Модуль статистики информационно-аналитической системы "Манускрипт": функции и демонстрация данных

The statistics module of the Manuscript information and analytical system: functionality and data visualization

Баранов В.А., д-р филол. наук, профессор, зав. кафедрой "Лингвистика"
victor.a.baranov@gmail.com,
Дубовцев С.В., программист
movsxw@mail.ru
Ижевский государственный технический университет им. М. Т. Калашникова,
Удмуртский государственный университет, Ижевск, Россия

Ключевые слова: лингвистический корпус, статистика

Summary

The statistics module is intended for the analysis of a corpus of medieval Slavic manuscripts on the manuscripts.ru portal using frequency and distribution of graphic and orthographic facts and of linguistic units in the transcriptions. The prototype can prepare a multi-parameter query, compare multiple copies from a parallel corpus, and find and display data in a text.

1. В [Баранов 2012] впервые было рассказано о создании в рамках информационно-аналитической системы "Манускрипт" (портал проекта http://manuscripts.ru) прототипа второго варианта модуля статистики, предназначенного для анализа рукописей с точки зрения встречаемости и распределения в них лингвистических единиц – частей словоформ, отдельных словоформ и их сочетаний, а также отдельных символов (http://manuscripts.ru/mns/cred.stat/). Особое внимание в публикации было обращено на возможность подготовки гибкого запроса на основе нескольких параметров и на использование модуля для анализа параллельных корпусов, содержащих списки одного текста.

Необходимость создания подобного инструмента для корпуса, содержащего средневековые письменные памятники, понятна: значительная формальная вариативность лингвистических единиц, наличие взаимозаменяемых средств выражения морфологического, словообразовательного, лексического значения, их различная частотность в рукописях заставляют исследователей ставить и решать задачи выявления факторов, влияющих как на активность использования альтернативных лингвистических единиц в различных памятниках, так и на их распределение в пределах одного памятника. Одним из эффективных путей

решения подобного рода задач является использование метода лингвотекстологии, который понимается нами как способ поиска лингвистических (как текстовых, так и системно-языковых) закономерностей на основе учета характеристик рукописей и их текстологически значимых частей[1]. Из работ последнего времени, выполненных в рамках этого метода, назовем анализ А. А. Гиппиусом форм *рѣхъ, рѣша* vs. *рекохъ, рекоша* в Повести временных лет [Гиппиус 2001] и создание Д. А. Добровольским и С. М. Михеевым компьютерной программы выявления словоформ, "распределение которых оказывается тождественным или сходным" в пределах некоторых фрагментов текста [Добровольский-Михеев 2010, 79].

2. Созданный статистический модуль системы "Манускрипт" позволяет выбрать один или несколько документов корпуса, ввести образец анализируемой словоформы, используя маски % (любое количество знаков) и _ (один любой знак), или отдельный символ, указать шаг подсчета и длину шага. В качестве единицы шага используются знаки, словоформы, страницы, листы, фрагменты, на которые размечен документ (например, стихи Евангелий, погодные записи летописей и др.). При необходимости могут быть применены дополнительные параметры запроса: поиск по точному или неточному совпадению с образцом словоформы и поиск с учетом морфологических значений текстовых прецедентов. Эти параметры запросной формы обеспечены имеющимися в системе "Манускрипт" процедурами а) построения условных форм (устраняется диакритика, лигатуры и др.), б) снятия графико-орфографической вариативности (см. правила устранения вариативности на http://manuscripts.ru/mns/slov.list_preobr) и в) автоматического присвоения текстовым прецедентам морфологических значений (выполняется с помощью лемматизатора системы «Манускрипт», http://manuscripts.ru/apex/f?p=104:1). Запросная форма позволяет также задать несколько словоформ, указав расстояние между ними в контексте.

---

[1] "Лингвотекстология – это комплекс взаимосвязанных теоретических и прикладных научно-исследовательских методик систематизации, анализа и интерпретации лингвистических данных корпуса списков одного текста и их фрагментов, а также произведений, содержащих генетически и/или типологически соответствующие друг другу фрагменты, направленный на выявление общих и частных, обусловленных текстологическими, территориально-временными, культурно-историческими и иными факторами закономерностей в функционировании и изменении формальных и содержательных характеристик языковых единиц" [Баранов 2010, 27].

Результатом выполнения запроса является график пошагового абсолютного или относительного количества искомой единицы в пределах одного или нескольких документов.

Предусмотренные возможности изменять параметры запроса призваны обеспечить сопоставимость данных текстов, содержащихся в различных по количеству листов и исполнению рукописях: между объемами текстов и объемами рукописей нет прямой корреляции вследствие разного размера листов, количества строк, величины почерка и других причин. Поэтому запрос по нескольким текстам одного объема с шагом в листах может дать не совпадающие по длине графики, что затрудняет сопоставление участков, расположенных на одном расстоянии от начала рукописи. Для устранения этого неудобства предусмотрена возможность указания шага в словоформах или знаках.

Одним из наиболее показательных режимов работы модуля является сравнение между собой рукописей, содержащих одни и те же тексты (например, Евангелия или Повесть временных лет), выровненные по фрагментам (стихам или погодным записям). Наложенные друг на друга графики позволяют искать соответствующие друг другу диапазоны текстов с одинаковой или близкой частотой использования некоторой единицы и части, которыми списки существенно различаются между собой при использовании альтернативных единиц.

3. Понятно, что анализ и интерпретация полученных графиков, позволяющих увидеть общую картину количественных аналогий или расхождений, требует дополнительной информации для идентификации данных и удобных интерфейсов, обеспечивающих просмотр единиц в контекстах.

Особое значение имеет информация о причинах отсутствия искомых единиц. Так, например, нулевое значение на некотором участке одного из сравниваемых списков параллельного корпуса при наличии искомой единицы в другом может быть вызвано несколькими причинами: пропуском искомой единицы, использованием альтернативной формы, а также отсутствием контекста вследствие иного состава списка или утраты части листов. Желательно уже на этапе просмотра результатов выборки различать эти случаи: об отсутствии некоторых частей текста в сравниваемом списке по сравнению с основным свидетельствуют отрицательные значения частей графика.

Большое количество фрагментов в рукописи (до 1000 и более) и выбор при запросе маленькой величины шага приводит к получению результатов, в которых

достаточно сложно на относительно небольшом экране рассмотреть участки графика. Для решения этой проблемы реализована возможность увеличения размера (растягивания) графика по горизонтали и по вертикали.

Предусмотрено два способа идентификации количественных данных относительно текста: а) указание места данных в рукописи – визуализация адресов единиц, б) демонстрация фрагментов текста, в которых представлены данные. При первом способе каждая точка графика получает подпись, содержащую номера первого и конечного фрагментов и их имена, диапазон адресов шага и количество искомых единиц в нем. При втором данные выводятся в виде таблицы, в которой запись соответствует шагу и содержит, кроме сведений подписи к точке графика, и контексты, в которых используются найденные единицы.

К запросной форме прототипа модуля дан свободный доступ из раздела "Инструменты" портала "Манускрипт: славянское письменное наследие", ее поля снабжены контекстными подсказками.

В настоящее время осуществляется тестирование режимов работы, обсуждается совершенствование компоновки элементов интерфейса и их дизайна. Но уже сейчас модуль предоставляет возможность осуществить анализ распределения лингвистических единиц или отдельных символов в рукописях корпуса, при этом демонстрируя иной порядок, методику и скорость работы. Традиционно при количественном анализе сначала осуществляется отбор, паспортизация и подсчет данных в явных или предполагаемых текстологически значимых фрагментах документа или в соответствующих друг другу частях списков одного произведения, а затем выполняется анализ их распределения и интерпретация результатов. С помощью компьютерных средств выборки и визуализации количественных данных после постановки задачи и определения факторов, предположительно влияющих на распределение материала, создается запрос с соответствующими задаче параметрами, по выборке которого автоматически строится диаграмма, а затем осуществляется анализ распределения (при необходимости – идентификация данных) и предлагается интерпретация выявленных закономерностей и отклонений от них.

Благодарности

Работа выполняется при финансовой поддержке Министерства образования и науки РФ в рамках государственного задания на выполнение работ ФГБОУ ВПО

"Ижевский государственный технический университет" (проект № 8.1613.2011 “Средневековый славянский текст как объект текстологического, лингвистического и структурного моделирования: обеспечение миграции полнотекстовых машиночитаемых исторических документов”).

Литература

Баранов 2010 – Баранов В.А. Корпус средневековых славянских письменных памятников и лингвотекстологические исследования в области исторической морфологии русского языка // Информационные технологии и письменное наследие: материалы междунар. науч. конф. (Уфа, 28–31 октября 2010 г.) / отв. ред. В. А. Баранов. — Уфа; Ижевск: Вагант, 2010. — С. 21-27.

Баранов 2012 – Баранов В. А. Лингвистические, методические и технологические вопросы создания и использования корпуса средневековых славянских текстов // Русистика: язык, культура, перевод: сб. докладов юбилейной междунар. науч. конф. (София, 23-25 ноября 2011 г.). – София : Изток-Запад, 2012. – С. 404-414.

Гиппиус 2001 – Гиппиус А. А. Рекоша дружина Игореви...: к лингвотекстологической стратификации Начальной летописи // Russian Linguistics. – Vol. 25. – 2001. – P. 147–181.

Добровольский-Михеев 2010 – Добровольский Д. А., Михеев С. М. Компьютерные алгоритмы лингво-текстологической стратификации Повести временных лет // Информационные технологии и письменное наследие: материалы междунар. науч. конф. (Уфа, 28–31 октября 2010 г.) / отв. ред. В. А. Баранов. – Уфа; Ижевск: Вагант, 2010. – С. 74-79.