Проблемы лингвистической разметки и анализа электронных критических изданий текстов письменного наследия в стандарте XML-TEI

А.М. Лаврентьев

Alexei.Lavrentev@ens-lyon.fr

Лаборатория ICAR Национального центра научных исследований (CNRS) и Лионского университета, Лион, Франция

XML-TEI, платформа TXM, токенизация, письменное наследие

Summary

In this paper we consider some problems of automatic linguistic annotation and analysis of textual heritage documents encoded according to the TEI XML guidelines. TEI XML is a popular standard for encoding electronic editions of textual heritage documents as it allows highly customizable semantically-oriented markup independent of a particular platform or software. TEI is aimed at facilitating data exchange and interoperability. However, rich editorial markup including various readings and interpretations at various levels of linguistic hierarchy may be a serious challenge if one wants to apply NLP (natural language processing) tools to such an edition. Based on the example of the *Base de Français Médiéval* Old French corpus and on the electronic edition of the *Queste del saint Graal*, we will

discuss the solutions to these problems that are implemented in the TXM platform import modules.

Стандарт электронной разметки текстов в формате XML в соответствии с рекомендациями международной Инициативы по Кодированию Текстов TEI (Text Encoding Initiative, http://www.tei-c.org) приобрел в последние годы достаточно широкое распространение в области электронного издания текстов, принадлежащих к фонду письменного наследия различных стран и культур. 150 проектов представлены на сайте TEI, а в действительности это число может быть значительно бóльшим. Преимуществами разметки в стандарте TEI являются ее опора на тщательно разработанную теорию структуры текста и документа, легкость персонализации и адаптации к конкретному материалу за счет модульной организации и специального механизма спецификации ODD, а также независимость от конкретной платформы или программного продукта. Вместе с тем чрезвычайная гибкость TEI создает определенные трудности для разработки программных средств обработки, анализа и публикации текстов, размеченных в этом стандарте, особенно если речь идет о средствах «широкого профиля», предназначенных для использования вне рамок отдельно взятого проекта. В частности, «глубокую» филологическую разметку, учитывающую разночтения и варианты интерпретации фрагментов текста на разных уровнях иерархии языковых

структур, может быть трудно совместить с использованием инструментов автоматической лингвистической разметки (токенизации, лемматизации, морфологической категоризации и т.п.).

В нашем докладе мы представим методику подготовки (нормализации) филологической разметки текстов Базы средневекового французского языка BFM (Base de Français Médiéval, http://bfm.ens-lyon.fr) в процессе их загрузки на платформу TXM с целью их дальнейшего лингвистического анализа. Мы также рассмотрим результаты опытов по адаптации данной методики к текстам других проектов, размеченным на основе рекомендаций TEI, но применяющим отличные от BFM решения в ряде ключевых для лингвистического анализа аспектов разметки.

Текстометрия (textométrie) возникла как научное направление во Франции в 1980-е годы. В ее рамках были разработаны эффективные методики анализа объемных корпусов текстов. Вслед за лексикометрией и статистическим анализом текста текстометрия предлагает статистически обоснованные методы и инструменты анализа для различных гуманитарных наук.

TXM – это модульная платформа с открытым исходным кодом, которая сочетает функции различных ранее разработанных программ текстометрического анализа. Она представляет новое поколение текстометрического инструментария, использующее современные корпусные

технологии (Unicode, XML, TEI, NLP). Подробная информация о платформе TXM представлена в публикациях [Heiden 2010; Heiden et al. 2010; Pincemin et al. 2010], а также на сайте http://textometrie.ens-lyon.fr/?lang=en.

База средневекового французского языка (BFM) – это корпус текстов старо- и среднефранцузского языка (IX – XV вв.), в настоящее время разрабатывающийся лабораторией ICAR Национального центра научных исследований Франции (CNRS) и Лионского университета. База насчитывает 75 текстов общим объемом более 3 500 000 текстоформ. Источниками BFM в основном являются авторитетные критические издания, однако в последнее время развиваются собственные издания, опирающиеся на лингвистически выверенные транскрипции оригинальных рукописей. В качестве примера можно привести интерактивное издание анонимного романа XIII в. «Поиски Святого Грааля» («La Queste del saint Graal») под редакцией К. Маркелло-Низья [Queste 2011].

С мая 2012 года доступ к BFM осуществляется посредством портала http://txm.bfm-corpus.org/bfm, построенного на платформе TXM.

Все тексты BFM размечены в формате XML на основе рекомендаций TEI, в соответствии со спецификацией, разработанной для нужд проекта [Guillot et al. 2010] с учетом перспективы лингвистического анализа. Платформа TXM не только служит для корпуса BFM средством доступа пользователей, но и

позволяет осуществлять ряд операций автоматической и полуавтоматической разметки (в частности, морфологической аннотации, выявления прямой речи).

Одной из наиболее сложных задач при использовании средств автоматического лингвистического анализа корпусов применительно к текстам, включающим глубокую редакторскую разметку, является корректная идентификация слов (токенизация) и предложений, к которым этот анализ должен применяться без потери самой редакторской разметки.

Следующий пример разметки, содержащий предложенный редактором фрагмент текста на месте лакуны, начинающейся в конце одного слова и заканчивающейся несколькими словами позже, абсолютно корректен с точки зрения рекомендаций TEI, однако весьма сложен для токенизации с учетом особенностей языка XML (запрет «перекрещивания» элементов, риск появления пробелов между тэгами и текстовыми узлами при обработке):

en<supplied>tra a cheval en la</supplied> sale une mout bele damoisele

Еще более серьезные проблемы возникают при разметке предложений, особенно в стихотворных текстах, где перекрещивание метрической и синтаксической структуры встречается очень часто.

Разумеется, можно «отфильтровать» все элементы, перекрещивающиеся с основными лингвистическими структурами (словами и предложениями), однако это может привести к потере существенной информации для запросов

и визуализации (например, о том, подвергалась ли словоформа редакторской правке).

Создание алгоритма токенизации, который корректно обрабатывал бы любой текст с глубокой редакторской разметкой в стандарте TEI XML, представляется практически невозможным. Тем не менее, можно добиться вполне удовлетворительных результатов при условии, что разметка исходного документа отвечает ряду простых правил. Например, «тэги, расположенные внутри слов, должны быть четко идентифицированы» или «если размеченный сегмент текста начинается внутри одного слова и захватывает несколько последующих, его необходимо разделить».

Также возможно составить списки тэгов TEI в зависимости от их позиции в лингвистической иерархии текста в рамках отдельного проекта. Например, предложения располагаются внутри элементов типа «абзац» **<p>** или «блок текста» **<ab>**, а не наоборот. Некоторые тэги можно считать эквивалентными слову (например, **<abbr>**, **<num>**, **<pc>**), а несколько элементов почти всегда располагаются внутри слова (**<am>**, **<c>**, **<ex>**). Ряд элементов содержат сегменты текста, которые не следует токенизировать, поскольку они не принадлежат к материалу источника (например, редакторские примечания и сноски в тексте критического издания). В рамках конкретного проекта эти списки могут быть существенно расширены и уточнены, в результате чего

число элементов, которые могут перекрещиваться с лингвистическими структурами, значительно сокращается.

Разметка текстов BFM соответствует четко сформулированной спецификации применения рекомендаций TEI, отраженной в документации ODD [Guillot et al. 2010]. Данная спецификация включает правила использования внутрисловных тэгов (например, исправленных редактором букв или знаков переноса), а также элементов, которые могут перекрещиваться со структурой предложений (например, «пустые» элементы <lb/> «новая строка» используются вместо <l> «стих», а элемент цитирования <q> рассматривается как граница предложения). Это позволило реализовать в платформе TXM эффективный алгоритм токенизации для текстов BFM.

В последнее время был успешно проведен ряд тестов по адаптации данного алгоритма к документам TEI XML, подготовленным в других проектах. Среди них можно назвать «Виртуальную библиотеку гуманистов» (http://www.bvh.univ-tours.fr) и издание черновиков романа «Бувар и Пекюше» Гюстава Флобера (http://dossiers-flaubert.ish-lyon.cnrs.fr). Адаптация состоит в применении специальных фильтров XSL на входе и на выходе процедуры токенизации. «Входной» фильтр удаляет тэги, которые не представляют интереса для эксплуатации с помощью TXM, а также упрощает и нормализует некоторые сложные структуры элементов XML. «Выходной» фильтр позволяет исправить ряд ошибок, которых практически невозможно избежать в процессе

первичной токенизации (например, деление слова при переносе в конце страницы, когда целый ряд тэгов может располагаться между его началом и концом).

В тексте доклада мы приведем конкретные примеры реализованных решений и продемонстрируем результаты, которые можно получить при эксплуатации текстов BFM и других проектов с помощью платформы TXM. Более подробно различные алгоритмы токенизации, реализованные на платформе TXM, описаны в [Heiden 2010].

Список литературы

Guillot, C., Heiden, S., Lavrentiev, A., Bertrand, L. Manuel d'encodage XML-TEI des textes de la Base de Français Médiéval, Lyon: Équipe BFM, 2010. – Адрес в Интернет http://bfm.ens-lyon.fr/IMG/pdf/Manuel_Encodage_TEI.pdf.

Heiden, S. The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. 24th // Pacific Asia Conference on Language, Information and Computation / Ed. Kiyoshi Ishikawa Ryo Otoguro. Institute for Digital Enhancement of Cognitive Development, Waseda University, 2010. P. 389-398.

Heiden, S., Magué, J.-P., Pincemin, B. TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement // Statistical Analysis of Textual Data - Proceedings of 10th International Conference JADT 2010 / Bolasco,

S. et al. (Eds.). – Rome: Edizioni Universitarie di Lettere Economia Diritto, 2010. – P. 1021-1032. – Адрес в Интернет http://www.ledonline.it/ledonline/JADT-2010/allegati/JADT-2010-1021-1032_025-Heiden.pdf.

Pincemin, B., Heiden, S., Lay, M.-H., Leblanc J.-M. and Viprey, J.-M. Fonctionnalités textométriques: Proposition de typologie selon un point de vue utilisateur. // Statistical Analysis of Textual Data - Proceedings of 10th International Conference JADT 2010 / Bolasco, S. et al. (Eds.). – Rome: Edizioni Universitarie di Lettere Economia Diritto, 2010. – P. 341-353. – Адрес в Интернет http://www.ledonline.it/ledonline/JADT-2010/allegati/JADT-2010-0341-0354_023-Pincemin.pdf.

Queste del saint Graal. Édition numérique intéractive / Ed. Marchello-Nizia, Ch. – Lyon: Équipe de la BFM, 2011. – Адрес в Интернет http://txm.bfm-corpus.org/txm.