Шкурская Н.М.
Белорусский государственный университет, Минск

КОМПЬЮТЕРНЫЕ ПРОГРАММЫ ОБРАБОТКИ КОРПУСОВ ТЕКСТОВ

Как известно, любое лингвистическое исследование опирается на анализ языкового материала. Чем больше объем материала, тем выше достоверность результатов и тем шире область применения этих результатов. Для многих областей лингвистики сбор новых языковых фактов считается основной задачей лингвистического описания. До недавнего времени письменные тексты подвергались лишь ручной обработке. Появление новых информационных технологий и технических средств значительно облегчило сбор и обработку языковых данных. Создаются специальные корпуса текстов, которые в дальнейшем служат как источником для создания новых словарей, так и богатым материалом для лингвистических исследований.

Технология работы с корпусом текстов предполагает наличие обязательной поддержки корпуса комплексом программ по обработке данных, обеспечивающих функции составления конкордансов, статистической инвентаризации, автоматической словарной обработки (составление полных и частичных словников по различным критериям – частоте, алфавиту и пр.)

А.Н. Баранов выделяет две основные стратегии, по которым строятся имеющиеся компьютерные программы, ориентированные на обработку корпуса текстов.

1. В соответствии с первой стратегией программа порождает для текста комплекс указателей, например, указатель словоформ, в котором для каждой словоформы указывается адрес в тексте, т. е. программа оперирует не столько текстом, сколько указателями к нему. Типичные примеры программ такого типа – программные пакеты UNILEX (Машинный фонд русского языка), американские программы ETC и WORD CRUNCHER. В Великобритании используется аналогичный по функции пакет OCP (Oxford Concordance Program), а в Германии – программа TEXTPACK.
2. При второй стратегии для поиска необходимых контекстов программа каждый раз последовательно просматривает текст, маркируя те фрагменты, которые удовлетворяют поисковому заданию [1, с. 117].

Обе стратегии имеют свои достоинства и недостатки. К достоинствам первой стратегии А.Н. Баранов относит тот факт, что программы типа UNILEX при составлении конкордансов работают не с текстами как таковыми, а с указателями к ним [1, с. 118]. Однако разбиение текстов на модули и составление больших указателей требует значительного рабочего времени и наличия больших ресурсов памяти.

При второй стратегии указатели, которые создаются к корпусу текстов, являются временными и уничтожаются по мере выполнения алгоритма. Они не требуют предварительной обработки корпуса, членения текстов на отдельные модули и т. п. Но программы такого рода должны использовать очень

продуктивные подпрограммы обработки текста, поскольку каждый поиск предполагает сплошной просмотр корпуса. Основной недостаток второй стратегии заключается в том, что значительное увеличение массива текстов в корпусе существенно замедляет работу программы.

Особого программного обеспечения требуют корпусы параллельных текста. Программа MULTICONCORD позволяет строить конкордансы и устанавливать соответствия между фрагментами оригинального текста и его переводами на другие языки. В настоящее время MULTICONCORD работает с корпусом из шести языков – английский, немецкий, французский, греческий, итальянский и датский (текст на языке-источнике и пять текстов на языках перевода). Для разных текстов целевые языки и языки-источники варьируются. Программа дает возможность производить поиск по разным языкам, словам, словоформам и словосочетаниям. Результаты поиска могут сортироваться по объему, алфавиту, по произведениям, авторам и т. д.

Основная проблема в построении корпусов параллельных текстов и разработке пакетов программ для их обработки заключается в уста-новлении соответствий между оригинальными текстами и переводами, так как при переводе устанавливаются лишь лексические соответствия, а в случае свободного перевода индексируются целые фрагменты предложений или текстов. Здесь уместнее говорить об использовании технологий систем машинного перевода с языком-посредником или универсальным языком, но данный подход пока остается только теорией.

Литература

1. Баранов, А.Н. Введение в прикладную лингвистику: Учеб. пособие / А.Н. Баранов. – М.: Эдиториал УРСС, 2001.
2. Захаров, В.П. Корпусная лингвистика: Учеб.-метод. пособие / В.П. Захаров. – СПб.: СПбГУ, 2005.