ЛИНГВИСТИЧЕСКИЕ ПРОБЛЕМЫ СОЗДАНИЯ РЕЧЕВОГО КОРПУСА. РУССКИЕ РЕЧЕВЫЕ КОРПУСА

М.Ю. СИМАКОВА

Московский государственный университет им. М.В. Ломоносова
Ленинские горы, д. 1, г. Москва, 119991, Российская Федерация
friesan@yandex.ru

В статье рассмотрены вопросы построения речевых корпусов как технологических, так и общего назначения. Также обсуждаются лингвистические проблемы, возникающие на этапах проектирования корпуса. Дана краткая характеристика русских речевых корпусов в основном общего назначения.

Ключевые слова: моделирование речевых корпусов, русские речевые корпуса.

Речевые корпуса широко используются в качестве лингвистического инструмента для решения разного рода фундаментальных и прикладных задач. Под речевым корпусом понимают структурированное множество речевых фрагментов, которое обеспечено программными средствами доступа к отдельным элементам корпуса. Речевой фрагмент, в свою очередь, представляет собой «оцифрованный фрагмент речевого сигнала, который сопровождается ассоциированной информацией определенного типа – аннотацией» [1]. Речевой корпус может быть эффективно применен не только при создании, обучении и тестировании систем распознавания и синтеза речи, но также быть полезным и для фундаментальной науки (с целью выявления и осмысления лингвистических фактов), для обучения (иностранным языкам), в лингвокриминалистике и медицинской диагностике.

Процесс создания речевого корпуса (РК) достаточно сложен. Материал для конкретного корпуса требуется отобрать, записать, оцифровать и описать согласно требованиям исследования, т.е. аннотировать. Содержание аннотации для каждого корпуса определяется его разработчиками, в соответствии с необходимой в данном проекте лингвистической и экстралингвистической информацией. На каждом из этапов создания корпуса исследователь сталкивается с различного рода лингвистическими проблемами. Наиболее трудоемким является процесс озвучивания корпуса, так как для сбора разнообразного речевого материала требуется обычно записать большое количество дикторов (их число зависит от типа корпуса, в среднем для исследовательских корпусов и корпусов для распознавания речи – это 50–100 человек), и последующего аннотирования полученных аудиофайлов. Помимо названных проблем существуют и другие. В частности, разработка эффективного программного обеспечения для работы с речевым материалом, стандартизация аннотаций, создание больших и информационно богатых РК, обеспечение их общедоступности и многопрофильности и, наконец, достаточное финансирование для осуществления всех намеченных целей.

В нижеследующей таблице в качестве примера корпусных разработок по устной речи приведены характеристики некоторых русских РК. Рассмотрены корпуса общего назначения. Стоит сразу отметить, что работа над некоторыми проектами еще не закончена.

Название	Краткое описание	Целевое назначение
Один речевой день. 2000-2002 гг. ИСА и РАН с участием компании Cognitive Technology Ltd.	Диалоги, каждая реплика аннотирована по нескольким основным уровням, единица описания - графическое слово целиком + «паралингвистические явления.	Изучение живой устной речи.
Интонация русского диалога Проект речевой группы филологического ф-та МГУ, 2005–2009 гг.	Записи естественного диалогического дискурса в средствах массовой информации. Базовая единица – отдельная реплика = Аудиофайл + просодическая транскрипция.	Изучение функций и формы фразовой просодии русской речи.
МУРКО Мультимедийный русский корпус Подкорпус НКРЯ. Общий доступ с 2010 г.	Аудиозаписи и кинофильмы, разрезанные на небольшие фрагменты. Базовая единица: Кликст — пара «клип + текст».	Мультимодальное изучение устной речи: звук + жест.
КРЛЯ Корпус русского литературного языка (СПб) Подкорпус НКРЯ.	Звучащие тексты + фонетическая транскрипция.	Разработка антропоморфных алгоритмов преобразования речевого сигнала в линейную последовательность лексических единиц
RuSpeech ИСА и РАН с участием Cognitive Technology Ltd. и сотрудников речевой группы кафедры ТиПЛ филологического ф-та МГУ, 2000–2002 гг.	Звучащая речь в режиме чтения + транскрипция.	Создание дикторонезависимой системы распознавания слитной русской речи.

Список литературы

1. Кривнова О.Ф. Речевые корпуса на новом технологическом витке // Речевые технологии 2, 2008. – 96 с.

2. Кодзасов С.В., Архипов А.В., Захаров Л.М., Кривнова О.Ф. База данных «Интонация русского диалога»: реплики-сообщения // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2007».

3. Bogdanov D. Russian large-scale speech corpus RuSpeech // SPECOM'.

4. Гришина Е.А. Национальный корпус русского языка как источник сведений об устной речи // Речевые технологии 3, 2008. – 96 с.

5. Богданова Н.В., Асиновский А.С. и др. Звуковой корпус русского языка «Один речевой день»: пути пополнения и первые результаты исследования // Конференция Диалог 2008. Электроные ресурсы.

6. Корпус русского литературного языка – КРЛЯ. http://www.narusco.ru/.