

ПРИМЕНЕНИЕ СОВРЕМЕННЫХ МЕТОДОВ КОРПУСНОЙ ЛИНГВИСТИКИ ПРИ АНАЛИЗЕ ТЕКСТА (НА ПРИМЕРЕ КОРПУСА BFM)

Райскина Валерия Александровна

студент Московского Городского Педагогического Университета, РФ, г. Москва

Дубнякова Оксана Алексеевна

научный руководитель, доц. Московского Городского Педагогического Университета, РФ, г. Москва

Научные исследования в рамках различных лингвистических направлений зачатую имеют объектом своего исследования текст или собрание текстов и подразумевают сначала подбор материала, а затем анализ и обработку больших текстовых объемов с целью выявления некоторых языковых закономерностей. Традиционные лингвистические методы анализа текста позволяют выполнить все вышеперечисленные задачи, но их невысокая эффективность обуславливает все более частое использование методов компьютерного анализа текста, который позволяет сократить работу лингвиста, при этом значительно увеличив объем обрабатываемых данных, а также избежать неточности и ошибки в подсчетах [2, с. 48]. Таким образом, компьютерный анализ текста делает возможным установление языковых закономерностей, основанных не на теоретических, но на эмпирических данных.

В рамках корпусной лингвистики под «**корпусом**» понимают «унифицированный, структурированный и размеченный массив языковых (речевых) данных в электронном виде, предназначенный для определенных филологических и, более широко, гуманитарных изысканий» [4, с. 52]. Преимущество применения корпусного анализа текста заключается в большой степени объективности исследования: такие функции как подбор, разметка, анализ текстов и выявление соответствий выполняются автоматически. Таким образом, задачей исследователя является не анализ материала, а обработка полученных данных, выведение языковых закономерностей и подведение итогов.

В зависимости от используемых корпусов и программ, методы корпусного анализа позволяют решать ряд задач в рамках лингвистических, социологических и исторических дисциплин. Так, корпус позволяет производить **поиск** лексических и грамматических языковых единиц и **отбор** слов, словоформ, грамматических категорий и словосочетаний. При поиске словоформ функция **морфологического дескриптора** позволяет проанализировать искомую единицу исходя из грамматических категорий слова.

Возможен поиск искомой словоформы во всех или в определенных параметрах **контекстах**. Данная функция применяется в практическом аспекте исследования для сбора фактического материала и для составления теоретической базы исследования (словарей, грамматик, справочных пособий). Смежная функция **конкорданса** (по А.С. Герду, — «упорядоченный список словоформ с указанием всех вхождений в заданный массив текстов») позволяют проследить употребление языковой единицы в контексте фиксированной длины, проанализировать совместное употребление искомых словоформ [3, с. 73].

Одной из основных функций лингвистического корпуса является возможности получения лингвостатистических данных о частоте использования тех или иных форм, слов или словосочетаний, то есть определить **частотность** употребления. Данная функция позволяет определить различия семантики синонимов, установить контексты, характерные для синонимичных слов, а также разграничить жанровые и стилистические особенности и

оттенки значений лексических единиц. В социолингвистическом исследовании частотность позволяет определить употребляемость языковой единицы определенной социальной, возрастной, гендерной группой. В историческом аспекте данная функция позволяет проследить этимологию и изменение семантики языковой единицы. Следствием частотного анализа текста является выделение **статистики**. Таким образом, в исследовании с применением корпусного анализа основанием научной гипотезы служат точные и конкретные эмпирические данные.

Каждый текст корпуса сопровождается **паспортизацией или метаразмечкой**, то есть полным библиографическим описанием. В зависимости от вида корпуса метаразмечка может содержать также и данные о стилистической и исторической принадлежности текста. Данная функция может быть использована в научных исследованиях по социолингвистике, лингвокультурологии, психолингвистике, стилистике.

Основной возможностью корпусного анализа является функция **создания собственного подкорпуса** (выборка форм по заданным параметрам поиска). Ввод параметральных данных, характерных для конкретного исследования позволяет составить собственный лингвистический корпус, который может служить как инструментом анализа имеющегося текста, так и собственной эмпирической базой исследования.

На данный момент существует несколько проектов исторических диахронных корпусов французского языка, одним из которых является проект «База средневекового французского языка» или **BFM** (*“La base du français médiéval”*). На февраль 2014 г. данный корпус включает 198 транскрипций текстов IX — конца XV вв., основанных на современных критических изданиях [5, с. 105]. Работа над созданием корпуса BFM началась в 1989 г. под руководством профессора Высшей Нормальной Школы Гуманитарных Наук (*ENS LSH*) К. Маркелло-Низья. В настоящее время над проектом работает группа сотрудников лингвистической лаборатории **ICAR** (*Intérazions, Corpus, Apprentissages, Représentations*). Данный корпус доступен для исследований как в режиме онлайн [9], так и для скачивания в свободном доступе.

Корпус BFM включал несколько проектов: **BFM 2014** — основной корпус, **CORPTEF** — корпус наиболее ранних текстов IX—XII вв. на поздней латыни и их перевод на старофранцузский язык, **BFMMSS** — транскрипции рукописных текстов и манускрипты, **GRAAL** — электронное издание романа XIII в. «В поисках Святого Грааля», разработанное под руководством К. Маркелло-Низья и А.М. Лаврентьева [7, с. 145—148].

Основной базой исследования корпуса BFM являются транскрипции средневековых рукописных текстов различных скрипт и диалектов (*flandre, franco-occitan, normand, liégeois, poitevin, champenois, picard*), жанров (*roman, miracle, serment, hagiographie, bestiaire*) и форм (*vers, prose, mixte*). Отличительна особенность данного корпуса заключается в представлении в проекте GRAAL транскрипций одного текста в нескольких изданиях. **Адаптированное издание** содержит транскрипцию, приближенную к современным нормам языка и предназначено для удобного прочтения текста. **Дипломатическое издание** более точно приближенно к оригинальному тексту, но также содержит некоторые расшифровки и пояснения. В **факсимильном издании** наиболее близко к оригинальному написанию переданы особенности графики средневекового текста. Кроме трех версий, в проект включен современный **перевод текста** и **фототипическое издание** манускрипта. Важной функцией проекта GRAAL является представление версий в многомерном формате (*multi-facettes*), что позволяет выводить на экран сразу несколько изданий, при этом сохраняя параллельное отображение текста построено.

Функционал корпуса помимо транскрипций составляют различные возможности поиска и отбора эмпирических и статистических данных. Данный корпус оснащен **метаразмечкой**, которая включает в себя библиографические данные о каждом произведении и позволяет произвести социолингвистическую диверсификацию текстов.

Базовым принципом корпуса BFM является строгое соответствие критическому изданию. В связи с этим при построении корпуса используется формат **XML**, который позволяет воспроизвести не только сам текст, но и ряд элементов критического аппарата (нумерацию строк, использование различных шрифтов, варианты текста, примечания). Для реализации

поиска и составления подкорпусов проект BFM использует язык компьютерного запроса **CQL** (*Contextual Query Language* или *Common Query Language*). Данный язык представляет собой ряд обозначений-символов, позволяющих производить поиск и отбор словоформ [6, с. 338]. Полный список принятых тэгов и обозначений размещен в свободном доступе на сайте смежного проекта лаборатории ICAR [10] в пособии “Manuel de TXM” [8, с. 116—117].

При исследованиях в корпусе BFM язык CQL позволяет производить несколько видов поиска. **Простой поиск графического написания** позволяет отобрать все случаи употребления конкретной формы слова. Например, запросы вида Graal, “Graal” и [word= “Graal”] используются для поиска конкретной формы слова “Graal”. **Поиск по лемме** возможен вследствие наличия в корпусе BFM морфологического дескриптора и позволяет произвести поиск различных форм слова. Таким образом, запрос вида [frlemma= “beau”] предоставит выборку всех форм имени прилагательного “beau”, а именно “beau”, “bel”, “belle”, “beaux”, “bels”, “belles”. **Поиск по грамматической категории** позволяет произвести выборку всех слов, обладающих искомой грамматической категорией. При этом необходимо использование принятых тэгов [8, с. 160]. Например, [frpos= “ADV”] — подборка всех наречий. **Поиск вариантов написания** диакритических знаков или регистра букв позволяет расширить поиск однокоренных слов. Так, запрос вида “franc.*”%cd позволяет отобрать все примеры с корнем “franc” включая варианты корня “franç” и написание с прописной буквы. Кроме того, в корпусе возможен **поиск пунктуации, части слова** или **вариативный поиск**, для чего используются комбинации символов CQL.

Кроме поиска в корпусе BFM представлена функция создание **индексов**, которые показывают частоту использования искомой языковой единицы.

Также в корпусе возможно использовать режимы **конкорданса** и **контекста**, которые показывают употребление искомой языковой единицы в полном и наглядном словарном окружении. Корпус также оснащен **лексикой-словарем** использованных слов в тексте.

Наряду с вышеперечисленными функциями, наиболее важная возможность для исследований в компьютерном корпусе — это создание, настройка и анализ собственного **подкорпуса текстов**. Применение комбинированных поисковых запросов и внетекстовых функций корпуса позволяют производить подбор примеров по узким параметральным данным. В данной статье мы продемонстрируем применение возможностей корпуса GRAAL в исследовании по исторической грамматике, а именно при анализе использования заглавных букв в средневековом тексте. Мы воспользовались функциями корпуса для того, чтобы отобрать все употребления заглавной буквы после прямой речи в дипломатическом издании романа «В поисках Святого Грааля». Для данного поиска применяется запрос следующего вида:

</q> [dipl=“[A-Z].*”]

Формула данного вида приводит к подборке из 294 вхождений по заданному запросу. Так как задачи данного поиска требуют синтаксически распространенные результаты, поиск наиболее удобно производить в режимах контекста и конкорданса. В окне **контекста** приведены примеры вхождений с указанием страницы в электронном издании, а также с выделенными красным цветом первыми словами после окончания прямой речи. Кроме того, режим контекста позволяет задать размер правого и левого контекстов: таким образом, можно восстановить и начало прямой речи, получая полное и исчерпывающее словарное окружение. Также формат XML позволяет перейти по каждому примеру к выбранному электронному изданию в параллельном окне, что создает оптимальное расположение страниц для анализа искомой языковой единицы как в поисковом списке, так и интегрировано в тексте.

В отличие от режима контекста, в **конкордансе** подборка вхождений форматируется по первому слову запроса, то есть по первому слову после окончания прямой речи для конкретного поиска. При этом, в данном режиме также указывается страницы электронного издания и представлена возможность перехода к изданию по каждому примеру.

Отметим, что подборка по данному запросу не является конечным действием, позволяющим

сделать выводы по поставленной цели, а именно, определить закономерность использования прописных и строчных букв после прямой речи. Дело в том, что запрос производится в дипломатической версии, текст которой с точки зрения оформления предложения адаптирован к современным нормам. Таким образом, все буквы после прямой речи — прописные. Тем не менее, параллельное воспроизведение дипломатической версии и фототипического издания позволяет отследить использование регистра букв в самом манускрипте. В итоге, совместное применение таких функций корпуса GRAAL, как поиск, контекст или конкорданс и параллельное отображение электронных версий, позволяет отследить написание прописных и строчных букв в средневековом тексте, при этом исключив трудоемкий процесс прочтения, расшифровки и анализа старофранцузского рукописного текста. Таким образом, корпусный анализ данного текста становится доступным для исследователей, не владеющих навыками прочтения и понимания старофранцузского письма, и вместе с тем позволяет пропустить этап ручного отбора эмпирической базы исследования.

Методы корпусного анализа текста позволяют значительно повысить эффективность исследовательской деятельности ввиду автоматизированной системы отбора, обработки и вывода результатов. Преимуществом использования методов корпусного анализа также является возможность производить подсчет и составлять статистики, подтверждающие или опровергающие гипотезы исследования [1, с. 135—137]. Таким образом, производя корпусный анализ текста, становится возможным обосновать исследование не на гипотезах и предположениях, но на точных эмпирических данных.

Корпус *“La Base du Français Médiéval”* представляет собой автоматизированную программу хранения и анализа текстов эпохи Средневековья. Благодаря обширной базе текстов данный корпус может быть использован в исследованиях по истории литературы, истории языка и стилистике. Возможности поиска при условии владения языком SQL позволяют проводить исследования в области исторической грамматики, орфографии, пунктуации, лексикологии и семантики. Представленные в корпусе средневековые манускрипты могут выступать объектами исследований по палеографии и каллиграфии, а также по книговедению, археографии, филологии, дипломатике, текстологии и источниковедению.

Список литературы:

1. Баранов А.Н. Корпусная лингвистика // Баранов А.Н. Введение в прикладную лингвистику. — М., 2001. — С. 112—137.
2. Богданова С.Ю. Возможности корпусной методологии в решении лингвистических задач / С.Ю. Богданова. — Иркутск.: Вестник ИГЛУ. — № 2 (18). — 2012. — С. 47—50.
3. Герд А.С. Автоматизация в лексикографии и словари-конкордансы / А.С. Герд // Филологические науки. 1981. — № 1. — С. 72—78.
4. Захаров В.П. Поисковые системы Интернета как инструмент лингвистических исследований / В.П. Захаров // Русский язык в Интернете: Сб. статей. — Казань: Отечество, 2003. — С. 48—59.
5. Лаврентьев А.М. Базы данных и корпуса текстов средневекового французского языка: подходы, проекты, технологии / А.М. Лаврентьев // Современные информационные технологии и письменное наследие: от древних рукописей к электронным текстам / Материалы междунар. науч. конф. (Ижевск, 13—17 июля 2006 г.) / отв. ред. В.А. Баранов. — Ижевск: Изд-во ИжГТУ, 2006. — С. 104—107.
6. Роберт У. Себеста. Основные концепции языков программирования (Concepts of Programming Languages) / Пер. с англ. — 5-е изд. — М.: Вильямс, 2001. — 672 с.
7. Guillot C., Lavrentiev A., Marchello-Nizia Ch. La Base de Français Médiéval (BFM): états et perspectives / Kunstman P., Stein A. (éds), Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23—26 février 2006. — Stuttgart: Franz Steiner Verlag. — P. 143—152.
8. Manuel de TXM Version 0.7. — Lyon.: ENS de Lyon, 2014. — 166 p.
9. Base de Français Médiéval — [Электронный ресурс] — Режим доступа. — URL: <http://txm.bfm-corpus.org> (Дата обращения 29.04.2015).
10. Textométrie: *Fédération des recherches* et développements en *textométrie* autour de la création d'une plateforme logicielle ouverte — [Электронный ресурс] — Режим доступа. —

URL: <http://textometrie.ens-lyon.fr/> (дата обращения 16.04.2015).