Основные тенденции развития многоязычной корпусной лингвистики

(Часть 1)

Потапова Р.К.,
доктор филологических наук, профессор

Корпусная лингвистика (КЛ) — это многоаспектный раздел прикладной лингвистики, «обслуживающий» целый ряд отраслей теории и практики вербальной коммуникации на базе новых информационных технологий.

Термин «Корпусная лингвистика» (*Corpus Linguistics)* отражает характер объекта, с которым имеет дело данная эмпирическая область исследований. Сюда относятся, прежде всего, тексты на естественных языках в машиночитаемом формате, образующие массивы, коллекции, а также специально оформленные корпусы (*Corpora)* [Леонтьева, 2006].

Текстовые корпусы начали создавать ещё в 60–70 гг. XX века, т.е. корпусная лингвистика (КЛ) существует уже более 30 лет. За это время созданы десятки банков текстовых данных в первую очередь для английского, а затем и для других европейских языков и языковых пар; на основе текстовых корпусов (ТК) созданы сотни словарей (*corpus-based dictionaries).*

COBUILD — первый словарь, основанный на корпусных данных, – вышел в 1987 г. и был принят как стандарт с учётом требований теории и практики. С тех пор все современные словари, особенно предназначенные для изучения неродного языка, основываются на материале имеющихся и всё время пополняющихся ТК.

В 1995 г. вышел в свет Collins COBUILD English Dictionary (CCED), в котором отразились существенные изменения в языке относительно первых публикаций COBUILD: некоторые слова и значения выпали, но в то же время в него вошли американизмы и ряд технических терминов.

ТК создаются не только для основных европейских языков (английского, французского, немецкого), но также и для менее распространённых языков (шведского, норвежского, финского). В настоящее время существует значительное число ТК: разнотипных и разноразмерных, одно-, дву- и многоязычных, для письменного и устного вариантов языка. Создаются также параллельные корпусы: англо-норвежский, чешско-английский, словацко-

русский, словацко-хорватский и др. Более того, современные грамматики и словари формируются только на основе корпусной поддержки.

Согласно стандарту Британского национального корпуса (БНК) были созданы текстовые корпусы для многих европейских языков. Характеристика «национальный», служащая для конкретизации варианта языка, описываемого в корпусе, стала применяться для обозначения представительного ТК любого языка.

Как правило, национальный корпус — это отдельная комплексная система, образование и ведение которой требует больших трудозатрат как со стороны лингвистов, так и с учётом программного обеспечения. Современные ТК-системы выбирают и проводят определённую лингвистическую политику и используют для этого последние достижения информационных технологий.

Назовём только некоторые ТК (наиболее известные из них можно найти в Интернете): COBUILD (основан в 1980 г., руководитель Джон Синклер); British National Corpus (BNC), или БНК; Bank of English (Банк английского языка); ALEX (банк английской и американской литературы по западной философии); БК (Брауновский корпус); ICE (International Corpus of English); Longman/Lancaster Corpus; London-Lund Corpora; OED (Oxford English Dictionary); CUMBRE Corpus (корпус современного испанского); Чешский национальный корпус; Словацкий национальный корпус; Китайский текстовый корпус.

В России первым опытом создания большого лингвистического корпуса был Машинный фонд русского языка [Андрющенко, 1989], целью которого было создание представительного корпуса с подкорпусами различных жанров и соответствующих программных средств, а также комплексная информатизация лингвистических исследований, включая создание грамматик и словарей. Несмотря на то, что в полном виде программа выполнена не была, удалось собрать коллекции текстов разного типа, перевести в машинный вид многие традиционные словари.

В настоящее время Фонд обслуживает внутренние задачи Института русского языка (ИРЯ) РАН: ведение Русского диалектологического атласа, создание автоматического конкорданса для текстов русского фольклора, политических текстов, текстов древнерусских источников XI–XVII вв. и др. Каждая из задач требует создания отдельного пакета программ. В состав Машинного фонда входит большое количество словарей: «Грамматический словарь русского языка» А.А. Зализняка, «Русский орфографический словарь», «Русский синтаксический словарь» Е.А. Золотовой и др. В Фонд вошли также коллекции русской художественной литературы (М.Ю. Лермонтов, Ф.М. Достоевский), коллекции русских поэтических текстов. Руководитель Фонда А.Я. Шайкевич самой важной научной задачей считает проведение дистрибутивно-статистического анализа текстов и создание объективного описания языка, используя языково-независимый метод формирования «естественных» классов [Shaikevich, 1997; Рахилина, Шаров, 2003].

В начале XXI века в России начата работа по созданию представительных корпусов для русского языка. Два проекта — БОКР (Большой корпус русского языка) и РС (Русский стандарт), которые должны были представить русский литературный язык во всех значимых жанрах и видах использования [Шаров, 2003], слились с «Корпусом ЦЛД — МЕУ» [Сичинава, 2002]. Последний создаётся с 2001 г. общественной организацией ЦЛД (Центр лингвистической документации, руководитель – В.А. Плунгян). Была создана Ассоциация «Национальный корпус русского языка», в которую вошла большая группа лингвистов Москвы, Санкт-Петербурга, Новосибирска и других научных цен-

тров России. Планируемый объём корпуса — 200 млн. слов. Подробнее об Ассоциации, её участниках и планах можно посмотреть в Интернете на странице www.ruscorpora.ru.

Кроме того, отдельные коллективы РФ продолжают свои работы по созданию специальных ТК [Корпусная лингвистика, 2003]. В Санкт-Петербургском университете проводятся регулярные конференции, посвящённые КЛ.

Текстовые корпусы (ТК) могут формироваться по разным основаниям: авторские, по жанрам, стилям, по дате источника, по научным направлениям и т.п. Создатели ТК должны определять, какие порции и пласты языка нужно в них представить, что зависит от конкретных задач и внешних условий (например, финансирования), а также от адресатов ТК [Шаров, 2003; Рыков, 2003].

Что касается источников формирования текстовых корпусов (ТК), то в настоящее время проблем не существует благодаря Интернету, технологиям автоматического чтения и сканирования, быстродействию компьютеров, практически неограниченным объёмам памяти и т.д.

С 1996 г. стал выходить Международный журнал по корпусной лингвистике, на страницах которого обсуждаются разные аспекты формирования и ведения текстовых массивов, описываются новые ТК, обсуждаются вопросы их аннотирования. ТК, снабжённые лингвистической информацией, называют размеченными или аннотированными. Чем богаче разметка (например, морфологическая, синтаксическая), тем большую ценность имеет корпус.

Так, в частности, в Национальном корпусе русского языка используется пять типов разметки: метатекстовая, морфологическая, акцентная, синтаксическая и семантическая. Две последние выполняются на небольшом фрагменте корпуса.

Недавно стал создаваться аннотированный корпус для русских текстов в ИППИ РАН [Богуславский, Григорьев и др., 2000], который состоит из нескольких подкорпусов. Тексты последних различаются уровнем аннотации:

— лемматизированные тексты, в которых для каждого слова указывается его основная форма и часть речи;

— тексты с морфологической информацией, в которых для каждого слова указывается его основная форма, часть речи и полный набор морфологических характеристик;

— тексты с синтаксической информацией, в которых для каждого слова указывается его основная форма, часть речи и морфологические характеристики, а для каждого предложения — его синтаксическая структура.

Выполняемая автоматически разметка корректируется лингвистами.

К 2000 г. создано не меньше 20 аннотированных корпусов для основных европейских языков. Из них, по крайней мере, три — с синтаксической информацией. Наиболее известны Perm Treebank, созданный в Пенсильванском университете в 1990 г. [Markus, Santorini, Marcinkiewicz, 1993], и создаваемый по его образцу Пражский банк деревьев зависимостей (PDT – Prague

Dependency Treebank). Эти работы постоянно освещаются в Пражском бюллетене по математической лингвистике [Bohmova, 2001; Hajicova, Pajas, Vesela, 2002].

PDT — это исследовательский проект Карлова университета в Праге. Схема аннотирования включает три уровня: морфологический, аналитический и тектограмматический. На первом из них проводятся стандартные для всех систем операции лемматизации и определения всех морфологических характеристик (используется примерно 3000 значений морфологических тэгов) для словоформ входного текста. На втором уровне строится поверхностная синтаксическая структура, называемая analytic tree structure (ATS): это промежуточное дерево зависимостей, в котором каждое слово и знак препинания представлены отдельными узлами с приписанными им характеристиками теньеровского типа (субъект, объект, адвербиал, атрибут). Перевод из линейных структур (с их скобочной записью) в древесную проводится полуавтоматически. Такой метод был испытан и отработан на трансформации деревьев составляющих английского языка из Пенсильванского банка в деревья тектограмматического уровня, принятые в PDT. Третий уровень строит тектограмматическую древесную структуру (Tectogrammatical Tree Structure — TGTS), представляющую собой глубинное синтаксическое дерево предложения. В нём в качестве узлов остаются только полнозначные слова; все функциональные слова «без собственного лексического значения» (предлоги, подчинительные союзы, знаки препинания и пр.) становятся атрибутами при оставшихся узлах. Полнозначные узлы «аннотируются» ролью в предложении (которая называется «функтором»). Функторов примерно 60: актант, пациенс, адресат, источник, эффект. Учитываются также разные типы пространственных, временных и иных признаков: средство, способ, степень, последствие, условие и др.

Большинство функторов приписывается вручную. Затем создаётся обучающийся модуль, который часть функторов строит автоматически, опираясь на правила и словарные данные, извлечённые из уже аннотированной части корпуса.

К 2002 г. из текстов текущей версии Чешского корпуса в 100 млн. слов проаннотировано в терминах ATS 100 тыс. предложений, средствами TGTS — 20 тыс. предложений; из них 2 тыс. предложений снабжены пометами о коммуникативной структуре (Topic-Focus Articulation — TFA). Последние работы чешских лингвистов обогащают глубинные синтаксические структуры ещё одним видом информации — введением кореферентных связей для личных и указательных местоимений [Kucova, Hajicova, 2004].

Данный уровневый подход к аннотированию текстовых корпусов принят, в основном, в русской и чешской школах КЛ. Он сближается с методами полного лингвистического анализа текстов в системах автоматического понимания текстов (АПТ). Вместе с тем он требует больших трудозатрат со стороны лингвистов, корректирующих результаты автоматических операций.

ТК — это источник различного типа знаний. Информация, содержащаяся в текстовых массивах, без лингвистической обработки не может быть использована. Для извлечения знания требуются мощные лингвистические технологии. Перед корпусной лингвистикой стоят те же проблемы, которые характерны для этапа анализа языковых ресурсов в системах автоматической обработки текстов:

а) сортировка и систематизация текстовых массивов;

б) сегментация текстов;

в) общелингвистический поверхностный анализ, или аннотирование, текстов;

г) внутренняя разметка: расстановка морфологических, синтаксических и семантических обозначений («тэгов»).

Чтобы быть полезным объектом для разных специалистов и чтобы предоставить лингвисту возможность выбрать или собрать нужный ему массив, КЛ систематизирует коллекции текстов — по эпохам, языкам, жанрам, стилям, тематике и т.п. Кодирование метаинформации о тексте документа и его внешних параметрах опирается в большой мере на уже разработанные технологии. Используются разные системы кодирования текстов (HTML, XML и особенно TEI: Text Encoding and Interchange); в частности, систематизация указанных выше русских корпусов основана на стандарте TEI. Этому, а также истории и полезным параметрам КЛ посвящены работы С.А. Шарова и указанная в них литература [Шаров, 2003].

Для системы COBUILD была разработана сегментация: «лёгкая» (выделяются заголовки и подзаголовки текстов) и «нежёсткая» (уточняются или снимаются различного рода пометы к текстам). Сегментация текста — процесс корпусного анализа, при котором части текста делятся сначала на предложения (или словосочетания), а затем вычленяются более мелкие единицы, например, обозначения дат, денежных сумм, названия компаний, адреса, номера телефонов и т.д. При этом синтаксический препроцессор объединяет их в группы непосредственных составляющих по заданным комбинациям признаков.

Грамматики в системе COBUILD построены по принципу *data-driven* («под управлением данных»). Данный принцип противопоставлен принципу *data-based*, когда лингвист задаёт грамматику интуитивно, а корпус используется для проверки её правильности и для подбора примеров. В грамматике *data-driven* существенна лексическая компонента: нет независимого выбора грамматических конструкций и подстановки в них лексем — они работают вместе, создавая определённое значение. Есть списки лексем, для которых характерна определённая «схема», например Vn that; V+C (verb + complement), V + O + A (Verb + Object + Adjunct).

Схемы в такой корпусной лингвистике не правила, а некое обобщение употреблений. В них не различаются синтаксис и лексика (нет такого формального автономного синтаксиса, категориями которого можно было бы манипулировать без обращения к значениям слов [Barlow, 1996]): VP [lose [POSS way]], VP [lose [REFL]], VP [let NP go], VP [let [REFL] go].

Схемы могут быть вложенными. Кроме того, они могут быть связаны с типом дискурса (с учётом включения говорящего и слушающего). В традиционной КЛ нет уровня автономного синтаксиса. Не различаются глубинный и поверхностный уровни синтаксиса. Не проводится различие между категориями «Лексика» и «Структура». Вместо этого имеется формальная часть «Схема — Значение», «Структура — Лексика». Соответственно, и поисковый аппарат в корпусах принимает структуры, состоящие частично из лексем, частично из «тэгов» (грамматических и других помет).

Специалисты, работающие в области корпусной лингвистики, как правило, руководствуются определённой концепцией. Так, например, по мнению В.Тойберта, КЛ противопоставляется когнитивной лингвистике [Teubert, 2001]. Для КЛ не существует «языка мысли». В.Тойберт отрицает всякие репрезентации, ментальные языки, атомы смысла и пр. как нечто нематериальное, символы, абстракции, которые нельзя интерпретировать. По его

мнению, ни в искусственном интеллекте (ИИ), ни в машинном переводе (МП), по сути, нет никакого когнитивного подхода.

КЛ не волнует истинность высказываний. Неважно также, что думает кто-то о реальной воде: слово «вода» означает то, чем и является вода. КЛ имеет дело с языком как социальным явлением. Значение — в словах и текстах. КЛ не интересуют значения изолированных слов вне релевантных для них контекстов. Цитата даёт больше, чем словарная дефиниция слова. Значение неотделимо от формы. Различие в значении всегда сопровождается различием в форме. Корпусный анализ может помочь определить образцы этих форм.

По мнению В.Тойберта, КЛ отказывается от всех теоретических достижений лингвистики после Ф. де Соссюра. В основном, это относится ко всем вариантам порождающей грамматики школы Н. Хомского и его последователей. Исключение составляет аппарат категориальной грамматики [Teubert, 2001]. Универсальная грамматика описывает только ядро языка и ничего не говорит о периферийных зонах, тогда как исследователи языка и изучающие язык нуждаются именно в конкретном материале разных синтаксических конструкций [Barlow, 1996].

Поскольку КЛ интересует не отдельное слово, а текстовые сегменты, разница между лингвистическими и энциклопедическими знаниями размывается. Так, если немецкое слово Machtergreifung означает просто захват власти какой-то группой, ранее исключённой из политической жизни, своими силами, недемократически, то сегмент braune Machtergreifung im Jahre 1933 безоговорочно означает захват власти нацистами. Объясняется это тем, что часто в разных контекстах они заменяли друг друга, были парафразами или анафорически связанными сегментами. Энциклопедическое знание — это не что иное, как дискурсивное знание. Нет значения вне языка, вне дискурса.

КЛ связана с проблемами автоматической обработки текста и, конкретно, с системами автоматического распознавания и понимания текстов [Потапова, 1997; Потапова, 2005] по нескольким признакам:

а) системы АПТ нужны именно для работы с большими массивами текстов. Чтобы добиться каких-то полезных результатов в работающей системе, необходимо знать и учитывать все свойства этих новых для лингвистики объектов — текста как целого и массива текстов. КЛ формирует, исследует и описывает их как информационный ресурс;

б) технологии и приёмы первичной обработки «сырых», непрепарированных текстов в прикладных системах (машинный перевод и другие системы АПТ) во многом совпадают с теми, которые приняты или отрабатываются в КЛ. Так, чтобы создать параллельный корпус, нужны алгоритмы и программы сегментирования текста на такие (значимые) единицы, которые могут быть сопоставлены друг другу;

в) массивы КЛ — это надёжный источник формирования словарей, в том числе двуязычных, и выуживания информации, которую надо включать в словарную статью (иллюстрации значений слова, сведения об актантной структуре слова и др.); это источник создания конкордансов, словников, тезаурусов и других инструментов, необходимых для автоматического анализа произвольных текстов. Составление словарей — одно-, дву- или многоязычных — должно подтверждаться массивами КЛ, если не полностью базироваться на них [Леонтьева, 2006].

Тем самым КЛ не исключает, а дополняет традиционную лингвистику, становится опорой общей лексикографии. Ведь лексикография работает не только с простыми единица-

ми и их контекстом, но и с большими текстовыми сегментами, единицы которых определены на лексическом и синтаксическом, включая порядок слов, уровнях (многословные единицы, термины, коллокации, обороты). Традиционная лингвистика всё больше нуждается в более крупных, чем слово, единицах и в обосновании их выделения обращением к КЛ; она тяготеет к изучению семантической связности (lexical solidarities, collocations, set phrases, valencies, case roles, thematic roles, semantic frames and scripts). КЛ проясняет понятие текстового сегмента эвристическим определением семантической связности: совместной встречаемостью схем (цепочек), которые тем самым связаны определёнными семантическими отношениями.

Статистика совместной встречаемости и явное выражение шаблонов (комбинации квантитативных и категориальных признаков) позволяют изучать «размытые» значения (fuzzy meanings). КЛ допускает втягивать пользователя в дискурс и включать его определения в универсум цитат и контекстов [International Journal of Corpus Linguistics 1996 и др.].

Эмпирической базой многоязычной корпусной лингвистики служит (виртуальный) массив всех текстов, когда-либо переведённых на другой язык, вместе со своими переводами. Теоретическая основа та же, что и для одноязычных корпусов, т.е. значением текстовой единицы считается парафраза, а полное значение текстового сегмента в этом дискурсивном универсуме заключено в истории (сумме) всех переводных эквивалентов данного сегмента.

Создание параллельных и многоязычных корпусов столкнулось с трудной задачей «выравнивания», т.е. разбиения параллельных текстов на единицы, которые можно сопоставить друг с другом.

Большинство программ выравнивания в параллельных корпусах основывается на том, что в переводе сохраняются те же границы предложений и абзацев, что и в исходном тексте. В действительности же разные типы текстов требуют перестановки или сокращения (например, в юридических текстах) числа предложений. Пословные соответствия (предлог — отсутствие предлога, падеж — предложная конструкция) составляют незначительное число. Минимальные единицы перевода могут состоять из одного слова или нескольких слов, переводимых как целое, а не пословно. Переводные эквиваленты соответствуют текстовым сегментам одноязычного корпуса. Значение единицы перевода содержится в её переводных эквивалентах на другие языки. Идентификация единиц перевода требует интерпретации: единый это эквивалент или комбинация нескольких. Текстовый сегмент является единицей перевода по отношению только к тем языкам, в которых он переводится как единое целое. Неоднозначные единицы перевода имеют столько значений, сколько есть несинонимичных переводных эквивалентов. Данная единица перевода языка A может иметь два несинонимичных эквивалента в языке B и три — в языке C. Объявить какие-то эквиваленты синонимами — это акт интерпретации; сначала надо понять текст, а это компьютерам недоступно. Практическое использование корпусной лингвистики — помощь переводчику путём обработки параллельных массивов. Последние — это хранилища переводов. Использовать их гораздо более эффективно, чем традиционные двуязычные словари, особенно

если в массиве учтены жанр и тип текстов: выбирается тот эквивалент, контекстная проекция которого больше всего совпадает с профилем текстового сегмента.

Анализ «по образцу», или прецедентный анализ, важен не только для систем МП, как отмечалось ещё в ранних работах по МП, но и как серьёзное подспорье при анализе свободных текстов. И всё же проблема формирования параллельных корпусов достаточно трудна — и не только содержательно, но и чисто технически. С одной стороны, нужно сделать эксплицитной всю релевантную информацию. С другой стороны, текст, отягощённый тэгами, становится нечитабельным. Любые изменения в размеченном корпусе — всегда проблема.

Многие сторонники КЛ считают, что для обработки многоязычных массивов текстов продуктивно использовать языково-независимые подходы [Greenstette, Segond, 1997]. В RXRC (Ranc Xerox Research Centre) создано несколько средств АОТ, работающих на основе автоматов с конечным числом состояний и трансдукторов (The transducer is a finite-state machine which consumes input while producing output). Эти простые методы обработки оказались применимы к очень большому количеству лингвистических структур.

Разработанные средства были использованы в нескольких прикладных задачах: задаче извлечения терминологии (information extraction), в системе помощи переводчику и в информационном поиске (cross-language information retrieval). Технология автоматов с конечным числом состояний имеет много достоинств: это хорошо изученные механизмы, поддающиеся разным математическим операциям, их можно по-разному комбинировать, вставлять в другие процедуры и т.д. Правила трансформаций могут включать контекст, тем самым не требуя специальных программных решений. Модульность и возможность включать контекстные условия в структуру данных позволяют быстро приспосабливать подобные пакеты (suits) АОТ к другим языкам. Пакеты включают языково-независимые правила сегментации (tokenizer), морфологические анализаторы, программы построения гипотез для неузнанных слов, программы приписывания частей речи (POS: part-of-speech taggers) и программы сборки именных групп (noun-phrases extractors). Такие пакеты созданы в RXRC для семи европейских языков, готовятся ещё для семи (русского, чешского, венгерского и др.).

Главное в подходе RXRC — разработка надёжных и всё более мощных технических решений, применимых к любым массивам текстов на естественном языке.

В настоящее время результаты корпусных исследований находят основное практическое применение в создании больших контекстно-ориентированных тезаурусов, которые увеличивают семантическую силу при работе систем информационного поиска. Так, в системе ACRONYM (Automated Collocational Retrieval of «Nyms») собираются концептуально родственные единицы, называемые Nyms («нимы», по аналогии с синонимами и др.) [Collier, Pacey, Renouf, 1998]. При этом не проводится никакая предварительная лингвистическая разметка (считается, что это слишком «дорогой» процесс на очень больших массивах), кроме перевода числовых цепочек в обобщённые категории. Проводится кластерный анализ, вычисляется мера подобия соответственно правых и левых контекстов для выделенных единиц (слов и словосочетаний), учитывается частота появления сходных контекстов и т. п. Сначала собираются группы родственных слов (нимов) первого порядка, что уже может хорошо работать для информационного поиска, затем рядом уточняющих процедур строятся нимы второго порядка, которые должны удовлетворить и лингвистов. Приведём пример построенного в системе ACRONYM списка нетривиальных «родственников» для четырёх английских слов:

Таблица 1. Пример организации данных в системе ACRONYM

Node	Nyms
Key	crucial important vital significant essential main fundamental major strategic specific
Medicine	medical medicines sciences mathematics biology science chemistry psychology physics clinical
Pretty	fairly quite incredibly extremely terribly really nice extraordinarily lovely sexy
Testing	tests test tested assessment monitoring screening research rigorous clinical curriculum

Таким образом, текстовый корпус — это особый, совершенно новый тип словесного единства. Можно выделить четыре базовых качества, делающих собрание текстов корпусом [Рыков, 2003]:
– расположение на магнитном носителе;
– процедуры отбора материала, обеспечивающие его репрезентативность;
– единство разметки на носителе;
– конечный размер.

Возможно формирование не только универсальных, т.е. представительных с учётом разных жанров для всего языка, но и специализированных (для каких-то задач) корпусов текстов (к их числу относятся КТ звучащей речи).

Создание устно-речевых баз данных (УРБД) является на сегодняшний день первоочередной задачей в свете актуальности проблемы автоматизации процессов распознавания и понимания речи, идентификации говорящего по голосу и речи, синтеза речи, устного перевода. В современном мире быстро развивающихся информационных технологий, когда «умные» дома с управляемыми голосом приборами стали реальностью, необходим «строительный материал» для подобных систем. Одними из таких «кирпичиков» и являются фонетические базы данных, или УРБД. Формирование репрезентативных УРБД является одним из условий успешного решения прикладных задач.

Формирование УРБД многоцелевого назначения применительно к различным языкам мира является одной из приоритетных задач современного речеведения [Потапова, 1997].

Подавляющее большинство конструируемых сегодня автоматизированных систем, работающих со звучащей речью, так или иначе используют устно-речевые базы данных. В частности, УРБД находят применение там, где используются вероятностные и статистические методы анализа и синтеза речевого сигнала. В первую очередь здесь следует упомянуть системы автоматического распознавания и синтеза речи, идентификации и верификации говорящего по голосу и речи, идентификации психофизического и эмоционального состояния говорящего по речи, а также обучающие системы. Далее, УРБД составляют основу автоматизированных систем, в задачи которых входит сбор и хранение речевых сообщений, поиск и выдача записанных речевых сообщений по запросу (например, автоматизированные системы приёма голосовых сообщений в колл-центрах, комплексы для тестирования каналов связи). В ряде других случаев использование УРБД, не будучи строго необходимым технически, оказывается разумной альтернативой разработке сложных процедурных решений.

Как правило, УРБД содержат большие объёмы численной информации, трудно поддающейся автоматическому структурированию и сжатию. В то же время в силу специфики систем, в которых применяются УРБД, в большинстве случаев эта информация должна быть доступна для обработки в режимах, близких к режиму реального времени. Поэтому структура УРБД должна обеспечивать максимальное быстродействие системы при разумной ресурсоёмкости. По причине большого объёма информации изменение, а следовательно, и оптимизация структуры действующей УРБД обычно является технически трудновыполнимой и крайне нежелательной операцией. С учётом многообразия задач, для решения которых применяются УРБД, это означает, что её структура должна быть универсальной и, как следствие, максимально простой.

При разработке УРБД неминуемо встаёт проблема выбора системы управления базами данных (СУБД) [Белолипецкий, Буря, 2004]. Здесь возможны следующие варианты: выбрать существующую хорошо зарекомендовавшую себя СУБД из числа присутствующих на рынке информационных технологий или разработать свою СУБД специально для этой задачи.

При выборе из возможных вариантов разработчики руководствуются обычно следующими требованиями к СУБД, на которой реализуется речевая база данных:

— СУБД должна осуществлять удобное хранение как обычного текста, так и больших бинарных данных (BLOB). Под удобством понимается простота программного интерфейса для извлечения и записи данных;

— СУБД должна поддерживать хранение данных большого объёма (и большого суммарного объёма). Объём речевой БД может в несколько раз превышать объём доступной оперативной памяти;

— СУБД должно обеспечивать достаточно быстрый доступ к данным в режиме чтения. Это требование вступает в противоречие с предыдущим требованием, поэтому потребуется выработать некоторый компромисс или определить, какое из этих требований является приоритетным. Следует также учитывать специфику работы с речевой БД. Наиболее распространённым является сценарий последовательного перебора всех речевых образов (т.е. их последовательного чтения из БД), поэтому обширное кэширование данных не приведёт к значительному росту производительности;

— СУБД должна иметь средства для различных статистических расчётов по текущему состоянию речевой БД;

— СУБД должна иметь средства для импорта и экспорта речевых бинарных данных (а также сопутствующих им текстовой информации);

— СУБД должна предоставлять возможность как однократного полнообъёмного прочтения бинарного речевого образа, так и поэтапного (постраничного) чтения. Это требование возникает из-за различных потребностей алгоритмов обработки данных, алгоритмов распознавания и процедур импорта/экспорта. В реляционных СУБД существует подобный механизм, позволяющий указывать оптимизатору запросов, требуется ли последовательное получение результатов запроса и однократное (спецификаторы FIRST_ROWS, ALL_ROWS);

— СУБД должна позволять получить речевые данные на внешнем носителе информации в стандартном формате звукозаписи для возможности использования большинства программ обработки звука. Это требование можно обеспечить либо постоянным хранением бинарных данных в файлах со стандартным форматом (наподобие типа данных BFILE в Oracle), либо с помощью развитых средств импорта/экспорта.

Желательно также предусмотреть средства, облегчающие (автоматизирующие) пакетный запуск алгоритмов различных видов обработки речевых данных.

Существующие реляционные СУБД мало подходят для хранения речевых данных; наиболее удобный механизм реляционных СУБД, позволяющий хранить речевые данные, – тип данных BLOB. К сожалению, использование этого типа данных для хранения речевых данных сопряжено с рядом проблем. Тип данных BLOB рассматривается разработчиками реляционных СУБД как дополнительный или даже необязательный. Вследствие этого операции для работы с ним малочисленны и неоптимизированны. Во многих реляционных СУБД нет таких операций для работы с типом данных BLOB, как вставка и удаление части данных, есть только полная перезапись и обнуление. Уже одно то, что в языке SQL нет никаких операций для работы с типом данных BLOB, показывает, насколько затруднено его использование.

У ряда реляционных СУБД есть возможность хранить данные во внешних файлах (например, тип данных BFILE в СУБД Oracle), что позволяет использовать средства файловой системы для работы с данными. Но и это не панацея. С тем же успехом можно и не пользоваться СУБД, а хранить речевые данные просто в файлах.

Несколько лучше отвечают требованиям речевой базы данных объектно-ориентированные СУБД (ООСУБД). Особенно заманчиво использовать для подобной цели ООСУБД типа Jasmine, ориентированные на разработку мультимедийных приложений. В этом случае не возникает проблемы с надлежащим типом данных, поскольку ООСУБД так или иначе предоставляют возможность создать свой тип данных (класс), для которого можно задать формат хранимых данных, описать нужные операции по обработке данных. Также мультимедийные ООСУБД работают с большими объёмами данных намного эффективнее реляционных.

Интерес к созданию корпусов звучащей речи был в значительной степени инициирован разработками в области автоматического распознавания речи, где исследователям приходится сталкиваться с акустической вариативностью звуковых единиц языка, которая имеет разнообразные причины: от системной контекстной вариативности, обусловленной коартикуляцией, до психофизиологического и эмоционального состояния говорящего, а также технических характеристик микрофона, который используется при записи речевого материала [Кривнова, Захаров, Строкин, 2001; Кривнова, 2008][1].

Первые речевые корпусы появились в середине 80-х годов в США, где их разработка финансировалась прежде всего Министерством обороны. При поддержке этого ведомства были созданы: TI-DIGITS корпус (1984 г.) для тестирования систем распознавания изолированных цифр и цифровых последовательностей; Road Rally для анализа и распознавания ключевых слов (word spotting) и King Corpus для систем идентификации говорящего (speaker recognition). В рамках государственной программы развития лингвистических технологий, известной как ARPA/DARPA (Advanced Research Projects Agency), это же министерство финансировало создание уже упоминавшегося выше корпуса TIMIT, который послужил прототипом для многих других речевых баз данных. При этой же финансовой поддержке были разработаны специализированные речевые корпусы Resourse Management

[1] Подробнее об УРБД для русской речи см. работы О. Ф. Кривновой и её соавторов.

(RM) и Wall Street Journal (WSJ) для исследований в области распознавания слитной речи, а также Air Travel Information Service (ATIS) для исследования спонтанной речи и понимания естественного языка в диалоговых системах [Кривнова, Захаров, Строкин, 2001; Кривнова, 2008].

Накопленный к концу 80-х годов опыт показал, что создание представительных речевых корпусов требует кооперативных усилий исследовательских институтов, индустриальных компаний и государственных спонсоров. Финансовые и временные затраты на разработку высококачественных ресурсов оказались очень велики. Эксперты отметили, что дорогостоящие, но необходимые для развития информационных технологий ресурсы не должны разрабатываться для какой-то одной специальной системы или задачи [Godfrey & Zampolli, 1997]. Они пришли к выводу, что ресурсы должны обеспечивать возможность их многократного использования разными пользователями, т. е. быть общедоступными, и более чем для одной цели, т. е. быть многофункциональными. В связи с этими требованиями возникла проблема стандартизации лингвистических описаний, согласования форматов представления информации в разных видах лингвистических ресурсов и их типологии (подробнее см. [Gibbon et al., 1997]).

УРБД разрабатываются для решения конкретной задачи. Круг возможных применений велик, однако конкретная задача задаёт непосредственные характеристики базы.

В 1991 году в США был создан лингвистический консорциум (LDC – Linguistic Data Consortium), который поддерживает создание новых языковых корпусов и распространяет ресурсы, полученные из разных источников. В частности, в настоящее время LDC предлагает речевые корпусы, которые в совокупности содержат многие сотни часов звучащей речи. Технологический центр в штате Орегон (CSLU – Center for Spoken Language Understanding) коллекционирует, аннотирует и распространяет телефонные речевые корпусы. Активность Центра поддерживается индустриальными спонсорами. Собранные корпусы доступны университетам по всему миру бесплатно. Этот центр располагает также многоязычным корпусом для оценки алгоритмов идентификации языка, который состоит из фрагментов спонтанной речи на одиннадцати разных языках мира. В 1995 году координационный центр лингвистических ресурсов (ELRA – European Language Resources Assosiation) был образован и в Европе (более подробные сведения об истории создания и задачах этой ассоциации можно найти, например, в обзорных статьях [Mariani, 1996; Teubert, 1996]). В распоряжении этого центра находятся речевые корпусы для большинства официальных языков Европейского союза: для британского и шотландского вариантов английского языка, голландского, датского, шведского, немецкого, французского, итальянского, испанского, – а также несколько многоязычных корпусов. В настоящее время в результате осуществления программы Copernicus ELRA распространяет также речевые корпусы 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.

Сравнение различных акустических баз данных позволяет сформулировать некоторые обязательные требования к современной фонетической базе данных, предназначенной для фундаментальных и прикладных исследований. УРБД для прикладных исследований, в частности, в области синтеза и распознавания речи, должны обеспечивать решение следующих задач [Скрелин, Щербаков, 2003]:

- Внесение в УРБД звуковых эталонов — оцифрованных записей речи нормативных дикторов в разных стилях речи, от спонтанной речи и чтения текстов, полученных на основе её расшифровок, до чтения списка слов. Другими словами, в БД необходимо включить звуковой материал, представляющий максимальную вариативность реализа-

ции языковых единиц (фонем и интонационных конструкций) в различных условиях речевой деятельности человека.

- Внесение сегментной информации и подробного фонетического описания включаемых звуковых образцов, поскольку необходимо снабдить этот материал подробным описанием: адресами границ звуков и интонационных единиц, словоформ и слогов (так как существуют различные методики распознавания и синтеза речи с учётом базовых единиц), а также фонемной и подробной фонетической транскрипцией.
- Обеспечение эффективного выполнения запросов к содержимому УРБД для поиска нужных звуковых фрагментов по их транскрипционным описаниям и указанным в описаниях признакам.

Недостаточная проработка реализации любого из вышеперечисленных пунктов существенно снижает ценность УРБД в целом [Скрелин, Щербаков, 2003]. В реальности же для создания УРБД можно использовать любой ПК, оснащённый звуковой платой, совместимой с SB16. Производят запись речевого материала, для чего могут либо приглашать дикторов и производить запись в лабораторных условиях, либо собирать материал из широко доступных источников, например, теле- и радиотрансляций, вещаний в интернете и т.п. Для записи Интернет-трансляций нужна программа, которая может выполнять функции магнитофона, и программа, поддерживающая и воспроизводящая формат потокового аудио из интернета. Например, CoolEdit 1.0, которая, выполняя функции записи, является одновременно и звуковым редактором, и заменяет собой Windows Media Player, RealPlayer и т.д. Чтобы облегчить сбор и хранение данных УРБД, разрабатывается специальная оболочка. Она представляет собой отдельную программу-приложение, которая обладает возможностями: записи/воспроизведения фрагментов; хранения информации о фрагменте; хранения информации о дикторе (если такая информация нужна); поиска информации по различным параметрам.

После записи речевого материала, ввода речевого материала в компьютер (оцифровки) и сохранения его, эксперт-фонетист производит транскрибирование материала; файл транскрипции имеет, как правило, формат txt. Затем эксперт-акустик производит сегментацию материала, сохранённого в файлах форматов wav и txt, с его последующим сохранением в две разные папки, поименованные соответственно WAVE и TEXT. Эксперт-фонетист создаёт правила перехода «звук–буква», причём звуки представлены специальным алфавитом, варианты которого создаются для каждого языка. В настоящее время одним из таких стандартов можно назвать фонетический алфавит Sampa (Speech Assessment Methods Phonetic Alphabet). Он представляет собой Международный фонетический алфавит, записанный символами ASCII с соответствующими изменениями под конкретный язык.

Некоторые речевые базы данных для русского языка создавалась в рамках европейских проектов SpeechDat(II) и SpeechDat(E) [http://www.auditech.ru]. Целью проектов, объединенных названием SpeechDat, является создание речевых баз данных в странах Европы посредством записи речи в реальных условиях через телефонный канал стандарта ISDN. Базы данных призваны служить общим ресурсом для 20 европейских языков и диалектов и способствовать разработке общих систем телесервиса.

В проектах SpeechDat, профинансированных Европейским союзом, были представлены крупнейшие промышленные и академические организации. Все базы данных, созданные в рамках этих проектов, имеют стандартный дизайн и прошли все этапы валидации.

Созданные в рамках проектов SpeechDat речевые базы данных удовлетворяют следующим требованиям:

- охватывают фонетически репрезентативные слова, слова-команды, словосочетания, числа, цифры, числовые последовательности, фонетически репрезентативные предложения;
- представляют различные стили произнесения (команды, речь-чтение и спонтанная речь);
- фиксируют окружающую акустическую обстановку;
- пригодны для разработки и обучения надёжных систем распознавания речи для телесервисов.

В речевой базе данных SpeechDat(II) представлено 48, а в базе данных SpeechDat(E) — 50 слов и выражений, как *СПОНТАННО ПРОИЗНЕСЁННЫХ*, так и *ПРОЧИТАННЫХ*. Продолжительность записи (диалога между диктором и компьютером) составляла 8–10 минут в зависимости от темпа речи.

Исходный словарь базы данных содержит списки наиболее употребительных слов и команд из компьютерной лексики, цифр и цифровых последовательностей, названий крупных городов и фирм, обозначающих время фраз, дат, денежных единиц, телефонных номеров, номеров кредитных карт, сочетаний «имя-фамилия», фонетически богатых слов и предложений, а также спеллинг (побуквенное произнесение) слов.

Технические характеристики записывающей установки были стандартизированы для всех речевых баз данных. Записи проводились в автоматическом режиме через реальный цифровой телефонный канал европейского стандарта ISDN. Сигнал имел формат: 8 бит, 8 кГц, А-закон. Качество соединения и линии связи характеризовалось отношением сигнал/шум. Непригодные по зашумлённости записи исключались.

Обработка речевого материала выполнялась экспертами по речевой акустике. Она заключалась в многократном прослушивании всех звуковых файлов и их аннотации, которая производилась в соответствии со спецификацией, разработанной для участников проекта SpeechDat(II).

Аннотация подразумевала внесение следующей информации в файл-метку:

- орфографическая запись высказывания;
- специальные пометки, указывающие на наличие возможных шумов, оговорок, обрывов записи;
- оценка качества записи;
- данные о дикторе (возраст, пол, региональный акцент);
- тип телефонного аппарата;
- тип акустического окружения.

Из всех слов, произнесённых дикторами разборчиво и без оговорок, был составлен лексикон (файл LEXICON) с указанием частоты встречаемости каждого слова и его фонематической транскрипции. Часть слов приведена с вариантами произнесения (разговорный вариант).

Полученный лексикон насчитывает около 16500 единиц. Фонематическая транскрипция лексикона выполнена в соответствии с системой символов Russian SAMPA (машинно-ориентированного языка). Кроме этого имеется файл акустического качества каждого речевого сигнала, файл информации о респонденте (пол, возраст, регионально-диалектальная принадлежность), файл содержимого базы данных.

Файл DISIGN содержит полное описание базы, её словаря, записывающей платформы, полную информацию о лексиконе (особенности произношения, частота встречаемости фонем и др.).

Поддержание стандартов качества созданных баз данных обеспечено двумя ступенями валидации, которая выполняется фирмой SPEX (Speech Processing Expertise Centre), созданной в рамках проекта SpeechDat для проверки качества и соответствия стандартам созданных баз данных.

В течение многих лет ведутся активные разработки в области формирования многоязычных УРБД на кафедре прикладной и экспериментальной лингвистки Московского государственного лингвистического университета (МГЛУ). При этом охвачены различные языки, включая языки этнических меньшинств Российской Федерации. В разработках участвует ряд языковых кафедр МГЛУ. В качестве примера приведём некоторые из УРБД.

УРБД для французского языка разрабатывалась на кафедре прикладной и экспериментальной лингвистики МГЛУ и в Центре фундаментального и прикладного речеведения МГЛУ (директор Центра — Р.К. Потапова) в рамках проекта «Корпусная лингвистика многоцелевого назначения».

В задачу входило формирование фонетической базы данных французского языка, представленной звучащими текстами. Первой задачей создания УРБД была разработка свода правил соотношения «звук-буква» для французского языка.

Правила были сведены в таблицы, которые включали рубрики: звук, буква/буквосочетание, примеры, примечание. В примечании давались исключения из правил и дополнительная информация.

При транскрибировании использовались фонетические шрифты: NewtonPhoneticNt, Phonetic TM, Phonetic TMUniv, WP Phonetic. Помимо этих правил в тот же корпус вошли таблицы французских гласных, согласных, полугласных, носовых гласных, а также таблица используемых в базе транскрипционных значков международного фонетического алфавита.

Первичный корпус базы представлен фрагментами французской речи, подлежащей сегментации дофразового и фонемного уровней (в зависимости от внутренней спецификации задачи). Записи проводились с помощью программ Cool Edit 2000 и Real Player Plus 8.0. Ряд записей представляет собой оцифрованные фонограммы текстов разного характера, полученные с использованием материала на аудиокассетах (условия оцифровки: 22050 Гц, 16 бит, моно). Тексты включали монологи, диалоги, полилоги, отрывки из театральных спектаклей и др. в исполнении 25 мужчин и 20 женщин. Общее время — 15 час.

Другие записи представляли собой сообщения новостей, взятые с разных порталов Интернета в прочтении 25 мужчин и 20 женщин. При записи новостей он-лайн возникли некоторые трудности: при загрузке файлом реального времени .rm происходили изменения бит-рейта, которые отразились на качестве звучания речи. Последующая стадия обработки звука позволяла компенсировать этот недостаток. После записи речь в файлах подлежала сегментации и помещению в отдельные файлы, соответствующие определённым сегментам.

Далее проводилось аннотирование. Для ряда текстов имелись и видеозаписи, что существенно расширило базу данных и послужило основой для последующих разработок в области создания мультимодальных БД.

В задачу **УРБД для арабского языка** входило формирование фонетической базы данных арабского языка, представленной звучащими текстами. База данных разрабатывалась в том же Центре в рамках проекта «Корпусная лингвистика многоцелевого назначения». Первой задачей создания УРБД была разработка корпуса правил соотношения «звук-буква» для арабского языка.

Помимо правил, база содержала: папку с файлами текста (txt); папку со звуковыми файлами (wav); папку со звуковыми файлами несегментированного материала (тренировочный комплекс).

Исходным материалом служили файлы, представленные в качестве примера в следующей таблице.

Таблица 2

Исходные файлы

№	Файл	Время звучания**	Дикторы	Источники звучащей речи
1	1_a	45:17	7m,1f	Aljazeera
2	1_b	22:46	5m	Aljazeera
3	2_a	30:10	8m,3f	Aljazeera
4	2_b	14:16	7m,	Aljazeera
5*	3_a	45:34	4m,3f	London course of Arabic
6*	3_b	15:31	4m,3f	London course of Arabic
7*	4_a	45:05	4m,3f	London course of Arabic
8*	4_b	11:28	4m,3f	London course of Arabic
9*	5_a	45:32	4m,3f	London course of Arabic
10*	5_b	13:04	4m,3f	London course of Arabic
11*	6_a	46:09	4m,3f	London course of Arabic
12*	6_b	12:37	4m,3f	London course of Arabic
13***	7_a	45:21	3f	Alarabia
14***	7_b	45:09	3f	Alarabia
15****	8_a	46:08	10 m,6f	Aljazeera
16****	8_b	28:16	10m,6f	Aljazeera

Общее время звучания — 8,5 часов. Дикторы — 59 мужчин и 20 женщин.

Ниже представлены примеры оцифрованных записей (22050 Гц, 16 бит, моно) с аудиокассет. Все файлы типа Windows PCM (wav).

Таблица 3

№	Файл	Время звучания	Дикторы	Источники звучащей речи
1	Aljazeera1	13:42	5m	Aljazeera
2	Bbc1	9:48	2m, 1f	BBC
3	Dw2	30:19	7m,1f	Deutsche Welle
4	Jaber Ibn Hayan	27:26	4m,2f	VOA

Общий объём звучащего материала УРБД составил 1 Гб (соответствует 6,5 ч звучания). Представлены голоса 79 дикторов (59 мужчин и 20 женщин) — носителей различных произносительных вариантов арабского языка. УРБД включает 2070 пар файлов (аудиоматериал/текст в орфографии и транскрипции). Аудиофайл представляет собой оцифрованную запись фрагмента арабской речи в формате WAV (Microsoft Wave). Такое представление данных позволяет легко проводить поиск и сопоставление, а также инкорпорировать информацию в любые автоматизированные речевые системы, что соответствует задаче формирования УРБД многоцелевого назначения.

В ряде случаев в качестве исходного материала использовались файлы сжатых аудиоформатов (в частности, WMA-8, качество FM Radio), которые в дальнейшем были также приведены к формату WAV с указанными выше параметрами. Максимальная длительность звучания одного фрагмента — 81,8 с., минимальная — 0,2 с. Каждый фрагмент включает голос одного диктора, записанный в одних и тех же условиях.

На рис. 1 представлено распределение значений длительности сегментов в УРБД (группированный ряд).

Формирование УРБД арабского языка осуществлялось по следующим этапам:

1. Проводилось многократное прослушивание звучащего материала в полном объёме аудиторами — специалистами в области арабского языка, а также аудиторами – специалистами в области экспериментальной фонетики.
2. Транскрибировался каждый прослушанный звучащий файл с использованием системы транскрипции, принятой в международной информационной сети (SAMPA for Arabic).
3. В процессе аудитивного анализа и транскрибирования использовались скомпонованные на 1-м этапе данного исследования файлы (N = 16), содержащие аутентичный арабский материал в исполнении дикторов — мужчин и женщин.
4. В ходе аудитивного анализа отбраковывалась часть материала вследствие наличия зашумлённости речевого сигнала. Итоговое время звучания речевого материала составило 8 часов.
5. Параллельно проводилась сегментация двух видов: акустическая (с использованием программ CoolEdit 1.0 и Sound Forge 4.5 c) и текстовая (с использованием текстового процессора Microsoft Word).

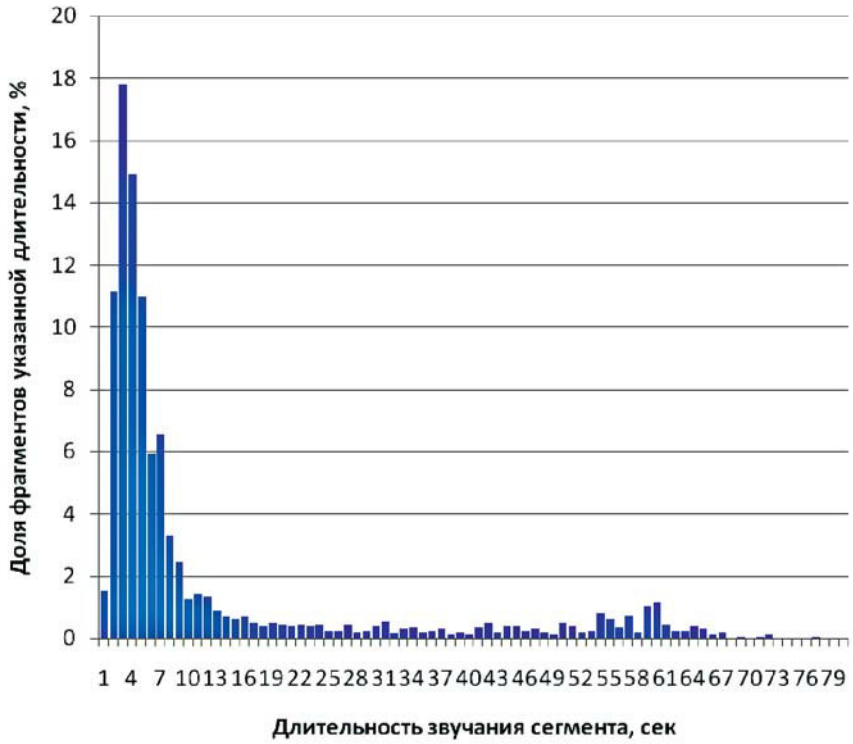

Рис. 1. Распределение значений длительности сегментов УРБД арабского языка

6. Полученные речевые сегменты (nΣ = 2070) подвергались повторному аудитивному анализу с целью подтверждения точности соответствия акустической и текстовой информации в рамках каждого сегмента.
7. Отсегментированный материал (звуковые и текстовые файлы) записывались на оптические носители информации (компакт-диски).

Среди устно-речевых фрагментов, представленных в УРБД, преобладают реплики диалогов (спонтанных или разученных), которые в совокупности составляют около 75% выборки. В основном этим объясняется доминирование в УРБД сегментов длительностью до 10 с. Примеры фрагментов представлены на рис. 2, 3. Оставшаяся часть материала представляет собой фрагменты монологической спонтанной речи (10%) или записи профессионального чтения текстов на литературном арабском языке (15%).

При сегментации больших участков, содержащих голос одного диктора, на фрагменты длительностью до полутора минут (в соответствии с ограничениями, поставленными при разработке структурной концепции УРБД) во всех случаях соблюдалось правило, согласно которому граница фрагментов не должна разрывать предложение (или дыхательную группу) в речевом потоке. Таким образом, фрагменты, длительность которых превышает 10 с., могут содержать от 1 до 4 единиц этого типа (что обусловливает наличие неярко выраженных максимумов на гистограмме в районе 30, 45 и 60 с.).

Все фрагменты, включенные в УРБД, затранскрибированы с использованием международного универсального фонетического алфавита SAMPA (см. http://www.phon.ucl.ac.uk/home/sampa/home.htm, а также http://en.wikipedia.org/wiki/SAMPA) с указанием долго-

ты гласных. В транскрипционную запись была также введена некоторая информация о грамматико-морфологической структуре слов (показано наличие артиклей, подвергшихся фонетической ассимиляции), поскольку эта информация может быть полезной в дальнейшем при использовании УРБД в целях изучения фонетической вариативности арабской речи. Транскрипция помещена в текстовые файлы (*.TXT), имена которых идентичны именам соответствующих аудиофайлов (*.WAV). Данная УРБД охватывает различные региональные и гендерные произносительные варианты арабского языка, а также различные виды речевой деятельности.

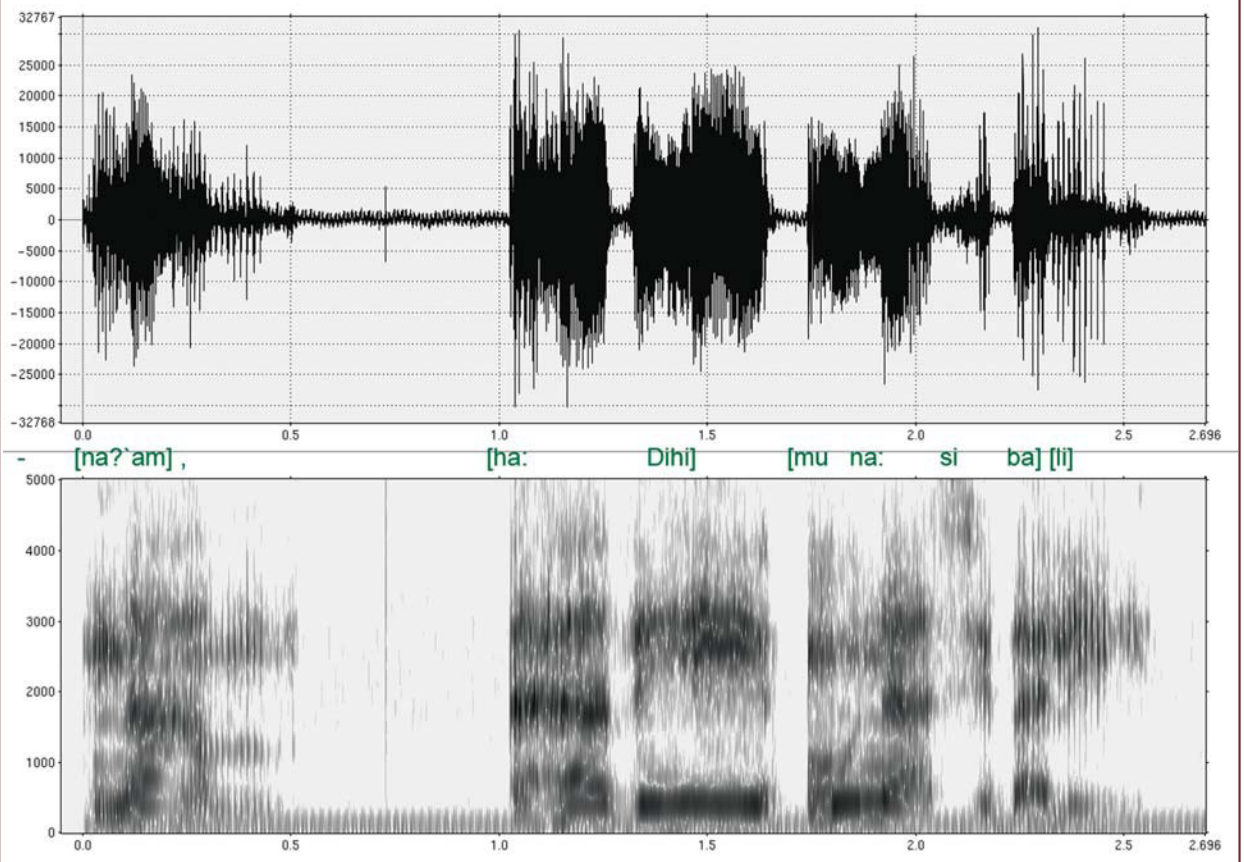

Рис. 2. Пример фрагмента УРБД (5B4MB-12.WAV, TXT), мужской голос

Аналогичные УРБД разрабатывались в Центре фундаментального и прикладного речеведения для турецкого, китайского языков и некоторых языков этнических меньшинств РФ.

Центры цифровых баз данных (les centres de ressources numйriques, CRN) созданы по совместной инициативе Управления научной информации и научного отдела «Человек и общество» Национального центра научных исследований Франции (Centre Nacional de la Recherche Scientifique, CNRS). Центр баз данных для устной речи (Centre de Ressources sur la Description de l'Oral, CRDO) и центры цифровых баз данных (CRN) сосредоточили

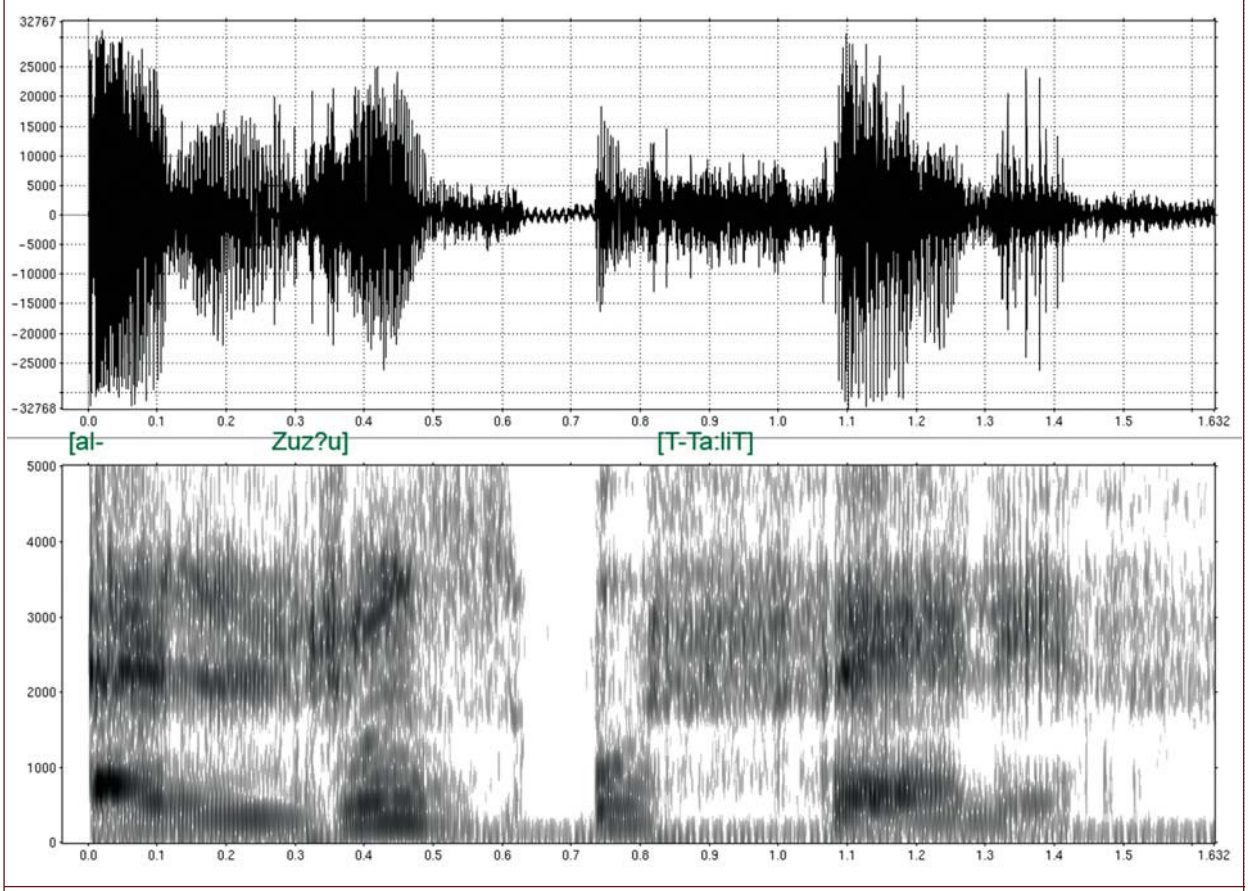

Рис. 3. Пример фрагмента УРБД (5B6F-1.WAV, TXT), женский голос

своё внимание на ресурсах устной речи. В 2006 г. Национальный центр научных исследований поручил Лаборатории языка и речи и Лаборатории языков и цивилизаций с опорой на устную традицию сформировать Центр баз данных для устной речи (CRDO) для таких задач, как каталогизация имеющихся массивов, централизация и обеспечение доступа к ресурсам и инструментам для изучения устной речи. Одним из главнейших компонентов этой работы стало формирование УРБД для различных языков мира.

В 2006–2007 гг. работы, проводимые CRDO, включали решение следующих задач:

— создание и отладка серверной структуры для хранения и обеспечения доступа к ресурсам УРБД;

— разработка структуры метаданных, описывающих содержание каждой УРБД с использованием лингвистических дескрипторов, соответствующих международным стандартам;

— обеспечение авторизованного доступа к УРБД для участников проекта CRDO;

— обеспечение возможности совместного доступа к УРБД, редактирования метаданных и другой информации [Bel, Blache, 2006: 13–14].

Создание серверной структуры проводилось в два этапа. На первом этапе была разработана структура реляционной базы данных для хранения и непосредственного редактирования метаданных. Эта реляционная УРБД была также призвана стать связующим звеном для всех будущих УРБД, формируемых CRDO.

При формировании структуры УРБД учитывались следующие требования: 1) структура УРБД не должна исключать любую априорную информацию; 2) данные, выдаваемые УРБД по запросу, должны быть структурированы согласно международным стандартам.

Информация об организациях, создающих УРБД, хранится в независимой базе данных (617 учреждений, см. http://teck.lpl.univ-aix.fr/institution/institution-recherche.htm). Эта база данных содержит также индексы организаций в перечне CCSD (сервер HAL; см. http://import.ccsd.cnrs.fr/doc/?consultLabs).

Основной массив информации (реляционная УРБД для хранения метаданных и собственно УРБД) был размещён на серверах высокого класса надёжности. Была также предусмотрена функция создания резервной копии информации на удалённом сервере для снижения риска потери информации. ПО серверов обеспечивает доступ к информации при помощи инструментария Apache, PHP, MySQL. Сайт системы имеет трёхъязычный интерфейс (французский, английский и китайский языки). На сайте введена в действие система авторизации доступа. Ряд операций доступен только зарегистрированным пользователям. В то же время любой посетитель сайта имеет свободный доступ к каталогу данных, инструментов и ресурсов, размещённых на сервере CRDO, а также к части относящихся к ним метаданных. В дальнейшем планируется также размещение в открытом доступе образцов аудио- и видеоматериалов УРБД, включая относящиеся к ним маркеры различных уровней (сегментная структура, просодическое оформление речи, данные о специфике артикуляции и т. п.).

Система встроенных запросов позволяет выбирать информацию из УРБД по параметрам, указанным в полях метаданных.

В дальнейшем CRDO планирует выполнить следующие работы:

— обеспечить возможность автоматического просмотра метаданных лингвистического характера в файле XML в формате OLAC для того, чтобы автоматизировать процедуру пополнения базы данных CRDO ссылками;

— ввести в действие систему текстовых информационных пространств Wiki для совместного редактирования веб-страниц в дополнение к метаданным;

— разработать процедуры ускоренного доступа к ресурсам УРБД;

— провести детальный анализ метаданных с целью унификации и стандартизации информационных структур УРБД, исключения ошибок и сбоев при поиске информации по запросам.

Литература

1. Автоматизированное рабочее место эксперта-фоноскописта. Электронная энциклопедия, версия V1.0: http://www.estra.ru

2. *Андрющенко В.М.* Концепция и архитектура Машинного фонда русского языка. М., 1989.

3. *Белолипецкий С.И., Буря А.Г.* Специализированные СУБД для поддержки речевых баз данных // Сетевой электронный научный журнал «Системотехника». № 2. 2004. М.: МГИЭМ, 2004.

4. *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.

5. *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.

6. *Богуславский И.М., Григорьев Н.В.* и др. Аннотированный корпус русских текстов: концепция, инструменты разметки, типы информации // Труды Международного семинара ДИАЛОГ-2000. М., 2000. Т. 2. С. 41–47.

7. Корпусная лингвистика в России / Сост. *Е.В. Рахилина* и *С.А. Шаров* // Спец. выпуск журнала НТИ. М., 2003. Сер. 2. № 6, 10.

8. *Кривнова О.Ф., Захаров Л.М., Строкин Г.С.* Речевые корпусы (опыт разработки и использование). М.: МГУ [http://www.dialog-21.ru/archive_article.asp].

9. *Кривнова О.Ф., Захаров Л.М., Строкин Г.С.* Речевые корпусы (опыт разработки и использование) // Сборник трудов Международного семинара Диалог'2001 по компьютерной лингвистике и её приложениям (в двух томах); Т.2. Прикладные проблемы. М., 2001.

10. *Кривнова О.Ф.* Области применения речевых корпусов и опыт их разработки // Сборник трудов XVIII сессии РАО. М.: ГЕОС, 2006.

11. *Леонтьева Н.Н.* Автоматическое понимание текстов: Системы, модели, ресурсы. М.:Академия/ Academia, 2006.

12. *Потапова Р.К., Линднер Г.* Особенности немецкого произношения. М.: Высшая школа, 1991. 319 с.

13. *Потапова Р.К.* Лингвистическое обеспечение Электронной Энциклопедии, предназначенной для экспертов-фоноскопистов (русский язык). М.: ЭСТРА, CDROM, 1998–1999.

14. *Потапова Р.К.* Новые информационные технологии и лингвистика. 4-е изд., сущ. доп. М.: Эдиториал УРСС, 2005. 368 с.

15. *Потапова Р.К.* Речь: коммуникация, информация, кибернетика. М.: Радио и Связь, 1997. 528 с.

16. *Потапова Р.К..* Тайна современного кентавра. М.: Радио и связь, 1992.

17. *Рыков В.В.* Корпус текстов — новый тип словесного единства // Труды Международного семинара ДИАЛОГ-2003. Протвино, 2003.

18. *Сичинава Д.В.* К задаче создания корпусов русского языка в Интернете // НТИ. М., 2002. Сер. 2. № 12.

19. *Скрелин П.А., Щербаков П.П.* Требования к современной фонетической базе данных для фундаментальных и прикладных исследований // Технологии информационного общества — Интернет и современное общество: труды VI Всероссийской объединенной конференции. Санкт-Петербург, 3–6 ноября 2003 г. СПб.: Изд-во Филологического ф-та СПбГУ, 2003. С. 62–63.

20. *Шаров С.А.* Параметры описания текстов корпуса, а также Корпусная лингвистика в России // НТИ. М., 2003. Сер. 2. № 5–6.

21. *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.

22. *Barlow M.* Corpora for Theory and Practice // 1JCL. Amsterdam, 1996. № 1.

23. *Bel B., Blache P.* Le Centre de Ressource pour la Description de l'Oral // Corpus. Les enjeux de l'annotation. Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence, 2006. Pp.13–18.

24. *Bertrand R., Blache P., Espesser R., Ferre G., Meunier C., Priego-Valverde B., Rauzy S.* Le CID — Corpus of Interactional Data: Protocoles, Conventions, Annotations // Corpus. Les enjeux de l'annotation. Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence, 2006. Pp.31–60.

25. *Bohmova A.* Automatic Procedures in Tectogrammatical Tagging. //The Prague Bulletin of Mathematical Linguistics. Prague, 2001. № 76. P. 23–34.

26. *Collier A., Pace y M., Renouf A.* Refining the Automatic Identification of Conceptual Relations in Large-scale Corpora. // Proceedings of the Sixth Workshop on Very Large Corpora. Montreal, 1998.

27. *Delais — Roussarie E., Post B., Portes C*. Annotation prosodique et typologia. Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence. Vol.25, 2006. p. 61–95.

28. *Greenstette G., Segond F*. Multilingual Natural Language Processing // IJCL. 1997. V.2. — № 1.

29. *Hajicovd E., Pajas P., Vesela K*. Corpus Annotation on the Tectogrammatical Layer: Summarizing of the First Stages of Evaluations // The Prague Bulletin of Mathematical Linguistics. Prague, 2002. № 77. P. 5–18.

30. International Journal of Corpus Linguistics (IJCL) / Ed. W.Teubert. Amsterdam, 1996–2001.

31. *Kibkalo A.A., Lotkov M.M*. Choice of Phonetic Alphabet for Russian LVCSR System // Proceedings of the International Workshop «Speech and Computer» SPECOM' 2003. (Moscow, 27–29 October, 2003) Moscow: MSLU, 2003. P. 102–105.

32. *Kucova L., Hajic ova E*. Prague Dependency Treebank: Enrichment of the Underlying Syntactic Annotation by Coreferential Mark-Up // The Prague Bulletin of Mathematical Linguistics. Prague, 2004. № 81. P. 23–34.

33. *Lee Y.-J., Choi D.-L., Um Y., Lee K.-H., Kim Y.-I., Kim B.-W*. Speech Resources at SITEC in Korea // Proceedings of the 10th International Conference SPEECH and COMPUTER (SPECOM' 2005) (Patras, Greece, 17–19 October, 2005) Patras, Moscow: MSLU, 2005. P. 579–582.

34. *Loseva E., Potapova R.* Speech variability of vibrants: phonetic database for English and German // Proceedings of the 10th International Conference Speech and Computer SPECOM' 2005, Patras, Moscow: MSLU, 2005.

35. *Marcus M.P., Santorini B., Marcinkiewicz M.A*. Building a Large Annotated Corpus of English: The Penn Treebank // Computational Linguistics. 1993. Vol.19. № 2. P. 313–30.

36. *Potapova R.K, Potapov V.V.* Database of forensic phonetics knowledges (as applied to electronic encyclopaedia for Russian experts) // Proceedings of the International Conference of IAFP, York, UK, 1999. P. 6–7.

37. *Shaikevich A*. The Computer Fund of Russian Language // IJCL.-Amsterdam, 1997. V.2. № 1. P. 163–167.

38. *Teubert W*. Corpus Linguistics and Lexicography // IJCL. Philadelphia, 2001.

39. htp://www.mdi.ru/aspnews/body/03.12.2001_39303.html

40. http://cfrl.ru

41. http://conf.infosoc.ru/03-r2f14.html

42. http://www.auditech.ru

43. http://www.auditech.ru

44. http://www.mdi.ru/aspnews/body/03.12.2001_39303.html

Потапова Родмонга Кондратьевна
академик Международной академии информатизации,
доктор филол. наук, профессор.
заслуженный работник Высшей школы РФ,
зав. отделением прикладной лингвистики,
зав. кафедрой прикладной и экспериментальной лингвистики,
директор Центра фундаментального и прикладного речеведения
Московского государственного лингвистического университета.
Специалист в области романо-германского языкознания,
общей и прикладной фонетики, теоретической, прикладной,
экспериментальной и математической лингвистики.
Автор свыше 450 научных и научно- методических публикаций.