Речевые корпуса на новом технологическом витке

О.Ф. Кривнова,
доктор филологических наук

Корпуса звучащей речи, которые называют также речевыми базами данных, представляют собой важнейший тип языковых ресурсов. Интерес к созданию речевых корпусов был в значительной степени инициирован разработками в области автоматического распознавания речи, где исследователям приходится сталкиваться с огромной акустической вариативностью звуковых единиц языка, которая имеет весьма разнообразные источники. Однако сегодня речевые корпуса имеют более широкое применение, и их разработка сама по себе постепенно превращается в самостоятельное и популярное направление речевых технологий. В статье рассматривается история разработок в этой области, их современное состояние, даётся краткое описание речевых и лингвистических корпусов для русского языка.

Речевой корпус как разновидность языковых ресурсов

Корпуса звучащей речи, которые называют также речевыми базами данных, представляют собой важнейший тип языковых ресурсов.

Речевой корпус — это структурированное множество речевых фрагментов, которое обеспечено программными средствами доступа к отдельным элементам корпуса. **Речевой фрагмент** как базовая единица корпуса представляет собой оцифрованный фрагмент речевого сигнала, который сопровождается ассоциированной информацией определённого типа (типов). Такая информация называется также **аннотацией** к речевому фрагменту.

В настоящее время задача создания больших, разнообразных и информационно богатых (многоуровневых) речевых корпусов, а также удобного и надёжного инструментария для их разработки и использования становится всё более актуальной как для компьютерных приложений, так и для фундаментальных научных исследований. Современные системы распознавания речи, дающие наиболее высокие показатели надёжности, базируются преимущественно на методах статистического моделирования речевых и языковых явлений и требуют обучения на больших массивах аннотированной звучащей речи, записанной от многих дикторов (не менее 100 человек).

Современный подход к синтезу речи по тексту, основанный на конкатенации акустических фрагментов разной размерности, также предполагает использование больших речевых корпусов [1]. Попытки применения статистических методов для формирования речевого сигнала при синтезе речи также способствуют возрастанию роли речевых корпусов в дальнейшем развитии речевых технологий разного профиля [2].

Специалисты считают, что корпусной подход (corpus-based approach) является определяющим для развития технологий синтеза, особенно при моделировании просодических характеристик речи и индивидуальных особенностей говорящего. Отмечаются также такие достоинства этого подхода, как формализация процедур обучения, применение итеративного обучающего процесса с исправлением возникающих и контролируемых ошибок, возможность контроля и объективной оценки работы различных прикладных систем на стандартизованной основе (на одних и тех же речевых корпусах). Практика показывает, что при наличии речевых корпусов и технологий обучения создание прототипической версии автоматического распознавателя или синтезатора речи занимает не так уж много времени. В литературе указываются сроки от двух месяцев до полугода. Для коммерчески ориентированных разработок это немаловажное обстоятельство.

Современные речевые технологии базируются не только на речевых корпусах, но нуждаются также и в более широких, богатых информационно, лингвистических корпусах, т.е. коллекциях специальным образом обработанных текстов, как письменных, так и устных, на данном языке. Какие принципы лежат в основе устройства текстовых лингвистических корпусов, для каких речевых задач их можно использовать и как — это отдельная тема, которая в данной статье не обсуждается (см., однако, в конце статьи краткое описание национального корпуса русского языка — НКРЯ).

Было бы неправильно думать, что речевые корпуса представляют интерес только для речевых технологий. Наличие представительных речевых корпусов в электронном формате, снабжённых специальной информацией, уровень развития современных программных средств обработки звучащей речи и постоянно возрастающие мощности компьютерной техники дают учёным-лингвистам недоступную ранее возможность для проведения крупномасштабных и статистически достоверных исследований на разнообразном речевом материале. Среди других социально важных применений речевых корпусов вне сферы собственно речевых технологий можно отметить задачи обучения иностранным языкам, лингвокриминалистику и медицинскую диагностику.

Из истории разработок

Первые речевые корпусы (далее РК) были созданы в первой половине 80-х годов прошлого века в США для американского варианта английского языка, где их разработка финансировалась Министерством обороны, а организация работ была поручена национальному институту стандартов и технологий NIST (National Institute of Standards and Technology). Основное назначение первых РК — тестирование и оценка работы систем распознавания речи на одном и том же стандартном речевом материале.

Во второй половине 80-х годов произошли значительные сдвиги в компьютерной технике: возросла мощность компьютеров и объёмы хранения данных; происходило массовое внедрение персональных компьютеров. К этому времени были подведены окончательные итоги крупномасштабных государственных проектов ARPA/DARPA (Defense Department's Advanced Research Projects Agency) США, которые были направлены на анализ и оценку перспектив распознавания слитной речи с большим словарём и человеко-машинных диалоговых систем с устным вводом информации [3]. Проведённые в рамках этих проектов исследования ярко продемонстрировали преимущества систем распознавания речи на основе теории распознавания образов, статистических методов и обучающих речевых корпусов (сравнительно с экспертными системами на основе лингвистических знаний и правил). Этот временнòй период можно считать началом формирования нового направления речевых технологий, связанного с созданием речевых корпусов.

При государственной поддержке в США в 80-е годы были созданы: TI-DIGITS корпус (1984) для тестирования систем распознавания изолированных цифр и цифровых последовательностей; Road Rally для анализа и распознавания ключевых слов (word spotting); King Corpus для систем идентификации говорящего (speaker recognition); корпус TIMIT (Texas Instruments & Massachusetts Institute of Technology, Acoustic-Phonetic Continuous Speech Corpus 1980–1990), который послужил прототипом для многих других речевых корпусов, в том числе и не англоязычных. Были разработаны также специализированные речевые корпуса Resourse Management (RM) и Wall Street Journal (WSJ, позднее CSRNAB (Continuous Speech Recognition of North American Business News)) для исследований в области распознавания слитной речи и корпус Air Travel Information Service (ATIS) для исследования спонтанной речи и понимания естественного языка в диалоговых системах. Краткая характеристика перечисленных корпусов даётся ниже в таблице 1.

Таблица 1

Краткая характеристика речевых корпусов 80-х годов XX в.

Название	Назначение: для использования в проектах	Язык	Год	Общая характеристика
TI-DIGITS	Распознавание цифр и их последовательностей	Амер — англ.	1984	?
ROAD RALLY	Распознавание ключевых слов в речевом массиве	Амер — англ.	Первая половина 80-х г.	?
KING CORPUS	Идентификация говорящего	Амер — англ.		?
RM	Распознавание запросов в области военно-морской службы	Амер — англ.	Вторая половина 80-х г.	160 дикторов, словарь 1000 слов, 21000 предложений
WSJ	Дикторонезависимое распознавание слитной речи	Амер — англ.		Чтение новостных текстов в различных сферах бизнеса, словарь 20000 слов
ATIS	Распознавание запросов в сфере обслуживания гражданской авиации	Амер — англ.		Спонтанные диалоги сфере обслуживания гражданской авиации
TIMIT	Для широкого использования; дикторонезависимое распознавание слитной речи; научные исследования	Амер — англ.	80–90-е годы	630 дикторов, отдельные предложения — 2432; словарь 6229 единиц, адаптированный Merriam-Webster Pocket Dictionary 964

Практика показала, что создание хорошего речевого корпуса представляет собой довольно сложную технологическую задачу, требующую значительных финансовых и кадровых вложений. Для её решения в 90-е годы XX в. были созданы специальные координационные центры по сбору, хранению, распространению и созданию общедоступных и стандартизованных языковых ресурсов, в том числе речевых. Среди них:

- **LDC** — Linguistic Data Consorcium, http://www.ldc.upenn.edu.
- **CSLU** — Center for Spoken Language Understanding, Oregon Graduate Institute, http://www. CSLU.ogi.edu.
- **ELRA** — European Language Resources Association, http://www.icp.grenet.fr/elra.

Более подробные сведения о центрах языковых и речевых ресурсов можно найти в [5]. С момента образования указанных координационных центров начался второй этап технологического развития РК.

Речевые корпуса на современном этапе технологического развития (конец XX — начало XXI века)

Коллекции речевых корпусов, которые предлагаются координационными центрами, с каждым годом увеличиваются, и всё больше специалистов участвует в их разработке. Одновременно растёт мощность, разнообразие и программное оснащение самих корпусов.

Самым мощным на сегодняшний день является центр LDC (США), который в 2008 году отмечает свой 15-летний юбилей. За прошедшие годы центр участвовал в создании и распространении более 50 000 лингвистических, в том числе речевых, корпусов на разном языковом материале[1]. В коллекции центра около 50 речевых корпусов, содержащих сотни часов звучащей речи, а также современный компьютерный инструментарий для обработки звучащей речи и создания речевых баз данных. На сайте центра размещена карта мира, на которой обозначены региональные исследовательские центры, участвующие в создании лингвистических корпусов разного профиля. Их число постоянно растёт, и это свидетельствует об образовании особого профессионального сообщества — Linguistic & Speech database community (*рис. 1*).

Рис. 1. Карта языковых ресурсов и разработчиков LDC

Можно назвать и другие признаки перехода речевых корпусов из обязательной составляющей других речевых технологий в самостоятельное технологическое направление. Так, на рубеже веков в фокусе внимания разработчиков и других заинтересованных специалистов оказались вопросы **стандартизации** методов, представления данных, аннотаций и инструментария корпусных ресурсов. Начало широкому обсуждению этих проблем было положено выходом книги «The Handbook of Standards and Resources for Spoken Language Systems». Ed. Gibbon D., Moore R., Winski R., 1997 [4]. В области речевых корпусов долгое время образцом для разработчиков служил американский корпус TIMIT, (табл. 1), а также подробнее [5].

[1] Университетские исследователи получают значительные скидки при приобретении корпуса из коллекции LDC.

С начала XXI века по инициативе LDC проводятся регулярные рабочие совещания и конференции по разным вопросам создания лингвистических баз данных. Очень важным событием оказалась дискуссия о типах и средствах лингвистических аннотаций в корпусах разного профиля и целевого использования, которая была организована отделом аннотаций LDC при поддержке IRCS (Institute for Research in Cognitive Science), США в декабре 2001; материалы этого совещания до сих пор доступны на сайте LDC: [IRCS Workshop on Linguistic Databases [Dec 2001]. Актуальная проблематика речевых корпусов рассматривалась предварительно на страницах двух специальных выпусков журнала «Speech Communication» в начале 2001 г. [6]. Представленные здесь публикации заслуживают отдельного обсуждения, назовём лишь их тематическую рубрикацию:

- представление речевых данных, структура и содержание аннотаций;
- связи между аннотациями и сигналами;
- структура и организация баз данных;
- проблемы компьютерной разработки и использования РК;
- фундаментальные проблемы методологии исследований и разработок, относящихся к аннотированным РК.

Многие участники вышеупомянутой дискуссии затронули и целый ряд важных организационных вопросов, см. например, доклад [7] с показательным названием «Writing a Corpus Cookbook». Была отмечена, в частности, необходимость подготовки методического руководства и рекомендаций для оптимизации проектов по созданию лингвистических корпусов. Подчёркивалось также, что в рамках существующего и возрастающего разнообразия ресурсов, в том числе электронных, трудно получить информацию о том, какие ресурсы уже существуют и доступны, хотя это необходимо по практическим и этическим соображениям.

В качестве реакции на эти актуальные проблемы в 2002 г. был инициирован проект **OLAC** (the Open Language Archives Community, http://www.language-archives.org/, который имеет сервисную службу на сайте LDC. Цель проекта и портала — устранить разрыв между потенциальными пользователями, разработчиками и массой несвязанной информации о цифровых лингвистических ресурсах, накопленных к настоящему времени мировым сообществом.

OLAC — это пример международной кооперации, сообщества организаций и отдельных лиц, которые участвуют в создании всемирной виртуальной библиотеки языковых ресурсов путём:
- формирования согласованного мнения относительно наиболее успешных разработок и проектов по созданию цифровых архивов и языковых ресурсов;
- создания сети интерактивных лингвистических архивов и средств для их размещения в Интернете, поиска и доступа к ним.

Классификация речевых корпусов

Аннотированные речевые корпуса — важнейший компонент исследований в области звучащей речи. Сегодня они созданы и создаются для большого количества языков, научных дисциплин и технологий. Опыт, накопленный в области их разработки и использования, позволяет выделить ряд признаков, которые могут быть положены в основу классификации речевых баз данных и учитываться при проектировании нового РК. Укажем наиболее важные характеристики (см. также [8]):

- ***целевое использование корпуса:*** специализированные, технологические, общие (репрезентативные), учебно-иллюстративные;

- ***тип речевого материала:*** дискретная речь, непрерывная речь-чтение, спонтанная речь, специальные и естественные диалоги;

- ***тип текстового материала:*** списки слов/слогов, наборы отдельных предложений, связные тексты; монотематические или политематические;

- ***тип речевого сигнала***: лабораторная речь, офисная речь, публичная речь, телефонная речь (обычная или через мобильный телефон); радио-, теле-речь, речь в условиях естественной внешней среды, иноязычная (акцентная) речь и т.д.;

- ***тип информации, ассоциированной с речевым сигналом (аннотации):*** орфографическая запись, фонемная / фонетическая транскрипция, просодическая транскрипция, акустико-фонетическая разметка сигнала: «событийная», сегментная, просодическая, включение других типов лингвистических аннотаций и комментариев, например, об индивидуальных особенностях произношения говорящего или эмоциональной окраске речевых фрагментов;

- ***тип статистической балансировки*** звуковых единиц языка: равномерная, репрезентативная, по специальной статистической схеме;

- ***наличие и типы дополнительной сигнальной информации,*** включённой в корпус наряду с речевым сигналом: простые, мультимодальные и специальные корпуса.

Горячие точки в технологии речевых корпусов

Судя по тематике и результатам текущих конференций и рабочих совещаний, горячими точками в технологическом прогрессе РК до сих пор являются финансовое обеспечение, необходимость кооперативных усилий, обеспечение общедоступности и многопрофильности речевых корпусов, стандартизация аннотаций и информационной структуры РК; разработка компьютерного инструментария для накопления, обработки, верификации речевых баз данных с активным привлечением возможностей сети Интернет. Кроме этих «канонических» проблем, специалисты обращают внимание на необходимость создания больших, разнообразных, информационно «богатых» (многоуровневых) речевых корпусов. Отмечается, в частности, растущая потребность в просодически размеченных и аннотированных РК, корпусах эмотивной и социально дифференцированной речи. Постоянно растёт интерес к мультимодальной коммуникации и мультимодальным корпусам и базам данных с компонентом звучащей речи [9], в этом году готовится специальный выпуск журнала LRE (Language Resources and Evaluation), посвящённый проблемам моделирования мультимодальной межличностной коммуникации.

Широкие, «коммуникативные» корпуса имеют не только технологическое, но и образовательное, общекультурное значение. В связи с этим нельзя не отметить, что ещё в 90-е годы XX в. корпорация Kay Elemetrics (США-Канада, www.kayelemetrics.com), мировой лидер в производстве аналогового и компьютерного инструментария для речевых исследований, начала разработку фонетических баз данных учебно-иллюстративного профиля с аудио- и видеокомпонентами, создав коллекцию таких баз для 50 разных языков. Приведём в качестве иллюстрации фрагмент подобной базы для одного из африканских языков (*рис. 2*).

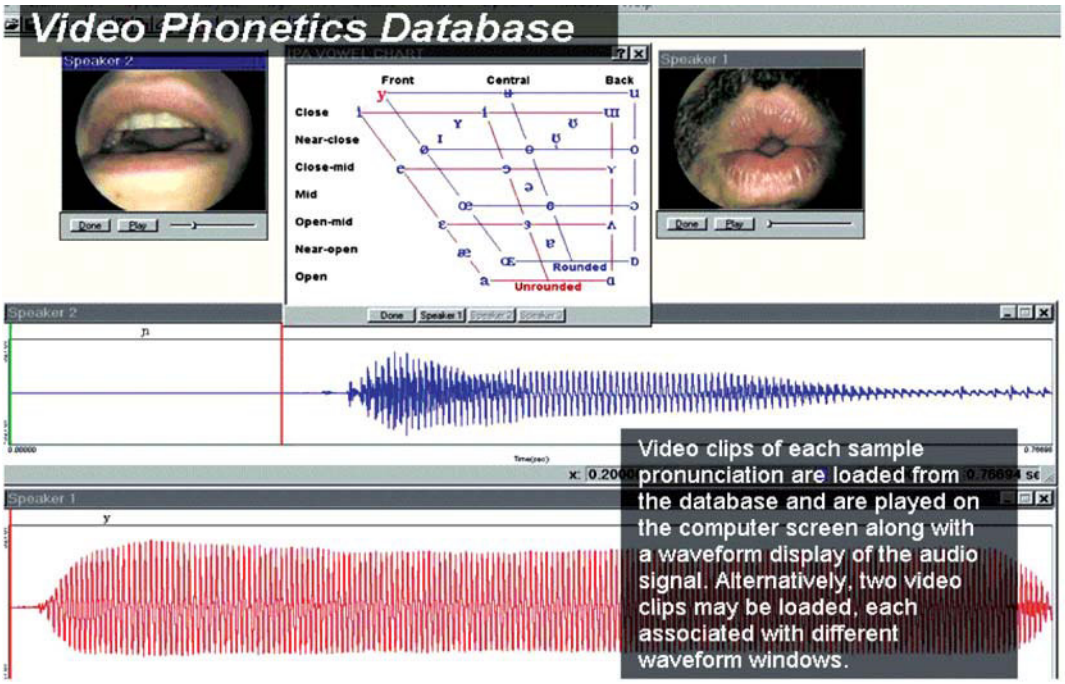

Рис. 2

Речевые корпуса для русского языка[2]

Как правило, речевые базы данных моноязычны. Речевые корпуса созданы не только для всех технологически важных языков (американского английского, немецкого, японского, китайского и др.), но и для большинства официальных языков Европейского Союза: для британского и шотландского вариантов английского языка, голландского, датского, шведского, немецкого, французского, итальянского, испанского; есть также несколько многоязычных корпусов. В результате осуществления программы Copernicus ELRA распространяет речевые корпуса и для языков Восточной Европы (польский, болгарский, эстонский, румынский и венгерский). На сайте Европейской Ассоциации в Интернете можно найти предложения и речевых корпусов для русского языка. Насколько нам известно, в их разработке принимала участие санкт-петербургская компания «Одитек» [10].

Речевой корпус ISABASE

В конце 90-х годов в Институте системного анализа (ИСА) РАН при участии специалистов речевой группы филологического ф-та МГУ был создан первый представительный речевой корпус для русского языка с разметкой речевых фрагментов на звуковые единицы. Корпус использовался не только в исследовательских целях, но и для построения автоматической системы распознавания дискретной речи [11]. Корпус моносигнальный, остальные характеристики см. ниже в таблице 2.

[2] В настоящей статье рассматриваются только те технологические речевые корпуса русского языка, которые описаны в специальных публикациях.

Таблица 2

Характеристики русского речевого корпуса ISABASE

Тип речевого материала		Дискретная речь	Дикторы/речевые фрагменты-предложения	Общий объём
Текстовый материал	**1**	Фонетически сбалансированный набор из 500 коротких предложений, монотематический	5 дикторов-мужчин и 4 диктора-женщины; 1863 фрагмента	4653 речевых фрагмента; 3713 слов
	2	Фонетически репрезентативный набор предложений, взятых из литературных текстов; политематический	15 дикторов-мужчин и 14 дикторов-женщин; 3280 фрагментов	
Типы аннотаций		Текст речевого фрагмента, фонетическ. транскрипция, ручная разметка сигнала на слова и фонемы	Транскрипционная система из 110 монофонов	

Речевой корпус RuSpeech

В 2000–2001 гг. в ИСА РАН по заказу корпорации Intel был создан также самый представительный на сегодняшний день речевой корпус русского языка **RuSpeech**, который может быть использован для разработки систем распознавания слитной русской речи [12, 13]. Общие характеристики корпуса приведены ниже в таблице 3.

Помимо самой речевой базы, важным результатом проекта ***Ruspeech*** стали отлаженная технология создания речевых корпусов и комплекс программных средств для обеспечения этой технологии. Среди последних можно отметить отладку автоматического транскриптора русской речи; создание инструментария для подготовки текстового материала с нужными фонетическими и статистическими характеристиками; создание автоматизированного рабочего места эксперта-фонетиста; программы пакетной записи дикторов; несколько программ для верификации результатов основных этапов разработки [12–15].

Укажем основные этапы проекта, которые полезно, на наш взгляд, иметь в виду потенциальным разработчикам РК:

- проектирование корпуса;
- подготовка текстового материала с возможной автоматизацией;
- подготовка фонетического обеспечения РК;
- разработка программного обеспечения для формирования речевого корпуса;
- подбор дикторского состава;
- организация записи и файлирования речевого материала;
- проверка качества записи;
- создание рабочего места эксперта-фонетиста и детальных инструкций по разметке и фонетической аннотации речевых сигналов;
- верификация аннотаций речевого материала, полученных автоматически;
- обработка результатов верификации;
- окончательное формирование и структурирование РК.

Резервы для расширения коллекции русских РК

Хотя число русских РК и баз данных с годами растёт, всё-таки это происходит очень медленно. Важным и чрезвычайно полезным резервом для будущих РК являются **фонотеки** русской звучащей речи, которые есть во многих центрах: научных (например, ИРЯ РАН им. В.В. Виноградова, ИРЯ им. А.С.Пушкина), образовательных (вузы) и культурных (музеи). Большая коллекция фонодокументов находится в Российском государственном архиве фонодокументов (РГАФД): более 200 000 единиц хранения, около 3,5 млн записей 1898–2001 гг., в том числе — восковые валики (фоновалики), оригиналы и копии 591 единиц; грампластинки, оригиналы и копии 135 000 единиц; матрицы, страховой фонд на граморигиналах 1420 единиц; тонфильмы, оригиналы и магнитные копии 1136 единиц; магнитные ленты, компакт-кассеты более 36 000 единиц; лазерные компакт-диски — более 900 единиц хранения.

Таблица 3

Дикторское и текстовое наполнение корпуса *Ruspeech*. Общие характеристики РК: моносигнальный, слитная речь, чтение

Общая характеристика	Тип речевого материала	Состав фрагментов	Дикторы/фрагменты
	Непрерывная речь; моносигнальный	50 часов записи; 30 CD, более 15 Gb; более 50 000 фрагментов-предложений	237 дикторов: 127 мужчин и 110 женщин разного возраста
Текстовый материал	1. Фонетически сбалансированный набор; политематический — нет	70 предложений, обеспечивающих полное (≥3 раз) монофонное покрытие	203 диктора: 111-м и 92-ж; каждое предложение произнесено всеми дикторами
	2. Фонетически репрезентативный (на аллофоном уровне) набор предложений, взятых из газетных и новостных текстов на интернет-сайтах; политематический	3060 предложений, обеспечивающих полное покрытие аллофонов из репрезентативного набора	203 диктора: 111-м и 92-ж по 180 предложений выборочно; каждое предложение произнесено 14 дикторами
		2000 фонетически разнообразных предложений	20 дикторов: 10-м и 10-ж по 200 предложений выборочно; каждое предложение произнесено одним диктором
Аннотации	Текст речевого фрагмента, каноническая и фактическая транскрипция, выверенная экспертами; данные о дикторе и эксперте-фонетисте — нет	Транскрипционная система из 114 монофонов	

Многие организации, обладающие фонотеками, в порядке самостоятельной инициативы проводят в настоящее время оцифровку имеющихся у них речевых материалов. Однако вопросы доступа и возможного использования этих материалов для широкого круга исследователей, и в том числе разработчиков РК, остаются открытыми и, к сожалению, даже не обсуждаются. Год русского языка, прошедший в 2007 г., никак на эту ситуацию не повлиял.

Вне пользовательской информационной зоны остаются и те РК, которые создаются по проектам, финансируемым государственными фондами РФФИ и РГНФ. Как правило, речевые базы, разрабатываемые в рамках этих проектов, где-то «исчезают» и известны узкому кругу учёных

и разработчиков по отдельным публикациям. Специалистам очевидна необходимость централизации и государственной поддержки в сфере создания, хранения, дистрибуции и обеспечения доступа к цифровым материалам русской звучащей речи.

Лучше обстоит дело с лингвистическими корпусами русского языка, важность которых трудно переоценить как для судьбы русского языка и культуры в целом, так и для научно-исследовательских работ разного профиля, хотя и здесь отмечается значительное отставание от США, Европы и Японии. С 2000 г. в России[3] ведутся работы по крупномасштабному проекту «***НАЦИОНАЛЬНЫЙ КОРПУС РУССКОГО ЯЗЫКА» (НКРЯ).*** Разработка корпуса осуществляется большой группой лингвистов из Москвы, Санкт-Петербурга и других городов России в рамках программы «Филология и информатика» РАН (с частичной поддержкой Российского гуманитарного научного фонда). В Интернете в свободном доступе открыт сайт *Национальный корпус русского языка,* объёмом более 140 млн слов. Поддержка сайта и поиск по корпусу осуществляются компанией «Яндекс», здесь же на сайте можно получить подробную информацию о задачах корпуса и его текущем состоянии, см. также [16]. Мы напомним лишь, что Корпус русского языка — это собрание грамматически размеченных русских текстов XIX–XXI вв. в электронной форме, удобной для автоматического поиска и научных исследований. Практически Корпус — это информационно-справочная система, основанная на собрании русских текстов в электронной форме.

В состав Корпуса входят тексты самых разных жанров, причём в сбалансированном объёме, — произведения художественной литературы, научные, научно-популярные, религиозные и иные сочинения, публицистика, производственно-технические, юридические и многие другие тексты. НКРЯ является максимально представительным отражением русского литературного языка во всём многообразии его письменных форм. Каждому слову и каждому тексту в Корпусе приписана лингвистическая аннотация (метатекстовая, грамматическая и семантическая) на основе специального стандарта, разработанного при участии ведущих российских лингвистов.

В 2006 г. в составе Корпуса появилось несколько новых составляющих: корпус **поэтических** текстов, снабжённых, помимо обычных аннотаций, морфологических и семантических, разметкой параметров стиха — рифмы, строфики, метрики, корпус **диалектных** текстов с разметкой специфических диалектных форм, а также особый подкорпус текстов **живой русской речи,** текстов **мультимедиа** (кинофильмов) и текстов **электронной коммуникации**.

Корпус предназначен для широкого круга пользователей: профессиональных лингвистов, преподавателей русского языка, журналистов, редакторов и издателей, школьников и студентов, иностранцев, изучающих русский язык. В то же время грамматически и семантически аннотированный корпус — это не только мощное средство для многоаспектного изучения русского языка, но и важный инструмент для создания и совершенствования компьютерных средств обработки русских текстов. В частности, для речевых технологий НКРЯ создаёт возможность составления различного рода словарей и статистических моделей языка разного уровня. Правда, к сожалению, сами по себе тексты, образующие базу информационной системы НКРЯ, доступны пока что только разработчикам Корпуса. То же самое относится и к записям звучащей русской речи, на основе которых создаётся подкорпус устных текстов. Для использования этих ценных материалов в сфере русских речевых технологий необходимы специальные соглашения относительно авторских прав. В заключение остаётся выразить надежду, что такие соглашения могут быть достигнуты.

[3] О русских языковых корпусах вообще см. [17], а также электронные публикации на сайте НКРЯ www.ruscorpora.ru.

Литература

1. *Hunt A., Black A.W.* Unit selection in a concatenative speech synthesis system using a large speech database // ICASSP-96. 1996, v. 1, pp. 373–376.

2. *Black, A., Zen, H., and Tokuda, K.* Statistical Parametric Synthesis, ICASSP 2007, Hawai.

3. *Клэтт Д.Х.* Основные результаты работ по проекту ARPA // Методы автоматического распознавания речи. Т. 2. М.: Мир, 1983. С. 333–360.

4. *Gibbon, D., Moore, R., Winski, R.* (Editors) Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, 1997.

5. *Кривнова О.Ф., Захаров Л.М., Строкин Г.С.* Речевые корпусы (опыт разработки и использование) // Труды семинара Диалог'2001 по компьютерной лингвистике и её приложениям. М., 2001.

6. Speech annotation and corpus tools // "Speech communication". Ed. S.Bird & J.Harrington, 2001, v.33, issue 1–2. http://www.ldc.upenn.edu/annotation/specom.html.

7. *Martin Wynne*. Writing a Corpus Cookbook, 2001, IRCS Workshop on Linguistic Databases [Dec 2001]. http://www.ldc.upenn.edu/annotation/ databases.html.

8. *Кривнова О.Ф.* Области применения речевых корпусов и опыт их разработки // Труды XVIII сессии Российского акустического общества РАО. Таганрог, 2006. С. 81–84.

9. LREC Workshops 2000–2008 on «Multimodal corpora: From Models of Natural Interaction to Systems».

10. *Викторов А.Б., Викторова К.О., Воронцова А.В.* и др. Речевые базы данных для задач автоматического распознавания речи и верификации говорящего // Современные речевые технологии. Сб. трудов IX сессии Российского акустического общества. М., 1999.

11. *Богданов Д.С., Кривнова О.Ф., Подрабинович А.Я., Фарсобина В.В.* База речевых фрагментов русского языка ISABASE // Сб. «Интеллектуальные технологии ввода и обработки информации». М., Эдиториал УРСС, 1998.

12. *Богданов Д.С., Брухтий А.В., Кривнова О.Ф., Подрабинович А.Я., Строкин Г.С.* Технология формирования речевых баз данных // Сб. «Организационное управление и искусственный интеллект». М.: Эдиториал УРСС, 2003.

13. *Arlazarov V.L., Bogdanov D.S. Krivnova O. F., Podrabinovitch A. Ya.* Creation of Russian Speech Databases: Design, Processing, Development Tools // International Conference SPECOM'2004. Proceedings. S-Pb. Russia, 2004. Pp: 650–656.

14. *Кривнова О.Ф.* Фонетическое обеспечение для построения речевого корпуса // Труды XIII сессии Российского акустического общества РАО. М., 2003.

15. *Захаров Л.М., Кривнова О.Ф., Строкин Г.С.* Подбор текстового материала и статистический инструментарий для создания речевых корпусов // Труды XI сессии РАО. М., 2001.

16. Национальный корпус русского языка: 2000-2005. Результаты и перспективы. М.: Индрик, 2005; www.ruscorpora.ru

17. *Резникова Т.И., Копотев М.В.* Лингвистически аннотированные корпуса русского языка (обзор общедоступных ресурсов) // Национальный корпус русского языка: 2000–2005. Результаты и перспективы. М.: Индрик, 2005. С. 31–61.

Кривнова Ольга Фёдоровна —

окончила филологический факультет Московского государственного университета им. М.В. Ломоносова по специальности «структурная и прикладная лингвистика». Работает на филологическом факультете МГУ им. М.В. в должности старшего научного сотрудника. Доктор филологических наук, имеет звания «Старший научный сотрудник», «Заслуженный научный сотрудник Московского университета». Член Фонетической комиссии при ОЛЯ РАН, секции «Акустика речи» Российского акустического общества, редколлегии периодического издания «Проблемы фонетики» ИРЯ РАН , редколлегии журнала «Речевые технологии». Имеет более 100 печатных работ.