УДК 811.161

ОБ ОСНОВНЫХ ЗАДАЧАХ СОЗДАНИЯ ПАРАЛЛЕЛЬНОГО РУССКО-БЕЛОРУССКОГО КОРПУСА УЧЕБНЫХ ТЕКСТОВ

А.В. Зубов

Минский государственный лингвистический университет

proscien@mslu.by

Поступила в редакцию 03.03.2011

Рассматривается структура и назначение параллельного тегированного русско-белорусского корпуса учебных текстов.

Ключевые слова: информация, параллельный корпус, тегирование, учебный текст.

Корпусом параллельных текстов обычно называют множество текстов на одном какомлибо языке и их переводов на один или несколько других языков. В разных странах мира создано уже достаточно большое число таких параллельных корпусов текстов [1; 2; 3].

В Минском государственном лингвистическом университете (МнГЛУ) уже в течение многих лет создаются небольшие по объему параллельные корпусы текстов: русско-белорусский, англо-белорусский и немецко-белорусский [4]. Такие тексты могут быть использованы для решения целого ряда лингвистических задач в разных сферах:

а) в лексикографии и лексикологии – для составления различных словарей, определения значений слов, установления ассоциативных связей слов в тексте, выявления терминов и терминологических словосочетаний и т.д.;

б) в грамматике – для определения частоты использования грамматических морфем и классов слов в текстах различного типа, для выявления наиболее употребляемых типов словосочетаний и предложений;

в) в лингвистике текста – для дифференциации типов текста, создания конкордансов, выявления связи между предложениями в абзацах и между абзацами и т.д.;

г) при автоматическом переводе текстов – для поиска контекстов слов, имеющих несколько переводных эквивалентов, для поиска переводных эквивалентов терминологических и фразеологических словосочетаний в параллельных текстах и т.д.

Извлечь автоматически необходимую информацию из параллельных корпусов текстов можно лишь в том случае, если исходный и переведённый текст были заранее размечены (тегированы, аннотированы). Поэтому в параллельных корпусах текстов, созданных в МнГЛУ, был применен стандарт тегирования CES (Corpus Encoding Standard), успешно реализованный при разработке европейских проектов MULTEX 135 и EAGLES (Expert Advisory Group on Language Engineering Standard) в сотрудничестве с американским партнёром Vassar Cillege и французским партнёром CNRS (CENTRE National de la Recherché Scientifique). При его использовании в оформлении информации к словоупотреблениям текста были внесены некоторые незначительные изменения.

Этот стандарт удобен также тем, что он специально создан для автоматического решения задач прикладной лингвистики, машинного перевода, лексикографии и т.п.

В соответствии с этим стандартом каждое словоупотребление белорусских и иноязычных текстов получало набор определенных единых лексико-морфологических признаков. Так, для существительного указывались коды класса слова, одушевлённость, число, падеж, личность, сокращение ли это или имя собственное. Для глагола – коды класса слова, вид, время, залог, переходность, спряжение, лицо, число, род. Аналогично имели свои коды и слова других классов слов. В отличие от тегов CES, в создаваемом параллельном корпусе каждое словоупотребление имело определенные структурные признаки. Для слов всех классов указывалось число слогов в словоупотреблении и место ударного слога в нём. Это важная информация для изучения поэтических текстов. Именно с опорой на такие признаки (теги) и строятся компьютерные программы для извлечения различной информации из текстов параллельного корпуса [5].

В последние годы большое внимание стало уделяться корпусам учебных текстов [6; 7; 8]. Такой параллельный корпус учебных текстов создаётся сейчас в МнГЛУ. Его создание связано с тем, что в Республике Беларусь два государственных языка – белорусский и русский. И, естественно, есть школы, где обучение проводится или только на белорусском или только на русском языках. В некоторых вузах есть дисциплины, которые читаются и на белорусском и на русском языках. Такая языковая ситуация создает в процессе обучения целый ряд проблем, как в процессе преподавания этих языков, так и при подготовке для школ учебников и учебных пособий.

Нами был предварительно проведен анализ большого числа параллельных белорусских и русских учебников для школ по самым различным дисциплинам (физика, география, математика, трудовое обучение, история Беларуси и др.). В результате выяснилось, что в учебниках зафиксированы следующие виды информации:

1. Теоретические темы.
2. Детализация отдельных тем по уточняющим аспектам: «Главное», «Новые понятия и термины», «Вспомните» и т.п.
3. Детализация отдельных тем для проведения дискуссий: «А вы знаете, что …», «Обсудим? Поспорим? Доберемся до истины» и т.п.
4. Главные выводы.
5. Материал для повторения.
6. Вопросы и задания по темам.
7. Упражнения.
8. Контрольные задания.
9. Практические работы.
10. Исторические сведения.
11. Основные события и даты.
12. Словари терминов.

Такой корпус будет создаваться с использованием уже упомянутого выше стандарта тегирования CES, который уже был использован нами при разработке англо-белорусского, немецко-белорусского и белорусско-русского (не учебного) подкорпусов текстов.

Создаваемый параллельный тегированный русско-белорусский корпус учебных текстов позволит в целях совершенствования учебного процесса в школах:

1. отбирать примеры употребления слов, словосочетаний и предложений в текстах изучаемого языка;
2. демонстрировать на конкретных примерах способы разрешения двуязычной неоднозначности;
3. составлять автоматически учебные словари по различным предметным областям;
4. создавать русско-белорусские терминологические словари для школ по различным предметным областям;
5. находить и классифицировать ошибки обучаемых.

Для проведения научных исследований по педагогике и в сопоставительном языкознании такой корпус текстов позволит:

1. автоматически выделять группы слов определенного словоизменения или словообразования;
2. находить и выделять слова с определенными грамматическими характеристиками;
3. выделять структурные модели словосочетаний и предложений исходного и переводного языков;
4. проводить сопоставительный анализ двух языков на синтаксическом уровне.

Список литературы

1. Добровольский Д.О., Кретов А.А., Шаров С.А. Корпус параллельных текстов // Научная и техническая информация, сер. 2. Информационные процессы и системы, 2005. № 6. С. 121–127.

2. Сичинава Д.В., Шведова М.А. Параллельные корпусы в составе Национального корпуса русского языка: технологии и решаемые задачи // Компьютерная лингвистика: научное направление и учебная дисциплина. Сб. научных статей. Выпуск 1. Гомель: ГГУ, 2010. С. 30–34.

3. Зубов А.В. Лингвометодические возможности русско-белорусского параллельного корпуса текстов // Русский язык. Исторические судьбы и современность. IV Международный конгресс исследователей русского языка. Труды и материалы. М.: МГУ, 2010. С. 516–517.

4. Зубов А.В. Структура и назначение параллельных иноязычно-белорусских корпусов текстов // Беларуская мова ў культурнай і моўнай прасторы Славіі. Матэрыялы Міжнароднай навуковай канферэнцыі. г. Мінск. 24–25 лістапада 2009 г. Мінск: Права і эканоміка, 2009. С. 313–316.

5. Зубов А.В., Филимонова Т.А. Компьютерные программы для извлечения лингвистической информации из параллельных текстов // Материалы ежегодной научной конференции преподавателей и аспирантов университета. 27–28 апреля 2010 г. В пяти частях. Часть пятая. Минск: МГЛУ, 2010. С. 30–32.

6. Камшилова О.Н. Исследовательский потенциал корпуса английских текстов Петербургских школьников: анализ интерязыка // Изв. Рос. гос. пед. ун-та им. А.И. Герцена. 2009. № 7. С. 114–123.

7. Соснина Е.П. О разработке и использовании российского учебного корпуса переводов [Электронный дискурс] – Режим доступа: http://ling.ulstu.ru/linguistics/chair/ lecturers/ sosnina/development/.

8. Cobb T. Analyzing late interlanguage with learner corpora: Quebec replications of three European studies // Canadian Modern Lang. Review. 2003. № 59 (3). P. 393–423.

PRINCIPAL TASKS IN THE DEVELOPMENT OF THE PARALLEL RUSSIAN-BYELORUSSIAN CORPUS OF EDUCATIONAL TEXTS

A.V. Zubov

The structure and the purpose of the parallel tagged Russian-Byelorussian corpus of educational texts is considered.

Keywords: information, parallel corpus, tagging, educational text.