ный бюллетень Ассоциации «История и компьютер». № 32 : материалы IX конф. АИК (апр. 2004 г.) — Москва ; Томск : Изд-во Том. ун-та, 2004. — С. 7–9.

Баранов и др. 2003 — *Баранов, В. А.* Электронные издания древних письменных памятников и технология создания полнотекстовых баз данных / В. А. Баранов, А. А. Вотинцев, Р. М. Гнутиков, О. В. Зуга, А. Н. Миронов [и др.] // Круг идей: электронные ресурсы исторической информатики : тр. VIII конф. Ассоциации «История и компьютер» / под ред. Л. И. Бородкина, В. Н. Владимирова. — М. ; Барнаул : Изд-во Алт. ун-та, 2003. — С. 234–270.

Использование формата TEI для обмена данными с полнотекстовой информационно-поисковой системой «Манускрипт»[1]

П. А. Вотинцев
Удмуртский государственный университет, Ижевск, Россия

The project contributes to the development of the means for the exchange of documents and their meta- and analytical description (under the XML-TEI format, http://www.tei-c.org/) with the full-text databases giving means for a multipurpose processing of the document objects and ensuring creation of electronic publications for various purposes on the Internet (Information Retrieval System "Manuscript", http://manuscripts.ru/).

Работа по созданию электронных изданий древних рукописных памятников в настоящее время ведется различными группами исследователей во многих странах мира. Использование компьютерных баз данных для анализа древних текстов представляется очень перспективным, особенно в связи с развитием Internet-технологий. Однако существует проблема преобразования данных разных форматов.

Проект предусматривает разработку средств обмена данными между форматами хранения документов и их мета- и аналитического описания (на основе формата XML-TEI[2]) и полнотекстовыми базами данных, предоставляющими средства для многофункциональной обработки объектов документа и обеспечивающими создание электронных публикаций различного назначения в Интернете (ИПС «Манускрипт», http://manuscripts.ru/).

[1] Работа по созданию ИПС «Манускрипт» ведется при поддержке Российского фонда фундаментальных исследований (грант № 05-07-90217в).

[2] The Text Encoding Initiative [Электронный ресурс]. — Режим доступа: http://www.tei-c.org/, http://www.tei-c.org/release/doc/tei-p5-doc/html/, свободный. — Загл. с экрана.

Итогом выполнения проекта должны стать:

– формат данных (на основе XML-TEI), адаптированный для описания древних текстов, рукописей и их фрагментов; при этом необходимо решить такие проблемы, как представление пересекающихся фрагментов в разметке XML, описание дат в неявном виде (например, первая половина XI века) и другие;

– средства загрузки документов в ИПС «Манускрипт» для последующей работы с ними, а также возможность соединения с уже описанными фрагментами, организованными в иерархии и в некоторых случаях связанными со словарями;

– возможность поиска по текстам, фрагментам, а затем и внутри фрагментов;

– инструменты редактирования текстов (фрагментов, представленных в указанном формате с возможностью сохранения);

– средства выгрузки документов.

Выполненная работа позволит объединить усилия нескольких коллективов для более активного и глубокого исследования рукописных памятников славянской культуры.

Размеченный корпус диалогов как ресурс моделирования диалога: организация и разметка Эстонского корпуса диалогов[1]

О. Герасименко, Т. Хенносте, М. Койт, Р. Кастерпалу,
А. Ряэбис, К. Страндсон, М. Вальдисоо, Э. Вутть
Тартуский университет, Эстония

The Estonian Dialogue Corpus is collected with the aim of developing the dialogue system using the natural language. The spoken dialogues (884 dialogues, 155000 running words) are used to study the rules and norms of the human-human communication; the corpus also includes human-computer dialogues (21, 2500 running words) collected by the Wizard of Oz method used to study the role behaviour of the users and information provider. The presentation considers the means and levels of transcription and annotation dialogues and also the application of the corpus.

Эстонский корпус диалогов создан лингвистами и компьютерными технологами Тартуского университета с целью исследовать речевое взаимодействие в естественных диалогах и моделировать общение между человеком и диалоговой системой, которая должна следовать нормам человеческой коммуникации.

1. Состав корпуса

Основная часть корпуса состоит из естественных устных диалогов (758 справочных телефонных диалогов и 106 непосредственных диалогов, всего 884 диалога объемом в 155000 слов).

Устные диалоги сохраняются в цифровом или оцифрованном формате .wav, а в текстовом виде представлены в транскрипции Джефферсон [Jefferson 2004], которая следует принципам анализа речевого взаимодействия (conversation analysis). Транскрипция фиксирует явления, позволяющие проследить динамическое построение диалога из реплик и интонационных единиц: движение

[1] Работу поддерживает Эстонский научный фонд (грант 5685).