решается с помощью автоматической обработки транскрипций, в которых имеется возможность предварительной ручной разметки отдельных словоформ. Использованное программное обеспечение и сами цифровые издания распространяются свободно на условиях открытых лицензий.

Материалом для настоящего доклада послужили два проекта, непосредственно не связанные между собой. Их объединяет относительное сходство языкового материала и примененная технологическая цепочка создания и обработки цифрового издания.

2. Русский перевод «Корана» (1716 г.)

Цифровое издание первого русского перевода «Корана» является составной частью международного проекта «Coran 12 – 21»[1], направленного на создание параллельного корпуса европейских переводов этой священной книги. В настоящее время корпус включает два издания на арабском, а также различные переводы на латинский, французский и итальянский языки, впервые опубликованные с XVI по XX в.

Первый полный перевод «Корана» на русский язык был осуществлен не с арабского языка, а с французского (издание Дю Рие, 1647) и был напечатан «повелением царского величества» в Санкт-петербургской типографии в 1716 г. Согласно общепринятой вплоть до конца XX в. точке зрения, автором перевода был П.В. Постников [Быкова и др. 1955], однако более поздние исследования показали маловероятность данной атрибуции [Запольская 2002].

Электронная транскрипция текста была проведена вручную Н.В. Луговской с экземпляра, хранящегося в РГБ (шифр МК Си-2°/16-К). К сожалению, ресурсы проекта не позволили оплатить оцифровку источника. Транскрипция в максимальной степени точно воспроизводит орфографию и пунктуацию источника, явные опечатки корректируются в сносках. Дополнительно используются стили заголовков для отражения структуры документа (деление на суры) и специальные коды для указания номеров страниц и стихов.

Текст транскрипции был автоматически переведен из формата .docx в TEI XML с использованием сервиса Oxgarage[2]. Дальнейшая обработка производилась автоматически на платформе TXM [Heiden 2010] с использованием модуля импорта XML TEI Zero и пакета скриптов XSLT, позволяющих привести структурную разметку в соответствие с рекомендациями TEI и добавить слой частично нормализованной орфографии.

Нормализация затрагивает упраздненные в результате реформы орфографии 1918 года буквы и конечный ъ. Как уже отмечалось, такая практика достаточно распространена в академических изданиях текстов XVIII в., однако в электронном издании нормализованная графика не заменяет, а дополняет оригинальную.

Помимо повышения комфорта чтения для неспециалистов нормализация орфографии позволяет использовать для автоматической морфологической разметки и лемматизации лингвистическую модель современного русского языка, созданную для программы TreeTagger [Schmid 1994]. Хотя далеко не все словоформы успешно «осовремениваются» с помощью использованного простейшего алгоритма, частеречная разметка оказывается правильной для 90%, а лемматизация – для 83% словоупотреблений (подсчет проведен на фрагменте объемом 500 словоупотреблений в начале текста).

1 https://coran12-21.org.

2 https://oxgarage.tei-c.org.

3. «Повесть о Лабелле и звере» (перевод 1758 г.)

Опыт более глубокой нормализации орфографии, позволяющей привести к современной форме подавляющее большинство словоупотреблений, был проделан на материале другого памятника, рукописного перевода на русский язык французской сказки «La Belle et la Bête» (заглавие в рукописи - «Повесть о Лабелле и звере»). Перевод был сделан и оформлен в виде подарочной рукописной книги четырнадцатилетней девочкой Хионией Демидовой для своего брата в 1758 г. (Зональная научная библиотека имени В.А. Артисевич Саратовского госуниверситета, рукопись ОРКР № 456). Оригиналом послужила сказка французской писательницы Мари Лепренс де Бомон (Marie Leprince de Beaumont), вошедшая в ее учебник для девочек «Magasin des enfants» (1756). Данный перевод представляет интерес как источник сведений о домашнем образовании и обучении иностранным языкам в России XVIII в., а также о том, как инокультурные реалии воспринимались подростком того времени. Подробное описание источника и проекта цифрового издания представлены в [Курышева и др. 2019], электронное издание доступно в Интернете на портале лаборатории IHRIM[3].

Небольшой объем памятника позволил в сравнительно короткие сроки подготовить пилотное электронное издание с критическим аппаратом и более качественной лингвистической разметкой, чем в случае «Корана».

При этом была использована та же технологическая цепочка, что и в проекте цифрового издания «Корана»: предварительная разметка в Microsoft Word, преобразование в TEI XML и импортирование в TXM с применением пакета скриптов XSLT и автоматической разметки TreeTagger.

В случаях, когда простая замена букв и удаление конечного ъ не позволяли получить современную словоформу, соответствующая форма набиралась вручную в дополнение к форме оригинала. Это позволило создать три слоя графической формы слов: оригинальный, нормализованный и модернизированный. Модернизированный слой скрыт от читателей, однако именно он используется для лингвистической разметки и позволяет использовать современную орфографию в запросах для построения конкордансов и частотных словарей.

Отдельную проблему составляет сегментация текста на словоупотребления. В рукописи встречается немало примеров, когда употребление пробелов не соответствует принятой в современном русском языке норме. Например, *съ начала* (наречие ‘сначала’), *невидя*. В этих случаях простые коды из специальных символов были использованы с тем, чтобы обеспечить отображение графики источника в дипломатическом слое и нормализовать сегментацию в остальных слоях.

4. Заключение

Процесс «ручной» модернизации достаточно трудоемок и не может быть масштабирован на крупные корпусы или коллекции текстов. Основная задача издания «Повести о Лабелле и звере» – служить прототипом для создания пользовательского интерфейса и выработки методики анализа многослойных транскрипций.

Простейший алгоритм, примененный в издании «Корана», показал свою эффективность и в то же время недостаточность для обеспечения высококачественной автоматической аннотации частей речи и лемматизации.

Совершенствование алгоритма автоматической модернизации орфографии может в дальнейшем существенно сократить необходимость «ручной предразметки» словоформ, которая потребуется только в случаях ошибочных или нестандартных написаний.

Мы надеемся, что предложенная методика транскрипции и разработанная технологическая цепочка смогут быть использованы в новых проектах цифровых изданий и корпусов текстов на русском языке в старой орфографии.

Цифровые издания имеют ряд преимуществ, хотя и не отменяют ценности печатных академических изданий памятников. Прежде всего, это доступность и богатство предоставляемого

3 https://txm-ihrim.huma-num.fr/txm/?command=Documentation&path=/LABELLE.

материала и инструментария, которые позволяют решать самые разнообразные задачи и отвечать требованиям различных категорий читателей – от узких специалистов до школьников и широкой публики. Кроме того, подобные издания открывают широкие возможности для новых исследовательских и педагогических проектов.

Литература

1. Быкова Т.А., Гуревич М.М. (1955), Описание изданий гражданской печати: 1708 – январь 1725 г. М.-Л.

2. Запольская Н.Н. (2002), Культурно-языковой статус личности и текста в Петровскую эпоху (опыт прогнозирующего анализа). *Славянская языковая и этноязыковая системы в контакте с неславянским окружением.* М., с. 422–447.

3. Курышева Л.А., Лаврентьев А.М. (2019), Об электронном издании рукописной «Повести о Лабеле и звере» (1758): первый русский перевод сказки «Красавица и зверь» на демонстрационном портале платформы TXM. *Сибирский филологический журнал,* 2019 (1), с. 54–61, DOI: 10.17223/18137083/66/4.

4. Савчук С.О, Сичинава Д.В. (2009), Корпус русских текстов XVIII века в составе НКРЯ: проблемы и перспективы. *Национальный корпус русского языка: 2006 – 2008. Новые результаты и перспективы*. СПб., с. 52–70.

5. Heiden S. (2010), The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, Nov. 2010. Sendai, Japan, Institute for Digital Enhancement of Cognitive Development*, Waseda University, pp. 389-398, available at URL: halshs.archives-ouvertes.fr/halshs-00549764 (дата обращения 20.05.2021).

6. Schmid H. (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees, *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK, available at URL: www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf (дата обращения 20.05.2021).

References

1. Bykova T.A., Gurevich M.M. (1955), Opisanie izdanij grazhdanskoj paechati: 1708 – janvar’ 1725 g. Moscow-Leningrad.

2. Heiden S. (2010), The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, Nov. 2010. Sendai, Japan, Institute for Digital Enhancement of Cognitive Development*, Waseda University, pp. 389-398, available at URL: halshs.archives-ouvertes.fr/halshs-00549764 (last access 20.05.2021).

3. Kurysheva L.A., Lavrentiev A.M. (2019), Ob elektronnom izdanii rukopisnoj “Povesti o Labele i zvere” (1758): pervyj russkij perevod skazki “Krasavica i zver’” na demonstracionnom portale platformy TXM. *Sibirskij filologicheskij zhurnal* 2019 (1), pp. 54–61, DOI: 10.17223/18137083/66/4.

4. Savchuk S.O, Sichinava D.V. (2009), Korpus russkih tekstov XVIII veka v sostave NKRJa: problemy i perspektivy. *Nacional’nyj korpus russkogo jazyka: 2006 – 2008. Novye rezul’taty i perspektivy*. Saint-Petersburg, pp. 52–70.

5. Schmid H. (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing.* Manchester, UK, available at URL www.cis.uni-muenchen.de/sschmid/tools/TreeTagger/data/tree-tagger1.pdf (last access 20.05.2021).

6. Zapol’skaja N.N. (2002), Kul’turno-jazykovoj status lichnosti i teksta v Petrovskuju epohu (opyt prognoziruemogo analiza). *Slavjansakja jazykovaja i etnojazykovaja sistemy v kontakte s neslavjanskim okruzheniem.* Moscow, pp. 422–447.