А.В. Зубов (Минск, МГЛУ)

О КОМПЛЕКСЕ КОМПЬЮТЕРНЫХ ПРОГРАММ ДЛЯ ИЗВЛЕЧЕНИЯ ЛИНГВИСТИЧСЕКОЙ ИНФОРМАЦИИ ИЗ ПАРАЛЛЕЛЬНОГО КОРПУСА УЧЕБНЫХ ТЕКСТОВ

Кафедра информатики и прикладной лингвистики Минского государственного лингвистического университета завершает работу по созданию параллельного русско-белорусского корпуса учебных текстов. Он включает тексты 7 школьных учебников («История Беларуси», «География», «Трудовое обучение», «Информатика», «Биология», «Обществоведение» и «Физика»), изданных в Республике Беларусь на русском и белорусском языках.

Особенностью этого корпуса текстов является то, что все словоупотребления этих текстов получают определенный набор тегов (индексов, признаков), указывающих на морфологическую и структурную организацию словоупотреблений этих текстов.

В качестве стандарта для такого тегирования был взят международный стандарт тегирования CES (Corpus Encoding Standart), который широко использовался при разработке европейских проектов по созданию корпусов текстов в США и Франции. В соответствии с этим стандартом каждое словоупотребление русских и белорусских текстов получало набор определенных лексико-морфологических признаков. Так, для существительного указывались коды класса слова, одушевленность, 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 нем. Это — важная информация для изучения поэтических текстов. Именно с опорой на такие признаки (теги) и строятся компьютерные программы для извлечения различной информации из текстов параллельных корпусов.

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

Анализ большого числа научных работ по использованию учебных корпусов текстов в процессе обучения и для подготовки учебных материалов показывает, что объем извлекаемой из таких текстов информации зависит как от целей создания учебных корпусов текстов, так и от структурной организации текстов, входящих в такие корпусы.

Для исследуемых текстов нами была изучена их структурная и семантическая организация. В итоге такого анализа выяснилось, что в них 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.

Большое внимание в наши дни уделяется возможностям использования корпусов текстов в лексикографии [1; 8]. В этом направлении нами создаются следующие компьютерные программы:

1. Программа создания алфавитно-частотного словаря по русским и белорусским текстам школьных учебников («Словарь по биологии», «Словарь по информатике» и др.).

2. Программа создания алфавитного русско-белорусского словаря на основе текстов исследуемых школьных учебников («Русско-белорусский словарь по биологии», «Русско-белорусский словарь по информатике» и др.).

Такие словари будут полезны русским школьникам, изучающим белорусский язык, и белорусским школьникам, изучающим русский язык.

Второй из этих типов словарей может стать основой для создания белорусских терминологических словарей по дисциплинам, изучаемым в школе. Основные принципы создания таких словарей были сформулированы Г.А. Цыхуном еще в 1997 году в работе [5, с. 4–7]. В том же сборнике приведен первый из таких словарей «Руска-беларускі слоўнік асноўнай матэматычнай тэрміналогіі школьных падручнікаў» [4, с. 16–23].

Такие белорусские терминологические словари школьных дисциплин могут помочь в создании общефилологического толкового словаря белорусского языка [6].

Большую помощь могут оказать параллельный русско-белорусский корпус текстов школьных учебников и при переводе белорусских технических тек-

стов на русский язык и русских технических текстов на белорусский язык [2; 7]. В простейшем случае, это перевод вновь создаваемых русских школьных учебников на белорусский язык.

Первым шагом на этом пути будет создание 2-х компьютерных программ:

1. Программа выделения контекстного окружения многозначных русских и белорусских слов.

2. Программа выделения терминологических словосочетаний.

Создаваемый параллельный корпус учебных текстов может быть использован и при создании новых учебников и учебных пособий по изучающим школьным дисциплинам. На первом этапе для этого будут созданы следующие компьютерные программы:

1. Программа выделения вопросов для проверки знаний по темам из использованных текстов учебников.

2. Программа, выделяющая основные события и даты, зафиксированные в текстах учебников.

3. Программа, выделяющая типы упражнений, задаваемых школьникам для проверки усвоения ими знаний.

Полезен создаваемый параллельный русско-белорусский корпус текстов и при проведении сравнительно-сопоставительного изучения русского и белорусского языков. Так как каждое словоупотребление этих текстов имеет полные наборы морфологических признаков, то можно для этого создать, например, такие компьютерные программы:

1. Определение наиболее употребительных типов предложений в текстах школьных учебников.

2. Определение частоты употребления в текстах учебников существительных, глаголов, прилагательных, слов других классов.

ЛИТЕРАТУРА

1. Беляева Л.Н. Лексикографический потенциал параллельного корпуса текстов // Труды Международной конференции «Корпусная лингвистика–2004». 11–14 октября 2004 г. — СПб., 2004. — С. 55–64.

2. Добровольский, Д.О. Корпус параллельных текстов и литературный перевод // НТИ. Серия 2. Информационные процессы и системы. — 2003. — № 10. — С. 13–18.

3. Егорова М.А. Корпусы и корпусная лингвистика переводчику // Социокультурные проблемы перевода. / Вып. 8. — Воронеж, 2008. — С. 84–94.

4. Тэрміналагічны бюлетэнь. / Вып. 1. — Мінск, 1997.

5. Цыхун Генадзь. Шляхі уладкавання беларускай тэрміналогіі // Тэрміналагічны бюлетэнь. / Вып. 1. — Мінск, 1997. — С. 4–8.

6. Шкода О.Я. Представленность специальной лексики в общефилологических толковых словарях белорусского языка // Прикладная лингвистика в науке и образовании. Сборник IV Международной научной конференции. 5–7 апреля 2012 г. — СПб., 2012. — С. 311–314.

7. Hansen-Schirra Silvia, Teich Elke. Corpora in human translation // Corpus Linguistics. An International Handbook. / Volume 2. — Walter de Gruyter. Berlin. New York, 2008. — P. 1159–1174.

8. Heid Ulrich. Corpus linguistics and lexicography // Corpus Linguistics. An International Handbook. / Volume 1. — Walter de Gruyter. Berlin. New York, 2008. — P. 131–154.