Корпус русских франкоязычных дневников XIX в. как материал для исследования взаимодействия языков и культур

Alexei Lavrentiev, Michèle Debrenne

М. Дебренн, А.М. Лаврентьев
M. Debrenne, A. Lavrentiev

Корпус русских франкоязычных дневников XIX в. как материал для исследования взаимодействия языков и культур[1]

A Corpus of 19th Century Russian Francophone Diaries as Data for Research on Interaction of Languages and Cultures

Аннотация. В докладе представлена работа по созданию корпуса франкоязычных дневников русских авторов XIX в. Рассмотрены проблемы адекватной исследовательским задачам разметки и предложены технические решения, позволившие оптимизировать работу коллектива проекта. Эксплуатация корпуса осуществляется с помощью платформы TXM – локальной версии и веб-портала.

Ключевые слова: билингвизм, дневник, французский язык, разметка TEI, платформа TXM

Abstract. This paper presents the work on a corpus of francophone diaries written by Russian authors in the 19th century. The problems of the markup necessary for research purposes are considered and technical solutions that allow optimizing the work of the project team are proposed. The local version and the web portal of the TXM platform can be used to work with the corpus.

Keywords: bilingualism, diary, French language, TEI markup, TXM platform

Франкоязычные дневники русской аристократии XIX в. предоставляют интереснейший материал для изучения билингвизма и взаимодействия русской и французской культуры. Создание корпуса таких дневников, снабженного адекватной исследовательским целям разметкой, может способствовать получению качественно новых научных результатов и обеспечить доступ к

[1] Работа проведена с финансовой поддержкой РГНФ проект (№ 16-24-08001) и французского Фонда домов гуманитарных наук (FMSH).

данным материалам широкому кругу представителей различных гуманитарных наук.

Такую цель ставит перед собой проект «Взаимодействие культур в пространстве русского франкоязычного дневника XIX века», проводимый коллективом исследователей Новосибирского государственного университета и лаборатории IHRIM Национального центра научных исследований Франции. Представленная в докладе работа сосредоточена на материале дневниковых тетрадей О. И. Орловой-Давыдовой (1814-1876), хранящихся в ГПНТБ СО РАН и озаглавленных «Journal d'Olga Davidoff». В корпус вошли пять тетрадей, содержащих записи на французском языке, датированные 1835 – 1845 гг, с небольшими более поздними вставками 1847, 1849 и 1869 гг.

На первом этапе проекта были определены виды разметки, необходимые для эксплуатации корпуса. К ним относится выделение имен собственных (топонимов и антропонимов), русскоязычных вкраплений (на кириллице или в транслитерации), различных видов ошибок и исправлений в тексте рукописи. Несколько маркеров разметки могут накладываться на один фрагмент текста (топоним может быть написан кириллицей и содержать исправление). Основу структуры корпуса составляет дневниковая запись. Все записи датированы (одним днем или периодом), однако указанные даты нуждаются в нормализации, так как наблюдается смешение старого и нового стиля и отдельные ошибки (несоответствие даты и дня недели). Начало каждой страницы и ее номер также маркируются, чтобы обеспечить возможность синоптического выведения на экран транскрипции и фотографии страницы рукописи.

Помимо транскрипции проект предполагает подготовку перевода дневника на русский язык. Разметка перевода сводится к минимуму, необходимому для постраничного выравнивания.

Для удобства участников проекта и с учетом технических возможностей был предложен рабочий процесс (workflow), включающий предварительную разметку в Microsoft Word с использованием технологии стилей и специальных сочетаний символов с

последующим автоматическим преобразованием в формат XML с разметкой, соответствующей стандарту международной Инициативы по кодированию текстов (TEI)[2]. Трансформация осуществляется с помощью онлайн сервиса Odette[3] [Glorieux 2015] и последующего применения специально разработанной стилевой таблицы XSLT. Сервис Odette позволяет распознавать структуру документа (по стилям заголовков) и преобразовывать стили символов в стандартные теги TEI. Дополнительное преобразование необходимо для обработки сложных случаев (наложение нескольких видов разметки, нормализация дат и т.п.).

Файлы корпуса в формате XML-TEI импортируются в корпус-менеджер TXM [Heiden 2010] с помощью недавно разработанного модуля импорта XTZ. Этот модуль позволяет в автоматическом режиме применять серию трансформаций XSLT (до и после токенизации) и создавать «синоптическое издание», включающее транскрипцию и фотографию источника. В процессе импортирования текст проходит автоматическую морфологическую разметку и лемматизацию с помощью программы TreeTagger[4], что существенно расширяет возможности качественного и количественного анализа.

На момент публикации сборника все пять тетрадей затранскрибированы и размечены в документе Word, тестовый корпус импортирован на платформу TXM. Он используется для проверки и корректировки транскрипции и предварительной разметки, а также для совершенствования автоматизации процедуры обновления корпуса и разработки сценариев его эксплуатации. Прототип издания корпуса размещен на демонстрационном портале TXM[5]. На основе размеченного текста дневников подготовлены два магистерских исследования, позволяющие уточнить использование имен собственных (антропонимов) в дневниках и

[2] http://www.tei-c.org

[3] http://obvil-dev.paris-sorbonne.fr/developpements/Odette

[4] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger

[5] http://portal.textometrie.org/demo/?command=documentation&path=/DAVYDOVA

особенности индивидуального идиолекта диаристов. Опубликован ряд статей с общим анализом особенностей французского языка О.И. Орловой-Давыдовой [Дебренн 2016a, Debrenne 2016b].

В дальнейшем корпус проекта будет пополняться за счет дневников других авторов. Первым из них является дневник поручика Александра Чичерина (1793-1813), работа над которым уже идет. Особенностью этого дневника является наличие большого числа авторских рисунков, тесно связанных с текстом. Анализ и аннотация этих рисунков и их отношений с фрагментами текста являются отдельной задачей, новой для текстометрических исследований.

Литература

1. *Дебренн М.* (2016a), Сопоставительный девиатологический анализ переписанных дневников О.И. Давыдовой и первичных текстов. Вестник НГУ, Серия Лингвистика и межкультурная коммуникация, 14 (3), с. 59-74.

2. *Debrenne M.* (2016b), The French Language in the Diaries of Olga Davydova. An example of Russian-French Aristocratic Bilingualism. In M. van Strien-Chardonneau & M.-C. Kok (Ed.), Escalle Le français, langue de l'intime à l'époque moderne et contemporaine, Amsterdam : Amsterdam University Press B.V., pp. 125-142.

3. *Glorieux F.* (2015), Processing texts to produced structured data [Le traitement de textes pour produire des documents structurés (XML/TEI)], http://resultats.hypotheses.org/267.

4. *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 – PACLIC24, Sendai, pp. 389-398.

References

1. *Debrenne M.* (2016a), Sopostavitel'nyj deviatologicheskij analiz perepepisannykh dnevnikov O.I. Davydovoj i pervichnykh tekstov [Comparative error-based analysis of the copied Davydova's diary and its original], Vestnik NGU, Serija Lingvistika i mezhkulturnaja kommunikacja [Vestnik SNU, Series Linguisitcs and intercultural communication], 14 (3), pp. 59-74.

2. *Debrenne M.* (2016b), The French Language in the Diaries of Olga Davydova. An example of Russian-French Aristocratic Bilingualism. In M. van Strien-Chardonneau & M.-C. Kok (Ed.), Escalle Le français, langue de l'intime à l'époque moderne et contemporaine, Amsterdam : Amsterdam University Press B.V., pp. 125-142.

3. *Glorieux F.* (2015), Processing texts to produced structured data [Le traitement de textes pour produire des documents structurés (XML/TEI)], http://resultats.hypotheses.org/267.

4. *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 – PACLIC24, Sendai, pp. 389-398.

Дебренн Мишель
Новосибирский государственный университет (Россия).
Debrenne Michèle
Novosibirsk State University (Russia).
micheledebrenne@gmail.com

Лаврентьев Алексей Михайлович
Национальный центр научных исследований (Франция).
Lavrentiev Alexei
Centre national de la recherche scientifique (France).
alexei.lavrentev@ens-lyon.fr