А.В. Зубов (Минск, МГЛУ)

ОБ ОСНОВНЫХ ЗАДАЧАХ СОЗДАНИЯ РУССКО-БЕЛОРУССКОГО КОРПУСА УЧЕБНЫХ ТЕКСТОВ

В последние десятилетия все большее развитие получает корпусная лингвистика. Ее появление связано с тем, что современные компьютеры все чаще используются для решения различных лингвистических задач. На этом пути ярко выявились те проблемы, которые компьютер не может решить, имея в памяти отдельные тексты или словари. Все чаще взоры исследователей обращаются к корпусам текстов, содержащим в себе множество текстов различного типа и определенную информацию о каждом тексте, предложении и слове. В современном понимании корпус текстов — «это электронное собрание текстов, размеченное таким образом, чтобы в нем можно было быстро найти слова и конструкции с заданными грамматическими и другими интересными лингвисту свойствами» [3, с. 7].

И тогда под **корпусной лингвистикой** понимается цикл исследований, связанных с правилами организации текстов в корпус, разработкой алгоритмов анализа таких текстов в рамках некоторой научной методологии [4, с. 128].

Сегодня к корпусам текстов предъявляются следующие требования:

1. Расположение множества текстов на магнитном носителе.
2. Наличие определенной процедуры отбора текстов в корпус текстов.
3. Единая методика представления сведений о текстах и их единицах на магнитном носителе.
4. Конечный размер корпуса текстов.

Корпусы текстов стали активно создаваться с начала 50-х гг. прошлого века. Создан Британский корпус английского языка объемом 600 млн. словоупотреблений, американский национальный корпус, венгерский, итальянский, хорватский, чешский, японский корпусы (каждый объемом в 100 млн. словоупотреблений), русский национальный корпус объемом в 150 млн. словоупотреблений. Последние данные о создаваемых европейских национальных корпусах представлены в статье [1]. Впервые в Беларуси создан корпус текстов белорусского языка объемом в 1 млн. словоупотреблений [2]. В эти корпусы включаются, в основном, тексты художественной литературы и публицистики.

Ни в одном из таких корпусов нет учебных текстов. Даже в одной из последних статей по совершенствованию национального корпуса русского языка [5] не предусмотрено изучение учебных текстов.

Кафедра информатики и прикладной лингвистики имеет опыт создания параллельных корпусов текстов. В процессе создания корпуса текстов белорусского языка совместно с Институтом языка и литературы им. Якуба Коласа и Янки Купалы мы одновременно подготовили три параллельных белорусско-иноязычных корпуса: белорусско-русский, белорусско-

английский и белорусско-немецкий. Но все они основаны на использовании художественных и публицистических текстов.

В то же время необходимость создания параллельного русско-белорусского корпуса учебных текстов вызвана самой языковой ситуацией в Республике Беларусь, имеющей два официальных государственных языка. В школах Беларуси многие дисциплины преподаются как на русском, так и на белорусском языках. Как в процессе преподавания таких дисциплин и создания соответствующих учебников и учебных пособий, так и при проведении научных исследований по сопоставительному изучению русского и белорусского языков неоценимую помощь может оказать параллельный русско-белорусский корпус учебных текстов.

Для проведения такой работы был предварительно проведен анализ большого числа белорусских и русских учебников для школ по самым различным дисциплинам (физика, география, математика, трудовое обучение, история Беларуси и др.). 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. Материла 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.

Так как предполагается, что нами будет создаваться тегированный (аннотированный, размеченный) текст, то необходимо было выбрать определенную систему тегирования создаваемых текстов, т.е. разработать систему букв и цифр, которые бы указывали на морфологические и структурные признаки соответствующего словоупотребления. В качестве такой системы был использован набор тегов CES (Corpus Encoding Standart), который уже применялся ранее в Европе для создания различных корпусов текстов. К тому же, этот код используется нами при разработке большого корпуса текстов белорусского языка, а также параллельных англо-белорусского и немецко-белорусского корпусов текстов. Дополнительно в эту систему тегов введены структурные признаки словоупотреблений (число слогов в словоупотреблении и место ударного слога в нем). Они могут понадобиться при изучении стихотворных 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 по различным предметным областям.

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.

ЛИТЕРАТУРА

1. Жулего, А.В. Современное состояние европейских национальных языковых корпусов / Компьютерная лингвистика: научное направление и учебная дисциплина. Сборник научных статей. — Гомель, 2010. — С. 112–116.

2. Кошчанка, У.А. Актуальны стан і перспектывы развіцця корпуснай лінгвістыкі і камп’ютэрнай лексікаграфіі ў Інстытуце мовы і літаратуры НАН Беларусі / Беларуская мова ў культурнай і моўнай прасторы Славіі. Матэрыялы Міжнароднай навуковай канферэнцыі. — Мінск, 2009. — С. 316–321.

3. Плунгян, В.А. Зачем нужен национальный корпус русского языка. Неформальное введение / Национальный корпус русского языка: 2003–2005. Результаты и перспективы. — М., 2005. — С. 6–21.

4. Рыков, В.В. Прагматически ориентированный корпус текстов / Компьютерная лингвистика и интеллектуальные технологии. Труды Межд. конференции «Диалог–99» (Москва — Таруса). — М., 1999. — С. 127–134.

5. Савчук, С.О. Тексты ограниченного обращения в составе национального корпуса русского языка / НТИ. Сер. 2. Информационные процессы и системы. — 2005. — № 3. — С. 23–28.