А. А. Алсуфьева, О. В. Киссилев
A. A. Alsufieva, O. V. Kisselev

ОПЫТ РАЗРАБОТКИ ЛИНГВИСТИЧЕСКОГО КОРПУСА НА ОСНОВЕ ПИСЬМЕННЫХ РАБОТ СТУДЕНТОВ РКИ ПРОДВИНУТОГО УРОВНЯ

DEVELOPMENT OF A RUSSIAN LEARNER CORPUS OF ACADEMIC WRITING

Аннотация. В статье рассматриваются вопросы разработки экспериментального учебного корпуса, составленного из письменных работ студентов продвинутого этапа обучения РКИ — Russian Learner Corpus of Academic Writing (RULEC): структура корпуса, принципы отбора и параметры текстов для корпуса. Также обсуждаются возможности использования корпуса для разного рода лингвистических исследований.

Abstract. The article reports on the development and the potential of the Russian Learner Corpus of Academic Writing (RULEC). The corpus design, the criteria of the selection of texts for RULEC, and the form of socio-linguistics questionnaire are discussed. The article will be of interest to those involved in second language acquisition research, teachers of advanced language courses and those interested in corpus-based applications.

1. Введение

Лингвистические корпуса занимают все большее место в различных лингвистических исследованиях, как в области теоретического языкознания, так и в области прикладных исследований (теория усвоения языка, корпусные методики обучения родному и неродному языкам и проч.)[1]. В 1990-е гг. одним из активно развивающихся направлений в корпусной лингвистике стало создание корпусов работ учащихся,

[1] Национальный корпус русского языка и проблемы гуманитарного образования / Под общ. ред.: *Н.Р. Добрушина.* М.: Изд. дом ГУ–ВШЭ, 2007.

изучающих язык как иностранный или второй (неродной), и лингвистический анализ этих работ. Такие корпуса получили название learner corpora – «корпуса работ учащихся», «учебные корпуса» или «ученические корпуса». В настоящей работе мы используем термин «учебный корпус».

Работа с учебными корпусами стала распространенной практикой, прежде всего, в области изучения и преподавания английского языка как неродного[2]. Пионерами этого направления считаются лингвисты Великобритании и Бельгии. К важнейшим работам, посвященным описанию учебных корпусов, следует отнести статью английского лингвиста Джефри Лича «Learner corpora: What they are and what can be done with them» (1998)[3] и статью бельгийской исследовательницы Силвиан Грейнджер «The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research» (2003)[4]. Тем не менее понятие учебного корпуса – корпуса, составленного из работ, написанных школьниками или студентами, изучающими русский язык, еще только входит в методику и практику преподавания русского языка.

Учебный корпус, как и любой другой текстовый корпус,– это не просто архив работ студентов, а специально организованная система, построенная на определенных принципах и включающая различные виды аннотаций, например, морфологическую разметку, синтаксическую разметку, кодировку ошибок, «социолингвистический паспорт» студента и т.д. Именно аннотированный учебный корпус представляет собой ценный ресурс и инструмент как для изучения лингвистических

---

[2] *Pravec N.* Survey of learner corpora // ICAME JOURNAL: Computers in English Linguistics, 2002. Vol. 26.

[3] *Leech G.* Learner corpora: What they are and what can be done with them // Learn English on computer. 1998. xiv – xx.

[4] *Granger S.* The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research // TESOL Quarterly. 2003. Vol. 37, No. 3.

феноменов в целом, так и для решения вопросов, связанных с обучением иностранному языку и проблемами его усвоения в частности.

2. Лингвистический корпус на основе письменных работ студентов, изучающих русский язык как второй

В предлагаемой работе рассматриваются принципы разработки корпуса, составленного из письменных работ студентов продвинутого уровня обучения русскому языку как иностранному. Насколько нам известно, такого рода корпуса – доступного в электронном формате – на материале русского языка пока нет (хотя работы по составлению подобных корпусов ведутся, например, в Университете Хельсинки и на базе Русского Национального Корпуса); наш учебный корпус можно считать новаторскими проектом экспериментального характера.

Работа по созданию учебного корпуса Russian Language Learner Corpus of Academic Writing, сокращенно RULEC, началась в 2009/2010 учебном году на кафедре русского языка и литературы Портландского государственного университета (Portland State University, Oregon, the U.S.A.). В настоящее время в базе корпуса имеется приблизительно 2000 письменных работ разных жанров, длиной от короткого абзаца до небольшой исследовательской работы, выполненных студентами продвинутого уровня владения русским языком. Общий объем корпуса более 450 тыс. словоупотреблений. Важной характеристикой RULEC является то, что корпус включает работы не только «традиционных» англоговорящих студентов, но и студентов, для которых русский язык является «унаследованным» (heritage speakers of Russian), т.е. студентов из русскоговорящих и украиноговорящих семей, эмигрировавших в США. Студенты обучаются в смешанных группах, посещая одни и те же занятия и выполняя одни и те же задания в рамках инновационной учебной программы «Русский флагман» (Russian

Flagship), являющейся частью государственного образовательного проекта США – The Language Flagship[5].

Для корпуса RULEC отбираются такие работы, которые предполагают создание самостоятельного письменного речевого произведения. Эти работы связаны с учебно-научной деятельностью студентов и представляют собой образцы академического письма на русском языке.

Одним из важных вопросов разработки RULEC был вопрос о выборе параметров для регистрации текстов: какой входной информацией должен сопровождаться каждый текст в корпусе. Безусловно, необходимо было учесть как социолингвистические характеристики студента (родной язык студента, уровень владения русским языком, пол), так и характеристики текста. Письменные работы студентов разнообразны по жанрам, способам изложения информации, каждая работа нацелена на развитие определенной коммуникативной функции (описание, аргументация, сравнение и т.д.), работы могут быть написаны как во внеаудиторное время (домашние работы), так и во время занятия, индивидуально или в группе. На основе указанных характеристик мы выделили следующие параметры для документации каждого корпусного текста:

(1) псевдоним, присвоенный студенту;
(2) пол;
(3) первый язык (русский, украинский, английский);
(4) уровень владения русским языком по стандарту ACTFL[6] (определяется с помощью методов и тестов, разработанных в США);

[5] «The Language Flagship leads the nation in designing, supporting, and implementing a new paradigm for advanced language education. Through an innovative partnership among the federal government, education, and business, The Language Flagship seeks to graduate students who will take their place among the next generation of global professionals, commanding a superior level of fluency in one of many languages critical to U.S. competitiveness and security». URL: **http://www.thelanguageflagship.org/**

(5) учебный год;
(6) название курса;
(7) учебная неделя;
(8) длина текста (предложение, абзац, развернутый текст);
(9) функция (функционально-смысловой тип текста);
(10) временное ограничение (без ограничения времени для выполнения работы или с ограничением);
(11) форма работы (индивидуальная/групповая).

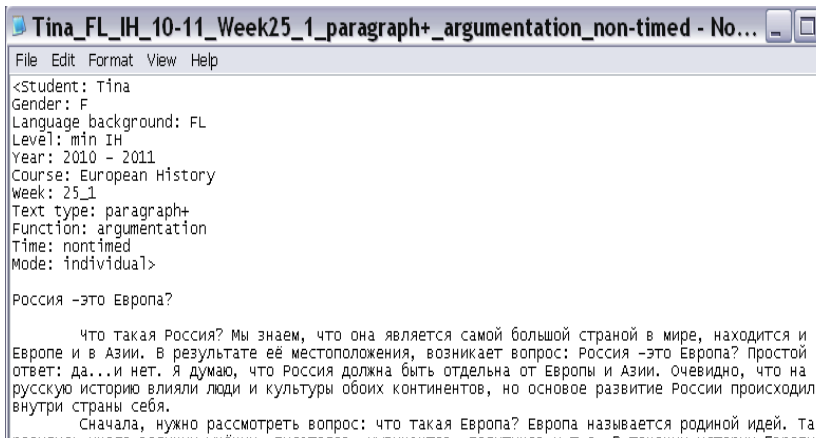

Рис. 1. Документация текста в корпусе

Важность учета социолингвистических данных очевидна. Корпусные исследования, учитывающие социолингвистические параметры (например, родной язык учащегося), могут помочь выявить особенности усвоения второго или иностранного языка в зависимости от того, как проходит процесс формирования языковой личности. Владение такой информацией – необходимое условие для развития методики преподавания русского языка как иностранного на продвинутом уровне.

---

[6] ACTFL – The American Council on the Teaching of Foreign Languages. URL: **http://www.actfl.org/**

Как известно, корпус наиболее эффективен как инструмент для лингвистических исследований, когда составляющие его тексты содержат не только метаописание, но и разного рода лингвистическую информацию. Работа по аннотации (tagging) корпуса ведется в настоящий момент; эта работа, однако, связана с определенными трудностями. Русская письменная речь учащихся (даже на уровне близком к уровню носителя языка) заметно отклоняется от лексико-грамматических и стилистических норм русского языка. Этот фактор определяет неизбежность ручного редактирования разметки и/или серьезной доработки существующих компьютерных программ для морфологического и синтаксического анализа русскоязычных текстов. В то же время эта проблема становится и важной и интересной исследовательской задачей для корпусных лингвистов, занимающихся разработкой программ для автоматизированного поиска ошибок.

Несмотря на отсутствие грамматической и семантической аннотации, RULEC может успешно использоваться в «сыром» виде для целого ряда исследовательских проектов. Прежде всего, материал корпуса дает возможность составить «лингвистический портрет» группы учащихся или отдельного учащегося. Также уже сейчас можно провести анализ лексического и синтаксического разнообразия текстов учащихся разного уровня владения русским языком; обратиться к проблемам лексической сочетамости (например, употребление неносителем языка конструкций с более или менее фиксированными лексическими компонентами: *обратить внимание, принять во внимание, провести анализ*); рассмотреть явления языковой интерференции (например, влияние английского предлога *through* на употребление русского предлога *через*); подойти к проблеме формирования орфографических и пунктуационных навыков на иностранном языке; проанализировать стратегии организации письменного текста на иностранном языке и многое другое.

Использование корпусных данных для учебных целей давно стало популярной практикой и в лингвометодике. Как

единодушно отмечается в исследованиях, посвященных корпусам работ студентов, в учебном корпусе заложен целый ряд возможностей для оптимизации учебно-методической работы[7]. Особенности употребления тех или иных языковых единиц, разнотипные отклонения от нормы, стратегии построения текста, проявляющиеся в работах учащихся, дают возможность не только проверить те или иные предположения о процессе усвоения языка и даже о самой природе языка, но и разработать эффективную методику обучения языку.

Так, используя тексты, составляющие RULEC, несложно выявить «проблемные места» в русской речи учащихся, то есть такие лексемы, грамматические формы и конструкции, которые вызывают особые затруднения у студентов в усвоении и корректном употреблении, что в свою очередь позволяет преподавателю откорректировать систему и типы упражнений для работы над «проблемным» языковым материалом. Кроме того, RULEC дает возможность сопоставить ошибки в работах англоговорящих студентов, изучающих русский язык как иностранный, и студентов, для которых русский язык является «языком наследия»; такие контрастивные исследования помогают определить языковой материал, представляющий трудность для обеих групп. Важность и необходимость такой информации для разработки учебных заданий и целых учебных комплексов для работы со смешанными группами очевидна: оптимизируется процесс обучения.

3. Заключение

В заключение отметим еще раз, что RULEC является единственным известным нам корпусным проектом в области методики и практики преподавания РКИ. Мы не беремся утверждать, что наш проект универсален или лишен недостатков – он прежде всего отвечает целям и задачам программы «Русский

[7] Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М.: Индрик, 2005.

флагман», разработкой которой занимаются авторы RULEC. Однако мы надеемся, что описание корпуса RULEC может послужить основой и моделью для создания учебных корпусов разного типа в области преподавания русского языка, русского языка как «унаследованного» и русского языка как иностранного.