А.В. Зубов (Минск, МГЛУ)

О ЗАДАЧАХ СОЗДАНИЯ ПОДКОРПУСА ТЕКСТОВ РУССКИХ И БЕЛОРУССКИХ УЧЕБНИКОВ ПО ЯЗЫКУ И ЛИТЕРАТУРЕ

Первый корпус текстов был создан американскими исследователями У. Френсизом и Г. Кучерой в 1963 году [1]. Этот корпус текстов («Брауновский корпус») включал 500 отрывков разных текстов печатной прозы США, каждый из которых содержал по 2 000 словоупотреблений. Они представляли 15 наиболее массовых жанров англоязычной прозы 60-х гг.

Вскоре после создания Брауновского корпуса текстов подобные корпусы текстов стали активно создаваться на разных языках. Так, уже в 1992 г. Международной организацией «Европейская корпусная инициатива» (ECI) зафиксировала около 50 крупных корпусов текстов на европейских языках, каждый объемом от 12 000 до 5 000 000 словоупотреблений.

С 2003 г. и в Республике Беларусь начинаются активные работы по созданию корпуса текстов белорусского языка. Первоначально такие работы проводились в рамках «Компьютерного фонда белорусского языка» [2], а затем — в сотрудничестве с Центром исследования белорусской культуры, языка и литературы Национальной Академии наук Республики Беларусь, Белорусского государственного педагогического университета им. М. Танка и Минского государственного лингвистического университета (МГЛУ) [3].

В 2010 г. кафедрой информатики и прикладной лингвистики МГЛУ закончено создание тегированного корпуса текстов белорусского языка объемом в 1 000 00 словоупотреблений. В него вошли художественные тексты (40%), тексты периодической печати (12%), научные и учебные тексты (40%), тексты деловой прозы (2%) и записи устной речи (6%). Одновременно с этим, созданы 3 параллельных тегированных подкорпуса текстов: белорусско-русский, белорусско-английский и белорусско-немецкий. Каждый из них включает по 300 000 словоупотреблений каждого языка. В состав этих подкорпусов входят, в основном, тексты художественной литературы и тексты деловой прозы [4].

Все больше внимания в мире уделяется созданию корпусов учебных текстов. В Минском государственном лингвистическом университете созданы первые четыре русско-белорусских подкорпуса учебных текстов по 4-м русским и белорусским школьным учебникам: «Обществоведение», «Информатика», «Биология» и «Физика». Каждый такой подкорпус содержит по 15 000 словоупотреблений русского и белорусского языков [5]. Для извлечения различной информации из этих подкорпусов были созданы 8 компьютерных программ.

Сейчас кафедра информатики и прикладной лингвистики МГЛУ начала работу по созданию русского и белорусского подкорпусов текстов школьных учебников по языку и литературе. Для проведения такой работы отобраны учебники: «Русский язык» и «Беларуская мова» за 9 и 10 классы, «Русская ЛИТЕРАТУРА» и «Беларуская літаратура» за 9 и 10 классы.

На первом этапе эти учебники были изучены с точки зрения их содержания. Выделены те составляющие этих учебников, которые и составят отдельные тексты тегированных подкорпусов:

1. Теория языка.
2. Теория литературы.
3. Сведения о писателях и поэтах.
4. Сведения об ученых.
5. Сведения о литературном произведении.
6. Примеры из художественных текстов.
7. Вопросы и контрольные задания.

Далее, по каждому тексту будут получены частотно-алфавитные словари. С опорой на них все словоупотребления отобранных текстов получат наборы морфологических тегов, показывающих их принадлежность к определенному классу слов, и все другие морфологические признаки, присущие словам каждого класса.

С опорой на создаваемые подкорпусы текстов школьных учебников могут быть автоматически решены следующие задачи:

– отбор примеров употребления слов, словосочетаний и предложений в текстах изучаемого языка;

– демонстрация на конкретных примерах способов разрешения двуязычной неоднозначности;

– автоматическое составление учебных словарей по различным авторам и произведениям;

– создание русско-белорусских терминологических словарей по языку или литературе;

– автоматическое выделение группы слов определенного словоизменения или словообразования;

– нахождение и выделение слова с определенными грамматическими характеристиками;

– выделение структурных моделей словосочетаний и предложений исходного и переводного языков;

– проведение сопоставительного анализа двух языков на синтаксическом уровне.

Наличие морфологических тегов у всех слов учебников позволит автоматически выявлять и систематизировать ошибки в письменных работах школьников [5].

ЛИТЕРАТУРА

1. Рыков, В.В. Прагматически ориентированный корпус текстов // Компьютерная лингвистика и интеллектуальные технологии: труды Междунар. конф. «Диалог–99». — М.: Наука, 1999.

2. Рубашко, Н.К. Компьютерный фонд белорусского языка / Н.К. Рубашко, Г.П. Невмержицкая // Информационные системы и технологии (IST’2004): материалы II Междунар. конф., Минск, 8–10 ноября 2004 г.: в 2 ч. — Минск, 2004. — Ч. 1. — С. 217–224.

3. Кошчанка, У.А. Актуальны стан і перспектывы развіцця корпуснай лінгвістыкі і камп’ютэрнай лексікаграфіі ў Інстытуце мовы і літаратуры НАН Беларусі /

У.А. Кошчанка, І.Л. Капылоў // Беларуская мова ў куль турнай і моўнай прасторы Славіі: матэрыялы Міжнарод. навук. канф., Мінск, 24–25 лістапада 2009 г. — Мінск: “Права і эканоміка”, 2009. — С. 316–321.

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

5. Зубов, А.В. Об основных задачах создания параллельного русско-белорусского корпуса учебных текстов // Вестн. Нижегород. ун-та им. Н.И. Лобачевского. — Нижний Новгород: НГУ, 2011. — № 6. –Ч. 2. — Т. 1. — С. 190–192.

6. Камшилова, О.Н. Исследовательский потенциал корпуса английских текстов Петербургских школьников: анализ интерязыка // Известия Российского государственного педагогического университета им. А.И. Герцена. Научный журнал. — № 7. — 2009. — С. 114–123.