

ИСПОЛЬЗОВАНИЕ КОРПУСОВ ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ В ОБУЧЕНИИ СТУДЕНТОВ-ЛИНГВИСТОВ

С 2005 года на гуманитарном факультете БГУ проводится обучение студентов по специализациям «Компьютерная лингвистика» и «Компьютерное обучение языкам» на отделении «Современные иностранные языки». Основными целями обучения являются накопление знаний в области автоматической обработки естественно-языковых текстов, а также выработка умений и навыков использования лингвистических компьютерных программ в профессиональной деятельности.

На занятиях по дисциплинам «компьютерная лингвистика» и «автоматическая обработка естественного языка» важное место занимает изучение корпусной лингвистики, поскольку языковой корпус – это «очень эффективный и полезный инструмент, которым могут пользоваться далеко не только узкие специалисты» [1, с.6], а корпусная лингвистика – молодая отрасль языкознания. Поэтому в нашей стране остро стоит задача создания как одноязычных корпусов для государственных языков (белорусского и русского), так и корпусов параллельных текстов, в которые входят тексты на одном (или на обоих) из этих языков.

Корпус параллельных текстов (КПТ, параллельный корпус) – это «подмножество текстов на языке-источнике и одно или несколько подмножеств текстов, которые являются переводами текстов языка-источника на языки-цели» [2, с. 65]. В зависимости от целей создания параллельный корпус может включать в себя как тексты одного стиля речи или даже одного автора, так и тексты самых разных стилей и жанров. Проецируя указанное свойство на тему данной статьи, можем утверждать, что при обучении студентов необходимо использовать КПТ, включающий максимально разнообразные тексты, то есть он должен обладать свойством репрезентативности в любой проблемной области. Ведь невозможно предсказать, в какой сфере реализует себя будущий выпускник и с какой стороны он может использовать КПТ в будущем. С этой стороны, например, параллельный корпус, представленный на сайте Национального корпуса русского языка [3], не полностью соответствует целям обучения, так как в нем представлены в основном тексты художественного стиля.

В изучении параллельных корпусов в рамках вышеназванных специализаций можно выделить четыре проблемные области:

1. Усвоение основных принципов составления и поддержки параллельных корпусов текстов.
2. Использование параллельных корпусов в профессиональной деятельности переводчика.
3. Использование параллельных КПТ при обучении иностранному языку.

4. Проведение лингвистических исследований при помощи параллельных корпусов.

Рассмотрим подробнее каждую из проблемных областей.

1. На лекциях и семинарских занятиях по компьютерной лингвистике студенты узнают основные понятия корпусной лингвистики, принципы создания и ведения корпусов текстов, приобретают навыки морфологической и синтаксической разметки, а также способы выравнивания текстов в параллельных корпусах. Обратим внимание, что создание корпусов следует базировать прежде всего на лингвистических механизмах, а не на статистических (как это часто происходит, если к работам над естественно-языковыми текстами лингвисты не привлекаются или привлекаются лишь изредка). Так, при выравнивании параллельных текстов «наиболее широко применяемые статистические методы, которые не требуют развитой словарной базы и могут использоваться для редких языков, часто дают ошибочные результаты выравнивания, требуя в последующем дорогостоящей ручной проверки и исправления» [4].

Важно сформировать у обучающихся не только взгляд на КПТ со стороны пользователя, но в первую очередь – как лингвиста-разработчика. Таким образом, на протяжении обучения у будущих специалистов формируется профессиональная компетенция в одном из направлений автоматической обработки текстов на естественном языке. Кроме того, студенты специализации «Компьютерная лингвистика» имеют возможность детально изучить одну из проблем при написании курсовой и дипломной работы. За 2010-2014 гг. было успешно защищено несколько дипломных работ, посвященных КПТ: «Использование параллельных корпусов текстов при переводе», «Использование параллельных корпусов текстов в системах автоматической обработки естественного языка» и др.

2. В данном аспекте рассматривается возможность использования параллельного корпуса в выявлении и анализе переводческих решений. Как известно, двуязычные словари не всегда полно отражают все возможные эквиваленты слов, а тем более словосочетаний. КПТ позволяет быстро найти и использовать в переводческой практике наиболее точный и адекватный вариант перевода определенной лексики или конструкции. Переводы в параллельном корпусе, как правило, сделаны профессиональными переводчиками, поэтому студенты при работе с ним усваивают наилучшие образцы перевода.

Кроме того, параллельные корпуса являются основой систем машинного перевода, основанного на примерах. В отличие от КПТ, представленного в табличной форме, система машинного перевода позволяет получить готовый перевод нового текста, который при необходимости редактируется. Такой подход значительно повышает производительность труда переводчика, однако для приемлемого автоматического перевода необходимо накопить огромное количество параллельных текстов. Поэтому важной задачей также становится автоматизация пополнения КПТ, создания автоматизированного рабочего места (АРМ) лингвиста и подготовка специалистов в данной отрасли.

3. Поскольку выпускники названных специализаций получают в том числе квалификацию преподавателя иностранного языка, проводится обучение студентов использованию КПТ в преподавательской практике. Корпус является источником огромного количества примеров для составления упражнений различного рода. Например, при изучении грамматики «возможность поиска по грамматическим критериям позволяет автоматически (и с несравнимо большей скоростью) получить примеры, из которых уже вручную можно выбрать наиболее подходящие... Корпус дает возможность отсортировать источник примеров» [5, с.311]. Ниже приведен пример упражнения, составленного одним из студентов специализации «компьютерная лингвистика» при помощи параллельного англо-русского корпуса, представленного на сайте Национального корпуса русского языка [3]:

Переведите предложения на русский язык, обращая внимание на трансформации грамматической основы при переводе. Сравните ваш вариант и вариант профессионального переводчика, используя параллельный корпус.

1) *He fled with a sinister reputation to South America, where he was said to have died unmarried. [Arthur Conan Doyle. The Hound of the Baskervilles (1902)]*

2) *I hadn't known his full name, and believed him to be what he was said to be, a new gardener working for Buildings and Grounds. [Kurt Vonnegut. Hocus Pocus (1990)]*

3) *He was believed to have a bedroom at the back. [John Galsworthy. The Man of Property (1906)]*

4) *Would he go back upon it? It was thought to be unlikely. All waited with interest. [John Galsworthy. The Man of Property (1906)]*

5) *Above all, he was known to be liberal and hospitable, and believed to be good-natured. [Walter Scott. Ivanhoe (1819)]*

4. Как отмечал А.Н. Баранов, «любое лингвистическое исследование в той или иной мере опирается на анализ языкового материала, языковых данных» [2, с. 63]. Параллельный корпус предоставляет возможности сбора языкового материала для исследований сравнительного характера, причем этот сбор занимает времени на несколько порядков меньше по сравнению с ручным сбором и обработкой материала. Собранный при помощи корпуса материал может быть успешно использован при написании курсовых и дипломных работ. Так, студентами специализации «Компьютерная лингвистика» были защищены дипломные проекты на темы «Сравнительный анализ валентности ментальных глаголов в русском и английском языках: корпусное исследование», «Валентностные возможности вещественных имен существительных в русском и английском языках: корпусное исследование» и др. Отметим, что сбор языкового материала для нужд лингвистов при помощи корпуса текстов достаточно прост и может быть использован при обучении студентов любого гуманитарного ВУЗа.

ЛИТЕРАТУРА

1. Плунгян, В. А. Зачем мы делаем Национальный корпус русского языка? / В.А. Плунгян // Отечественные записки – 2005. – № 2. – С. 296-308.
2. Баранов А.Н. Введение в прикладную лингвистику: Учебное пособие / Н.А Баранов – М.: УРСС Эдиториал, 2001. – 360 с.
3. Национальный корпус русского языка [Электронный ресурс] / Институт русского языка им. В. В. Виноградова РАН. – М., 2003-2014. – Режим доступа: <http://ruscorpora.ru>. – Дата доступа: 02.09.2014.
4. Потемкин С.Б., Кедрова Г.Е. Выравнивание неразмеченного корпуса параллельных текстов / С.Б. Потемкин, Г.Е. Кедрова // Диалог – международная конференция по компьютерной лингвистике [Электронный ресурс]. – 2008. – Режим доступа: <http://www.dialog-21.ru/digests/dialog2008/materials/html/67.htm>. – Дата доступа: 02.09.2014.
5. Добрушина Н. Р. Как использовать Национальный корпус русского языка в образовании? / Н.Р. Добрушина // Национальный корпус русского языка: 2003–2005. – М.: Индрик, 2005. С. 308—329.