А. Е. Жданович

ИСПОЛЬЗОВАНИЕ КОРПУСА ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ ДЛЯ АВТОМАТИЧЕСКОГО ПЕРЕВОДА УСТОЙЧИВЫХ СЛОВОСОЧЕТАНИЙ

Корпус текстов является основным понятием корпусной лингвистики, разрабатывающей методологию создания и использования электронных лингвистических корпусов. Корпус текстов представляет собой, по мнению Т. МакЭнери и Э. Вилсона, «собрание языковых фрагментов, отобранных в соответствии с четкими языковыми критериями для использования в качестве модели языка». Как отмечает А. Н. Баранов, целесообразность создания и использования корпусов текстов определяется двумя основными предпосылками:

1) данные разного типа находятся в корпусе в своей естественной контекстной форме, что предоставляет возможность их всестороннего и объективного изучения;

2) достаточно большой, или репрезентативный объем корпуса гарантирует типичность данных.

Технологии корпусной лингвистики базируются на достоверных данных о фонетической, морфологической, синтаксической и семантической структуре языка и речи, которые могут быть получены только из достаточно большого массива текстов. Для исследования процесса перевода в корпусной лингвистике используется такие виды корпусов, как параллельный корпус текстов и сопоставительный корпус текстов. Параллельный корпус текстов состоит из набора текстов на языке оригинала и их соответствующих текстов на языке перевода. Сопоставительный корпус включает в себя два разных набора текстов на одном языке: первый состоит из оригинальных текстов, второй – из переведенных текстов на данный язык. Таким образом, сопоставительный корпус – это моноязычный корпус, состоящий из двух подкорпусов на одном языке. С. Гренжер говорит о том, что, несмотря на явное преимущество сопоставительного корпуса, представленное двумя сопоставляемыми одноязычными подкорпусами, такой вид корпуса имеет и недостаток, который заключается в том, что сложно выявить параметры сопоставления. Некоторые виды текстов имеют культурные особенности и не обладают точными соответствиями при переводе. Отличительной чертой параллельного корпуса является разметка, позволяющая каждому предложению исходного языка сопоставить предложение языка перевода. Такой корпус позволяет установить переводные соответствия между искомыми элементами корпуса, выявить их частотность и представить список контекстов, в которых употребляются исследуемые единицы.

Исследователь Г. Г. Белоногов в 1975 году впервые высказал идею использовать для перевода текстов определенные фрагменты уже переведенных текстов. По мнению ученого, суть данного подхода заключается в том, что в памяти компьютера накапливаются корпусы исходных текстов и их переводов, выровненных между собой на различных уровнях (абзацев, предложений,

словосочетаний). В процессе перевода такая система пытается отыскать переводимый сегмент в массиве исходных параллельных текстов. Если он найден в исходном массиве текстов-оригиналов, то система выбирает его перевод в массиве переведенных текстов (Д. А. Жуков «Мы – переводчики»). Японский ученый М. Нагао подробнее сформулировал использование для перевода уже сохраненных корпусов переведенных текстов. Свой подход к автоматическому переводу он назвал «машинный перевод на основе заложенных примеров или машинный перевод по аналоговому принципу» и обосновал его следующим образом: «Человек не переводит простое предложение с помощью глубокого лингвистического анализа. Скорее, человек при переводе сперва членит содержание предложения на некоторые отдельные фразы <…>, затем переводит эти фразы на другой язык и, наконец, надлежащим образом складывает эти фрагменты перевода в одно длинное предложение. Такой перевод каждого фрагмента предложения будет переводиться по принципу аналогового перевода, со ссылкой на надлежащие примеры».

Ученый Д. О. Добровольский в своей работе «Корпус параллельных текстов как инструмент анализа литературного перевода» говорит о необходимости выявления структур, «которые по своей природе требуют от переводчика нетривиальных решений». К таким структурам автор предлагает отнести нестандартное употребление лексических единиц, нерегулярные синтаксические конструкции, излюбленные слова автора, несвободные словосочетания, идиомы и конвенциальные метафоры, авторские метафоры, культурно-специфичные элементы (названия, имена, титулы и т.п.). Автоматический перевод таких структур вызывает определенные трудности. Этим можно объяснить тот факт, что широко известные промышленные системы машинного перевода до сих пор не использовались для перевода текстов с несвободными словосочетаниями. В то же время попытки применить для автоматического перевода устойчивых словосочетаний корпусы параллельных текстов позволили получить неплохие результаты. Первым шагом в этом направлении явилось исследование Д. В. Степановой, основная идея которого основывается на следующих предположениях: 1) в каждом несвободном словосочетании предложения текста на ИЯ можно выделить его опорное («ядерное», «узловое») слово, являющееся основой словосочетания; 2) в рамках взятого из параллельного корпуса текстов перевода исходного предложения можно выделить несвободное словосочетание, опорным словом которого является один из переводных эквивалентов опорного слова исходного несвободного словосочетания или наиболее значимое с лингвистической точки зрения слово этого словосочетания; 3) в базе данных системы перевода для текстов достаточно узкой предметной области, заданы, во-первых, структурные схемы (минимально на уровне классов слов) исходного и переведенного несвободного словосочетания, во-вторых, выявленные на предварительных этапах лингвистических исследований текстовые признаки, определяющие границы несвободных словосочетаний в рамках исходного предложения и его перевода (Д. В. Степанова «Построение принципиального алгоритма автоматического выделения английских терминологических словосочетаний и их перевода на русский язык с помощью корпуса параллельных текстов»).

По мнению А. В. Зубова, описанный выше подход может быть усовершенствован и использован для автоматического перевода идиоматических выражений (А. В. Зубов «Подходы к переводу идиоматических выражений»). При этом ученый подчеркивает, что основная трудность заключается в том, что перевод идиоматического выражения не всегда будет содержать опорное слово, являющееся одним из переводных эквивалентов опорного слова исходного идиоматического выражения, как это было при переводе терминологических словосочетаний. Основная трудность при переводе идиоматических (фразеологических) единиц с помощью корпуса параллельных текстов заключается в правильном выделении слов, связывающих исходную единицу и ее перевод на другой язык.

В последние десятилетия на помощь переводчикам приходят разнообразные системы автоматического перевода текста с одного языка на другой, однако не все компьютерные системы могут осуществить адекватный перевод текстов с фразеологическими единицами. Автоматический перевод устойчивых словосочетаний возможен в рамках технологий, основанных на больших двуязычных (многоязычных) корпусах параллельных текстов.