А.В. Зубов (Минск, МГЛУ)

ИСПОЛЬЗОВАНИЕ КОРПУСА ТЕКСТОВ ДЛЯ ИЗУЧЕНИЯ ПОЭТИЧЕСКИХ ТЕКСТОВ

В последние годы компьютеры все больше используются для решения различных лингвистических задач. На этом пути ярко выявились те проблемы, которые компьютер не может решить, имея в памяти отдельные тексты или словари. Все чаще взоры исследователей обращаются к корпусам текстов, содержащим в себе множество текстов различного типа и определенную информацию о каждом тексте, предложении и слове.

Корпус текстов – "это электронное собрание текстов, размеченное таким образом, чтобы в нем можно было быстро найти слова и конструкции с заданными грамматическими и другими интересными лингвисту свойствами" [7, с. 7]. Расшифровывая более подробно это понятие исследователи отмечают, что некоторое множество текстов можно считать корпусом текстов, если они удовлетворяют следующим условиям [8]:

1. Расположение множества текстов на магнитном носителе.
2. Наличие определенной процедуры отбора текстов в корпус текстов.
3. Единая методика представления сведений о текстах и их единицах на магнитном носителе.
4. Конечный размер корпуса текстов.
5. Репрезентативность множества текстов, входящих в корпус.

Требует объяснения и еще одно понятие, фигурирующее в приведенном выше определении – "размеченное". Предполагается, что каждое словоупотребление корпуса текстов должно иметь достаточно полный набор лексико-грамматических, синтаксических, семантических структурных и других признаков. Такие тексты называются размеченными, аннотированными или тэггированными.

В идеале, хорошие корпуса текстов должны содержать тексты различных временных периодов, различных функциональных стилей и жанров. Корпуса текстов стали активно создаваться с начала 50-х годов прошлого века. Созданы Британский корпус английского языка объемом 600 млн. словоупотреблений, американский национальный корпус, венгерский, итальянский, хорватский, чешский, японский корпуса (каждый объемом в 100 млн. словоупотреблений), русский национальный корпус объемом в 150 млн. словоупотреблений [4, с. 11]. Сейчас в МГЛУ совместно со специалистами Института языка и литературы НАН РБ создается корпус текстов белорусского языка объемом в 1 млн. словоупотреблений.

Допустим, что у исследователя в корпусе текстов имеется определенное, достаточно репрезентативное множество поэтических текстов разных временных интервалов и принадлежащих перу различных авторов. Какую же информацию можно извлечь из такого корпуса, позволяющую решать проблемы, относящиеся к поэтическому литературоведению? Ответ на этот вопрос предполагает, что каждое словоупотребление корпуса текстов имеет,

как минимум, набор лексико-грамматических (класс слова, род, число, падеж и т.п.) и структурных (число слогов в словоупотреблении и место ударного слога в словоупотреблении) признаков.

Первый ряд данных, которые могут быть получены из такого корпуса, представляет собой чисто статистические данные об особенностях текстов произведений отдельных авторов (в разные периоды их творчества), группы авторов (или авторов определенной литературной школы). Такие данные – неоценимый материал для различных стилистических и сопоставительных исследований.

К числу таких данных можно отнести следующие: 1) алфавитный словарь поэта; 2) частотный словарь поэта; 3) словарь рифм поэта; 4) обратный частотный словарь по произведениям поэта; 5) частотный список употреблений частей речи в произведениях поэта; 6) частота употребления таких грамматических категорий как род, число, падеж, время, вид и т.п.; 7) частотный словарь словоформ поэта по количеству слогов в словоформах; 8) словарь словоформ поэта по месту ударного слога в словоформах; 9) частота употребления структуры строк на уровне частей речи; 10) частота употребления структуры предложений на уровне частей речи; 11) частота употреблений структуры строф на уровне частей речи; 12) частоты употребления ритмических структур строк; 13) частоты употребления строф с различной ритмической структурой.

Имея в компьютерной памяти подобную информацию, можно проводить достаточно содержательные многоаспектные исследования.

Простейшая для компьютера задача – это изучение эволюции стилистических характеристик отдельных авторов в разные периоды их жизни или же изучение стилистических особенностей группы авторов с опорой на указанные выше статистические параметры.

Использование этих характеристик позволяет проводить и более сложные исследования. Так, с опорой на такие данные выполнено исследование по изучению закономерностей эволюции русской рифмы за период с 1851 года по 1980 год [6]. Исследование проводилось на 7860 рифмах, взятых из различных поэтических произведений за указанный период, опубликованных в журналах. Подобное же исследование [5] имело целью изучение эволюции рифм ("точная" – "неточная") в стихах одного периода (1808–1835) у А. Пушкина и П. Вяземского.

Интересное направление в исследовании поэтических текстов представлено в работе [1]. В ней путем изучения употребительности словесных (на уровне ударных и безударных слогов) и ритмико-синтаксических формул стихов более 20 поэтов (от А. Пушкина и М. Лермонтова до А. Вознесенского) путем выделения их ведущих типов и их вариаций определены формальные стили отдельных поэтов и эпох.

В работе [3], изучая ритмику "Хотинской оды" М.В. Ломоносова, первой русской оды, написанной четырехстопным ямбом, и оды немецкого поэта Й.К. Гюнтера "На победу принца Евгения". Е.В. Казарцев доказал жанровую и тематическую близость сочинений этих авторов.

Более сложная методика, использующая отмеченные выше статистические параметры (а именно, использование частоты употребления личных местоимений, частоты употребления грамматической и ритмической структуры строк, типа рифм и структуры строф на уровне входящих в них типов предложений), позволила получить интересные данные об одах XVIII века (М. Ломоносова, А. Сумарокова) [2].

Интересное исследование об особенностях употребления в поэзии А.Блока глагола "проходить (пройти)" представлено в работе [9]. Изучая частоту употреблений этого слова в различных контекстных окружениях, М.Д. Якубовская приходит к выводу, что "в рассмотренных контекстах глагол *пройти* не равен самому себе: его семантическое содержание или у́же, или шире, чем в естественном языке" [9, с. 281] и что это слово у А.Блока несет образную функцию или приобретает образное значение, которого вне блоковского контекста не имеет.

С опорой на корпус текста можно провести и много других интересных исследований.

ЛИТЕРАТУРА

1. Гаспаров, М.Л. Ритмико-синтаксическая формульность в русском 4-х стопном ямбе / М.Л. Гаспаров // Проблемы структурной лингвистики. 1983. – М.: Наука, 1986. – С. 181–199.

2. Иванов, Вяч. Вс. Из наблюдений над одой XVIII века / Вяч. Вс. Иванов // Лингвистика и поэтика. – М.: Наука, 1979. – С. 174–187.

3. Казарцев, Е.В. Ритмика первой духовной оды М.В. Ломоносова в контексте проблемы генезиза русской силлаботоники / Е.В. Казарцев // Формальные методы в лингвистической поэтике. Сборник научных трудов. – Санкт-Петербург: СПбГУ, 2001, – С. 37–48.

4. Машинная обработка текстов и квантитативные методы в современной германистике. – Самара, 2005. – 31 с.

5. Минералов, Ю.И. О путях эволюции русской рифмы / Ю.И. Минералов // Studia metrica et poetica: УЧ. ЗАП. Tart. гос. ун-та. Вып. 420. – Tartu. – часть II. – 1977. – С. 40–58.

6. Шепелева, С.Н. Некоторые закономерности эволюции русской рифмы (1851-1980) / С.Н. Шепелева // Проблемы структурной лингвистики. 1983. – М.: – Наука, – С. 209–215.

7. Плунгян, В.А. Зачем нужен национальный корпус русского языка. Неформальное введение / В.А. Плунгян // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. – М.: ИНДРИК, 2005. – С. 6–21.

8. Рыков, В.В. Прагматический ориентированный корпус текстов / В.В. Рыков // Компьютерная лингвистика и интеллектуальные технологии: труды межд. конф. "Диалог-99" – М.: Наука, 1993. – С. 211–220.

9. Якубовская, М.Д. Об одном глаголе в поэзии А. Блока / М.Д. Якубовская // Лингвистика и поэтика. – М.: Наука, 1979. – С. 277–281.