Уровни представления данных в транскрипции средневековых европейских рукописей на основе XML – TEI

А. М. Лаврентьев

Высшая гуманитарная школа, Лион, Франция

Традиционные научные издания средневековых западноевропейских рукописей можно разделить на две большие категории: критические и дипломатические. Критические издания, касающиеся в первую очередь литературных текстов, характеризуются стремлением передать содержание оригинального произведения, сведя к минимуму ошибки переписчиков и представив текст в удобной для современного читателя форме. Дипломатические издания, типичные для административных и юридических документов, отличаются достоверной передачей текста источника, даже если он содержит явные ошибки. Конкретные характеристики различных типов изданий зависят от языка текста и от национальной издательской традиции. В своей работе мы остановимся на традиции издания средневековых французских текстов во Франции, однако предлагаемые нами принципы применимы к изданию текстов западноевропейской рукописной традиции в целом.

Оставим в стороне вопрос установления текста на основе сопоставления нескольких рукописных источников.

Во Франции сложилась относительно устойчивая традиция подготовки критических изданий, опирающаяся на принципы, провозглашенные Ж. Бедье [Bédier 1928]. При наличии нескольких рукописных источников выбирается «лучшая рукопись» и издается с минимально необходимыми исправлениями в том, что касается буквенного состава слов текста. При этом, однако, пунктуация и разделение слов с помощью пробела приводятся в соответствие с современными правилами, расшифровываются сокращения, вводится ряд снимающих неоднозначность диакритических знаков, нейтрализуются позиционные и каллиграфические варианты начертания букв [Bourgain 2002]. Традиция дипломатических изданий менее стабильна. Многие издатели применяют практически все те же операции нормализации, что и в критических изданиях, однако ряд ученых стремится к более строгому

воспроизведению данных рукописного источника, вплоть до символов сокращений и каллиграфических вариантов букв.

Традиционное издание, к какому бы типу оно ни относилось, не в состоянии удовлетворить потребностей всех заинтересованных исследователей. Критическое издание с его тенденцией к нормализации не устраивает палеографов и лингвистов, интересующихся историей графических систем. «Гипер-дипломатическое» (или «имитативное») издание с воспроизведением сокращений вариантов букв, сокращений и «ненормативного» словоразделения затрудняет чтение и содержательный анализ текста, в которых заинтересованы историки, литературоведы и большинство лингвистов. Существует небольшое количество печатных «синоптических» изданий, предлагающих несколько параллельных версий текста, а также факсимиле источника, однако такие издания весьма дороги и неудобны в использовании, если объем текста превышает несколько страниц.

Информационные технологии позволяют решить проблему неудобства использования и оптимизировать затраты на подготовку синоптического издания. Читатель электронного издания может выбрать наиболее подходящую для него форму представления текста и при необходимости легко переходить от одной формы к другой. Многие операции при подготовке различных форм представления текста в электронном издании могут быть автоматизированы, что существенно снижает его стоимость.

Тем не менее само по себе использование информационных технологий не снимает необходимости научного анализа и интерпретации данных, отражаемых в издании. Более того, новые технические возможности требуют обновления методики подготовки текстов к изданию. С учетом стремительного развития технологий, частой смены форматов и программного обеспечения выработка общепринятых, научно обоснованных стандартов кодировки и разметки текстов в электронных изданиях приобретает особое значение. Использование международно признанных стандартов должно облегчить адаптацию издания к технологическим изменениям, а научная обоснованность разметки – обеспечить его ценность для широкого круга исследователей.

Международная Инициатива по кодированию текстов (Text Encoding Initiative, TEI, URL: http://www.tei-c.org), разрабатывающая универсальные рекомендации по разметке электронных изданий, основанные на стандарте XML, играет в этом отношении первостепенную роль. Отдельная рабочая группа TEI готовит рекомендации для транскрипции рукописей, однако опубликованные к настоящему времени рекомендации не отвечают в полной мере потребностям многоуровневых научных изданий средневековых западноевропейских рукописей.

По нашему мнению, различные формы представления транскрипции средневековой рукописи должны соответствовать уровням нормализации при лингвистическом анализе. Эти уровни сопоставимы с тремя типами транскрипции устной речи: орфографическим, фонематическим и фонетическим (на уровне аллофонов). Традиционные правила транскрипции, применяемые в критических изданиях, можно охарактеризовать как *орфографические*: они не просто передают графемы источника, но в ряде случаев вводят дополнительные знаки, облегчающие чтение текста (например, различение *u* и *v* по принципу гласный / согласный, использование акцента над ударным *é* в конце многосложных слов). Пунктуация и графическая сегментация (разделение слов пробелами) на этом уровне транскрипции приведены в соответствие с современными нормами.

Ряд дипломатических изданий приближается к *графематическому* уровню транскрипции: его принцип состоит в передаче одним знаком различных вариантов графемы. Сокращения на этом уровне транскрипции расшифровываются, однако восстановленные при этом буквы типографически выделяются (например, курсивом). Графематический принцип применим и к знакам пунктуации: если анализ графической системы рукописи показывает противопоставление двух типов пунктуации: «сильной» (любой знак препинания с последующей прописной буквой) и «слабой» (знак препинания с последующей строчной буквой), то сильная пунктуация может передаваться точкой, а слабая – запятой. При наличии в рукописи более сложной системы пунктуации могут быть использованы дополнительные знаки. Графическая сегментация может быть нормализована в

графематической транскрипции для удобства чтения, однако возможно использование специальных знаков для обозначения «аномалий» (с точки зрения современной системы). Так, агглютинация (слитное написание двух слов) может обозначаться знаком «+», а деглютинация (пробел «внутри» единого слова) – знаком «_».

Наконец, так называемые «имитативные» издания соответствуют *аллографическому* уровню транскрипции: они отражают лингвистически значимые варианты букв, знаков препинания и символов сокращений. Графическая сегментация на этом уровне транскрипции строго соответствует источнику.

Систематическое различение трех уровней транскрипции было обосновано в [Haugen 2004] и впервые реализовано в проекте Архива средневековых нордических текстов (Medieval Nordic Text Archive, MENOTA, URL: http://www.menota.org). В рамках проекта Базы средневекового французского языка (Base de Français Médiéval, BFM, URL: http://bfm.ens-lsh.fr) предложения проекта MENOTA были адаптированы к материалу французских рукописей и уточнены с учетом уровней лингвистического анализа.

В докладе мы остановимся на технической стороне транскрипции рукописей в проекте BFM, в частности, на вопросах эргономии и решении проблемы синхронизации различных уровней транскрипции при внесении исправлений и дополнительной разметки.

Список литературы

Bédier 1928 – *Bédier, J.* La tradition manuscrite du *Lai de l'Ombre*, réflexions sur l'art d'éditer les anciens textes / Joseph Bédier // Romania. –1928. – Vol. 54. – P. 161-196.

Bourgain 2002 – *Bourgain, P.* Conseils pour l'édition des textes médiévaux. Fascicule III, Textes littéraires / Pascale Bourgain, Françoise Vielliard. – Paris : C.H.T.S. ; École nationale des chartes, 2002. – 253 p.

Haugen 2004 – *Haugen, O. E.* Parallel Views: Multi-level Encoding of Medieval Nordic Primary Sources / Odd Einar Haugen // Literary and Linguistic Computing. – 2004. – № 1. – P. 73-91.

Levels of representation of data in transcriptions of medieval European manuscripts using TEI

Alexey M. Lavrentiev

École normale supérieure Lettres et sciences humaines, Lyon, France

In this paper we consider various traditions of editing medieval French texts (critical, diplomatic and “imitative”) and suggest a distinction of transcription levels based on a linguistic analysis. These levels can be implemented in multi-layer electronic editions that would meet the requirements of various types of users (paleographers, linguists, literary scholars, historians, etc.).