Традиционные и специализированные сервисы исторического корпуса "Манускрипт"

Виктор Аркадьевич Баранов
Ижевский государственный технический университет им. М. Т. Калашникова

0. В докладе будут рассмотрены вопросы технологического и лингвистического проектирования, создания и эксплуатации исторических корпусов. Решения на каждом из этих трех этапов определяются особенностями текстового материала, его объективно ограниченным объемом, а также целями использования корпусов.

1. Средневековые тексты имеют отличающиеся от современных графические системы, неупорядоченную орфографию, компилятивную структуру и сложный состав, существуют в разновременных списках. А количество дошедших до нашего времени авторских и анонимных документов определенного исторического периода не позволяют создать сопоставимый с современными по объему, полноте и степени сбалансированности корпус. Тем не менее, с точки зрения разметки, обеспечивающей формирование подкорпусов и демонстрацию данных, исторический корпус аналогичен современному. В то же время, например, в отличие от современных корпусов, в которых большую роль играет метаразметка текстов, позволяющая формировать на основе разнообразных параметров подкорпусы, в историческом корпусе аннотации документов (рукописей и текстов) выполняют в первую очередь справочную роль.

2. При проектировании исторического корпуса и подготовке лингвистических данных в центре внимания должны быть следующие вопросы:

- степень точности передачи документов;
- правила унификации и нормализации графико-орфографических вариантов,
- разметка и аннотирование как документов, так и их частей;
- выравнивание списков одного произведения;
- автоматизированная лемматизация текстов и нек. другие.

Исторический корпус должен обладать как стандартными средствами поиска, отбора и демонстрации данных, так и специализированными сервисами для работы с материалом. Формы запроса и вывода должны обеспечивать:

- нахождение и упорядоченную визуализацию лингвистических единиц с учетом и без учета их графического варьирования;
- демонстрацию количественных и дистрибутивных данных о единицах выборки;
- визуализацию распределения искомых единиц как в пределах подкорпусов, так и в пределах списков;

- различные способы переходов между данными и разные формы демонстрации контекстов;

- настраиваемые параметры сравнения и сопоставления выборок;

- поиск и вывод выровненных на разных уровнях (от структурно-аналитического до синтаксического и лексического) списков текста и др.

3. Два исторических корпуса проекта "Манускрипт", содержащие транскрипции славянских средневековых письменных памятников периода XI–XV вв. (manuscripts.ru) и все документы М. В. Ломоносова (XVIII в.), вошедшие в Полное собрание сочинений, (lomonosov.pro) демонстрируют реализацию решений лингвистических и технологических задач.

3.1. Запросные формы обеспечивают решение достаточно традиционных для корпусов задач: а) отбор документов для анализа, б) поиск лингвистических единиц на основе маски и/или их значений, в) упорядоченный вывод словоформ и/или лемм одного или нескольких произведений, г) просмотр контекстов в виде стандартных конкордансов или ссылок на фрагменты, д) демонстрацию текстового материала в виде полных документов или их частей, е) визуализацию соответствующих друг другу фрагментов разных списков одного произведения и получение некоторых других видов справочных материалов.

3.2. Специализированными сервисами корпуса являются модули статистики и n-грамм, первый из которых позволяет выявить абсолютную и относительную частоту использования искомых единиц в документе или документах и их распределение на протяжении списков, второй – обнаружить наиболее частотные сочетания словоформ в рукописях.

Оба модуля предоставляют пользователю гибкие параметры запроса. Модуль статистики дает возможность выбрать документы и единицы подсчета (знаки, словоформы, леммы), их количество и расстояние между ними в тексте, указать границы фрагментов подсчета в некоторых единицах (знаках, словоформах, страницах, листах, фрагментах), выбрать точность совпадения с маской и др. Диаграмма распределения найденных единиц позволяет выявить участки документа(ов) с большим или меньшим количеством искомых единиц, сравнить частотность в разных списках одного произведения, просмотреть контексты.

Запросная форма модуля n-грамм позволяет выявлять сочетания из двух, трех и более словоформ, находящихся в контакте или на некотором расстоянии и в любом следовании относительно друг друга, учитывая или игнорируя служебные слова, и выводить сортированный по количеству перечень таких сочетаний, демонстрирующий

регулярную и единичную дистрибуцию словоформы или леммы в пределах документа или группы документов.

* * *

Традиционные и специализированные сервисы любого исторического корпуса должны быть ориентированы на нахождение и единичных фактов, и фактов, количество которых значимо с точки зрения статистики, дистрибуции или распределения в корпусе, интерпретация которых позволила бы подтвердить выводы, полученные традиционными методами, и поставить и решить новые задачи исторического языкознания.