ВОЗМОЖНОСТИ ПРОГРАММЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ «МУЛЬТИПОИСК» В ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ

А.В. Жудро
МГПУ им. И.П. Шамякина (г. Мозырь)

Использование компьютерной техники для автоматической обработки текстов сегодня является необходимостью для решения задач, которые ставит перед нами современная наука. Компьютерная неподготовленность филологов и неоснащенность школ и гуманитарных вузов компьютерами ранее задерживали применение этой техники. Сегодня благодаря повсеместному использованию компьютерной техники лингвистические исследования выходят на новый уровень. Выделяют следующие основные направления применения вычислительной техники в лингвистических исследованиях и при обучении иностранным языкам: машинный перевод, автоматизация лексикографических работ, отдельные виды автоматизации собственно лингвистических исследований, автоматический поиск библиографической информации.

Таким образом, компьютер может выполнять многие сложные работы, на которые лингвистам приходится затрачивать очень много времени. Автоматический поиск библиографической информации представляется наиболее перспективным направлением применения ЭВМ в связи с лингвистикой, без него современному ученому все труднее и труднее справляться с обрушивающимся на него потоком информации.

Проведение значительного количества лингвистических исследований предполагает поиск материала в аутентичных текстах рассматриваемого языка. При использовании функции поиска в распространенных текстовых редакторах типа Microsoft Word, Word Pad, Блокнот возникает следующая проблема: поиск по запросу может выдавать сотни и даже тысячи результатов, которые физически невозможно просмотреть в ограниченное время. При этом отсутствует возможность оформления результатов поиска

в один файл. Для решения этой проблемы разрабатываются системы, позволяющие группировать результаты поиска и автоматически разбивать их на подмножества (кластеризация результатов поиска), либо системы, выдающие наиболее устойчивые словосочетания (коллокации) со статистической оценкой их значимости. Разработанная программа «Мультипоиск» призвана решить подобные проблемы.

Программа «Мультипоиск» предназначена для автоматического поиска языковых данных (словоформ) в электронных документах. Программа имеет ряд преимуществ перед иными аналогами. «Мультипоиск» не требует открытия текстовых документов, в которых ведется поиск. Программа поддерживает работу с наиболее распространенными видами текстовых документов (расширения txt, doc, rtf). Программа не ограничивает количество текстовых файлов, в которых ведется поиск. Программа дает возможность получать статистические данные по словоформам (например, сколько раз в документах встречается та или иная словоформа). Имеется возможность просмотра контекста того или иного слова в отдельном окне программы. Также имеется возможность исключать из списка запроса ненужные словоформы. Все перечисленные преимущества позволяют сэкономить время, а также грамотно и удобно организовать работу над исследованием.

Апробация программы проводилась в ходе ряда исследований на кафедре английского языка и методики преподавания иностранных языков УО МГПУ им. И.П. Шамякина: «Семантика предложений, организованных глаголами созидания и разрушения», «Эвфемизация лексико-семантического поля «Профессия» на материале английского языка», «Особенности фразеологических единиц с компонентами частей тела человека на материале английского языка» и др. На поиск необходимой информации вместо месяцев тратились минуты. Особую ценность программа представляет для исследований в области семантики.

Дальнейшая работа с программой предполагает создание корпуса электронных лингвистических текстов на базе кафедры, что значительно расширит тематику разрабатываемых проблем для курсовых проектов студентов. Это позволит включить в ряд разрабатываемых студентами курсовых работ также тематику корпусной лингвистики – раздела языкознания, занимающегося разработкой, созданием и использованием текстовых (лингвистических) корпусов. Термин введён в употребление в 60-х годах XX века в связи с развитием практики создания корпусов, которому, начиная с 80-х годов, способствовало совершенствование вычислительной техники.

Под лингвистическим корпусом подразумевается совокупность текстов, собранных в соответствии с определёнными принципами, размеченных по определённому стандарту и обеспеченных специализированной поисковой

системой (в нашем случае – программа «Мультипоиск»). Иногда корпусом называют просто любое собрание текстов, объединённых каким-то общим признаком (языком, жанром, автором, периодом создания).

Целесообразность создания текстовых корпусов объясняется: представлением лингвистических данных в реальном контексте; достаточно большой представительностью данных (при большом объёме корпуса); возможностью многократного использования единожды созданного корпуса для решения различных лингвистических задач.