Звуковой корпус как способ мониторинга и фиксации разных форм естественного языка*

A speech corpus as a tool for monitoring and fixation of various forms of natural language

Богданова Н. В. (nvbogdanova_2005@mail.ru), **Асиновский А. С.** (a.s.asinovsky@gmail.com), **Русакова М. В.** (mvrusakova@gmail.com), **Рыко А. И.** (aryko@yandex.ru), **Степанова С. Б.** (stsvet_2002@mail.ru), **Шерстинова Т. Ю.** (sherstinova@gmail.com)

Факультет филологии и искусств Санкт-Петербургского государственного университета, Санкт-Петербург, Россия

Доклад посвящен разработке методов мониторинга и фиксации звукового материала естественного языка, принципов организации информационной среды и программного инструментария для нужд интегрального моделирования, а также описанию первых готовых блоков Звукового корпуса русского языка (ЗКРЯ).

1. Введение

Язык современного города представляет собой, вне всякого сомнения, фактор социальной и психологической дифференциации — настолько в нем переплетены и взаимосвязаны различные социолекты и идиолекты. Это и делает его столь интересным объектом для самых разных исследований. О «лингвистической истории большого города» много писал еще Б. А. Ларин, усматривая в ней «борьбу языков, отражающую непрестанное столкновение и скрещивание <…> разнородных культур» [6: 177]. По мнению Б. А. Ларина, языковое разнообразие города характеризуется, с одной стороны, *многоязычием*, обусловленным «встречей разноязычных коллективов», а с другой – *многодиалектностью*, поскольку «в каждом слое городского населения, кроме первичного, “своего”, наречия, необходимо располагают еще каким-либо универсальным языковым типом, приобщающим к большой социальной среде» [6: 190]. Наилучшим образом все это многообразие речи горожан может быть представлено с применением корпусного подхода к собранию речевого материала. Именно такой подход был реализован при создании Звукового корпуса русского языка (ЗКРЯ), сбалансированного как лингвистически, так и психо- и социолингвистически.

С другой стороны, одной из основных задач современной лингвистики является накопление и систематизация новых данных о том, как соотносятся лингвистические описания, сделанные на основе традиционных понятийных и терминологических систем, и физически наблюдаемая звуковая материя естественной устной русской речи. Поворот лингвистики от структуралистского описания языковой системы и обслуживания прикладных областей знания к говорящему и слушающему человеку выразился прежде всего в понимании бесперспективности дальнейшего моделирования методом «черного ящика» в процессе решения прикладных задач и «дальнейшего шлифования метода классических дефиниций» [7: 232]. «Нужна принципиально другая “дефинициология”, чем та, которая досталась нам от аристотелевой и математической логики». Необходимо понять, «что такое точное описание в применении к языковым явлениям. Такое понимание нельзя импортировать из других областей знания, включая математику. Чтобы понять операционную природу лингвистической точности, надо предварительно углубить наши неформальные знания о *природе* (курсив автора – *Авт.*) языковых явлений, а уж затем выработать соответствующие формализмы точного их представления» [7: 232].

Идея о существовании, наряду с грамматикой языка, своеобразной грамматики речи, обладающей своими собственными единицами и специфическими правилами их функционирования и сочетания (своей парадигматикой и синтагматикой), в лингвистике отнюдь не нова и в течение последнего столетия высказывалась неоднократно. Так, еще Ф. де Соссюр писал, что «лингвист должен также рассматривать

* В настоящее время работа проводится при поддержке РГНФ — проект 07-04-12163в «Разработка информационной среды для мониторинга устной русской речи».

взаимоотношения книжного языка и обиходного языка (по сути, взаимоотношения языка и живой повседневной речи — *Авт.*), ибо развитие всякого литературного языка, продукта культуры, приводит к размежеванию его сферы со сферой разговорного языка» [11: 44]. Подобную мысль находим и в трудах Л. В. Щербы, ср.: «нужно прежде всего различать у русских, т. е. у говорящих и пишущих на общерусском литературном языке, два языка: один слышимый и произносимый (снова, по-видимому, имеется в виду живая речь — *Авт.*), а другой написанный, которые находятся один к другому в известных отношениях, но не тождественны — элементы одного не совпадают с элементами другого» [15: 1]. И далее автор пишет, что «если надо различать эти два языка, то надо, очевидно, различать и их грамматики» [15: 12].

Одним из первых при таком подходе встает вопрос об инвентаре тех единиц, которые должны стать стержнем любой грамматики, в том числе и грамматики речи. Уже первые наблюдения над спонтанным материалом показали, что на нем преломляются все исходные языковые метапонятия, на которых строится обычно его анализ. Привычные понятия *фонемы*, *морфемы*, *слова* и *предложения* оказываются неприложимыми или плохо приложимыми к спонтанной речи. Фактически на этом материале все традиционные метапонятия языка (единицы его описания) так или иначе разрушаются, на их месте создается нечто новое, что не всегда легко поддается определению и описанию. Ср.: «…происходящее при переходе от чтения к говорению “переключение кодов” приводит к расширению допустимых пределов варьирования, т. е. размытости фонетических характеристик таких единиц, как фонема, слово (на уровне фонетической реализации), синтагма, размытости, степень которой возрастает по мере возрастания степени неофициальности общения» [13: 133].

Несмотря на значительные успехи в различных областях лингвистики, связанных (в самом общем смысле) с функционированием языка в процессе коммуникации, на сегодняшний день не существует сколько-нибудь целостного *многоуровневого* описания звучащей речи. Настоящий проект направлен на исследование разнообразных закономерностей разворачивания естественной устной речи в первую очередь на русском языке и представляет собой один из первых подходов к исследованию звучащей речи с применением новых методологических возможностей и с формированием новых терминологических решений в перспективе.

2. Материал корпуса

Материалом настоящего исследования является естественная звучащая русская речь. Формат языкового материала может быть определен как *Звуковой корпус русского языка* в его естественной, звучащей, форме. К настоящему времени подготовлены 3 основных модуля, состоящие из 6 речевых подкорпусов (см. табл. 1).

Таблица 1. Блоки речевого материала

Один речевой день			
ORD	«Один речевой день» (Повседневная речь)	34 информанта + 560 коммуникантов	235 часов звучания
Сбалансированный материал			
MED	Речь медицинских работников	32 диктора 210 текстов	6 часов звучания
JUR	Речь юристов	40 дикторов 322 текста	16 часов звучания
RKI	Речь преподавателей РКИ	20 дикторов 70 текстов	3,5 часа звучания
STUD	Речь студентов	5 блоков	7 часов звучания
Интерферированная речь			
RIA	Интерферированная речь (русская речь американцев)	48 дикторов	8 часов звучания

А. *ORD. «Один речевой день»*

Принципиальное значение для нашего проекта имеет модуль «Один речевой день», занимающий в нем центральное место. *Целью* создания данного блока ЗКРЯ явилось изучение речевого поведения носителя русского языка в течение дня (с использованием методики 24-часовой записи) в зависимости от ряда факторов:

- его социально-психологические характеристики,
- его коммуниканты (= социальная роль),
- место, где протекает общение,
- время суток.

Выборка информантов при записи материала ORD — пока несбалансированная, хотя, вероятно, она в некоторой степени отражает социальный и психологический срез современного общества. Из 34 участников эксперимента было 12 мужчин и 22 женщины, в возрасте от 15 до 63 лет. Одновременно была осуществлена запись около 560 их коммуникантов (приблизительность цифры объясняется невозможностью в некоторых случаях определить принадлежность голоса одному или разным коммуникантам). Все они относятся к различным социальным группам и имеют разный уровень образования — от среднего специального до высшего, в том числе с ученой степенью. Сферы деятельности информантов также оказались разнообразны — это и учеба, и занятость на производстве, и научно-преподавательская деятельность, и бизнес, и торговля, и юриспруденция, и медицина, и строительство, и некоторые другие[1].

[1] Подробнее о модуле ОРД см. [12].

*В. **Сбалансированный материал***

Иные принципы были положены в основу создания второго блока Звукового корпуса русского языка [1]. Этот блок изначально достаточно строго сбалансирован по разным параметрам — социологически, психологически и собственно лингвистически.

Социологическая балансировка материалов корпуса заключалась в изначальном подборе информантов с разными социальными характеристиками, среди которых, помимо традиционных признаков пола, возраста, образования и проч., следует отметить *уровень речевой компетенции* (УРК) говорящего, который характеризует умение человека решать разные коммуникативные задачи, свободу его в выборе речевых средств и навыки построения устного монолога различного характера. Определяют УРК в основном два признака говорящего индивида: уровень образования и *профессиональное* или *непрофессиональное* его *отношение к речи*. Предполагается, что непрофессионально относятся к речи те люди, для кого язык/речь является лишь *средством общения* (большинство так называемых «наивных» носителей) или средством общения и *объектом изучения* (школьники или «кабинетные» ученые-филологи). Их профессиональная деятельность обычно не связана с активной публичной речевой практикой. Профессиональное отношение к речи характеризует тех, для того язык/речь – это не только средство общения или даже объект изучения, но еще и *орудие труда* (актеры, дикторы, лекторы, публичные и общественные деятели, преподаватели, особенно — преподаватели-филологи). Высшее образование и профессиональное отношение к речи нормально формируют высокий уровень речевой компетенции говорящего, высшее образование и непрофессиональное отношение к речи — средний УРК, отсутствие высшего образования и непрофессиональное отношение к речи – низкий УРК.

Психологическая балансировка данного блока Звукового корпуса (осуществленная лишь частично) заключалась в подборе информантов с разными психологическими характеристиками. В основе такого подбора лежал психологический тест Г. Айзенка на определение интровертности-экстравертности и эмоциональной неустойчивости личности[2].

Лингвистическая балансировка материала заключалась в том, что все тексты, составляющие Корпус, построены в рамках комплекса *коммуникативных сценариев*, обычно реализующихся в нашей повседневной бытовой речи:

- *чтение* (сюжетный/несюжетный исходный текст);
- *пересказ* (сюжетный/несюжетный исходный текст);
- *описание изображения* (сюжетное/несюжетное);
- *свободный рассказ на заданную тему* (знакомая/незнакомая тема).

Представляется, что все характеристики подобных бытовых монологов определяются двумя их признаками, находящимися в отношениях обратно пропорциональной зависимости: степень *лингвистической мотивированности* текста неким исходным стимулом и степень *спонтанности* вторичного речевого произведения. Чем более монолог мотивирован тем или иным первичным текстом, тем он менее спонтанен, и наоборот.

Дополнительными характеристиками исходного стимула, способными повлиять на свойства спонтанного монолога, стали в данном проекте *сюжетность-несюжетность предтекста или изображения* и *степень знакомства* говорящего *с темой свободного монолога*, заданной вопросом. Эти дополнительные характеристики не меняют степени лингвистической мотивированности и, соответственно, степени спонтанности вторичного текста, но все же оказывают влияние на выбор говорящим тех или иных речевых средств и в целом на лингвистическую природу вторичного текста. Можно предположить, что в данном случае решающими являются характеристики уже не (или не только) первичного текста, но и самого говорящего — уровень его речевой компетенции или психологический тип личности.

Думается, что соблюдение принципов, положенных в основу создания данного блока Звукового корпуса русского языка, позволяет получить достоверный и представительный речевой материал, пригодный для анализа в различных аспектах и дающий представление об особенностях речи того или иного социума.

3. Программное обеспечение исследования

Формирование звуковых корпусов и их многоуровневая разметка стали возможны в последнее время благодаря развитию информационных технологий в гуманитарных науках. Работа, связанная с интерпретацией звучащей речи, исключительно трудоемка, но благодаря специальным программам она принципиально выполнима.

При создании нашего корпуса используются следующие программные средства:

- программа профессионального фонетического анализа Praat;
- профессиональный аудиоредактор Sound Forge 8.0;
- программа многоуровневого лингвистического аннотирования ELAN;
- программа лексикографической обработки данных KartaTeka (собственная разработка);
- программа для создания баз данных Accsess.

Исследовательский интерфейс связывает все модули в единую информационно-исследовательскую среду.

[2] Подробнее об использовании данного теста в наших исследованиях см. [5].

3.1. Программа профессионального фонетического анализа Praat

Программа Praat, созданная сотрудниками факультета фонетики Амстердамского университета П. Бёрсма и Д. Вининком (Paul Boersma, David Weenink)[3], предназначена для лингвистов, исследующих звучащую речь. Она предоставляет ряд возможностей для сегментации звукового потока, анализа и синтеза речи, для манипуляций со звуком в целях проверки различных гипотез, связанных с организацией звуковой формы языка, а также дает возможность создания иллюстративного материала для публикации результатов исследований.

Поскольку Praat позволяет осуществлять многоуровневую разметку звучащей речи, именно эта программа использовалась нами для соположения и интеграции информации, относящейся к собственно акустическому, фонетическому, словесному и фразовому уровням[4].

3.2. Программа многоуровневого лингвистического аннотирования ELAN

Программа ELAN является средой профессионального лингвистического аннотирования аудио- и видеоматериалов, которая поддерживает фактически неограниченное количество уровней аннотации, любые шрифты и кодировки данных, сложные иерархические структуры связей между данными, экспорт и импорт аннотаций в основные форматы представления данных. Программа разработана в институте психолингвистики им. Макса Планка в Голландии специально для исследования звучащего языка, речевого поведения и жестикуляции и является удобным средством для обработки, документирования и аннотирования разнообразных мультимедийных корпусов[5].

ELAN поддерживает:

- визуализацию аудио- и/или видеосигналов одновременно с полученными аннотациями;
- временну́ю привязку аннотаций к медийному потоку;
- сложные связи аннотаций друг с другом;
- неограниченное количество задаваемых пользователем уровней аннотации (Tiers);
- различные шрифты и кодировки;
- экспорт данных в виде текстовых файлов табличного вида (tab-delimited text);
- импорт и экспорт между ELAN, PRAAT, ToolBox, Shoebox и другими популярными лингвистическими программами;
- поисковые опции.

На рис. 1 представлен звуковой фрагмент *Я настолько себя плохо чувствую, я так устал* с разметкой формата ELAN.

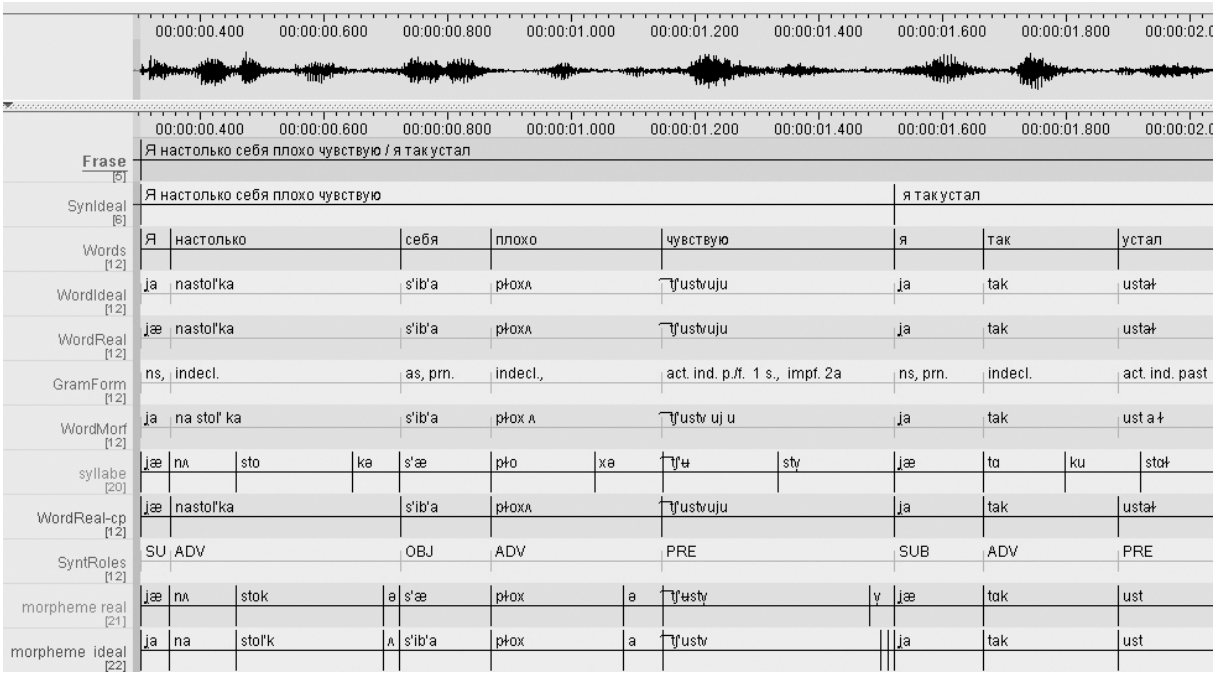

Рис. 1. Разметка данных в формате ELAN

[3] Создатели Praat регулярно обновляют версии своей программы на сайте www.praat.org и предлагают бесплатное ее использование для некоммерческих целей.

[4] Подробное описание принципов выделения и аннотирования уровней см. в [10].

[5] Программа является свободно распространяемой и может быть скачана с сайта института психолингвистики им. Макса Планка: http://www.lat-mpi.eu/tools/elan/.

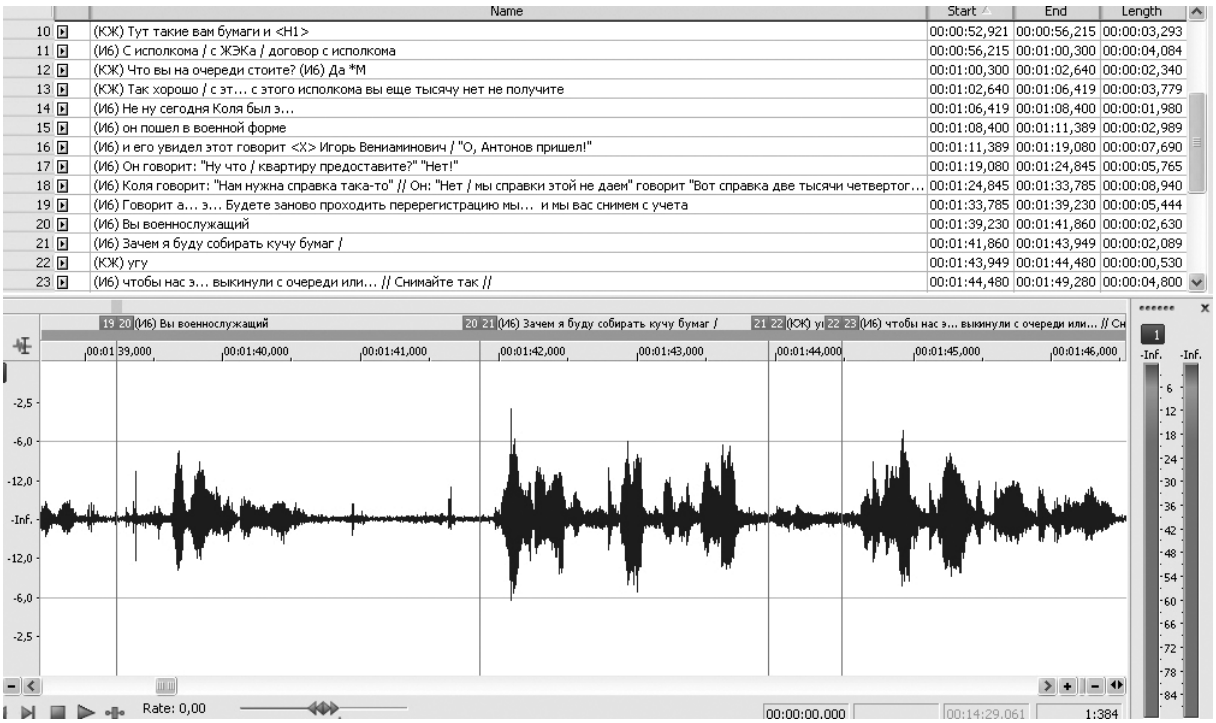

Рис. 2. Разметка данных в формате Sound Forge

3.3. Сравнение форматов аннотирования данных

Использование разных форматов аннотирования обусловлено спецификой используемых программных средств. Так, программа лингвистического аннотирования ELAN является крайне удобной для представления многоуровневой разметки разнородного материала, хранит данные в удобном для их последующей автоматической обработки виде. Именно формат ELAN принят за основной для разметки нашего корпуса. Программа PRAAT используется для фонетической обработки материала — заполнения аннотаций на фонетических уровнях, а программа Sony Sound Forge хорошо работает с аудиофайлами большой длительности и позволяет совместить аннотирование с их точной сегментацией (см. рис. 2). Все используемые форматы аннотирования данных являются совместимыми и могут быть взаимно перекодированы.

3.4. Электронная картотека E-Kar

Программа E-Kar автоматически создает конкорданс по выбранным текстам и позволяет решать многообразные задачи классификации и описания языковых единиц. В частности, она позволяет собрать по тексту все словоформы, в нем встречающиеся, посчитать их частоты (см. рис. 3), предъявить для этих словоформ любое лингвистическое расширение по тексту или группе текстов (см. рис. 4), имеющихся в электронной коллекции.

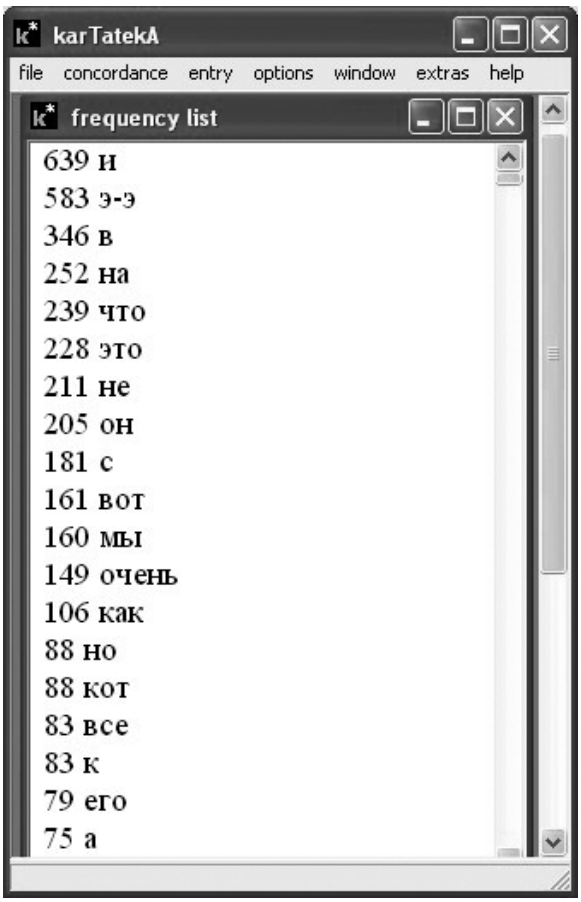

Рис. 3. Частотный словник словоформ

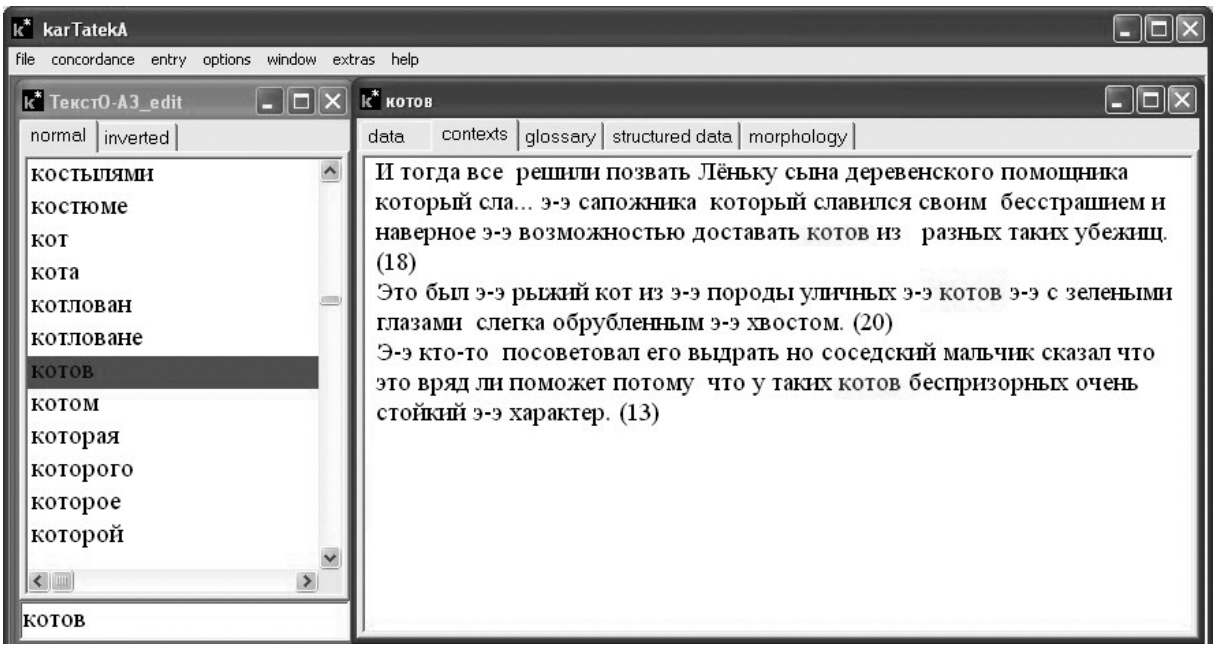

Рис. 4. Словоформа с контекстом

Результаты автоматической работы программы (конкорданса) и последующей работы эксперта дают возможность новых содержательных форм интерпретации лингвистического материала.

Конкорданс дает необходимую для лингвиста информацию, которая помогает классифицировать словоформы по морфологическим признакам и проводить лемматизацию, а также идеографическую или тематическую классификацию.

3.5. Специализированная база данных

Для самого крупного звукового модуля исследовательской среды «Один речевой день» (ORD), включающего в себя около 260 часов аудиоматериала, подготовлена специализированная база данных SpeechDay, реализованная в среде MS Access 2003.

На настоящий момент база данных состоит из 7 заполненных таблиц. Все таблицы можно условно разделить на 2 группы: *фактические данные* и *результаты научно-исследовательской работы и их интерпретация*. Некоторые таблицы содержат «смешанные» данные[6].

4. Некоторые результаты исследований на материале ЗКРЯ

Хотя работа над созданием ЗКРЯ еще далека от своего завершения (впрочем, вряд ли её можно и нужно завершать, учитывая основную цель проекта — отражение и фиксация постоянно изменяющейся повседневной русской речи), материал корпуса и процесс его аннотирования уже подвергся исследованиям, о результатах которых стоит здесь упомянуть.

Так, специальный анализ материала [14] показал, что бóльшая часть записей ORD (44 %) была сделана информантами на работе или учебе, со значительным отрывом далее следуют записи семейных разговоров по вечерам (10 %), разговоры в кафе или ресторанах (10 %), по дороге куда-либо (9 %), утром за завтраком (5 %). Остальные разговоры (за обедом, во время спортивных или культурных мероприятий, в гостях и т. п.) в жизни наших испытуемых занимают гораздо меньше времени. При этом различие между мужчинами и женщинами заключается в основном в том, что мужчины потратили на разнообразные мероприятия почти на 9 % больше времени, чем женщины; естественно и времени на дорогу у них ушло больше (почти на 5 %). Женщины же «это время» потратили на разговоры дома вечером (на 7 % больше), на вечеринках и за ужином (примерно на 2 % больше по каждой категории) и утром (на 3 % больше). Впрочем, с психологической и социологической точек зрения такой результат не является неожиданным.

Интересные результаты дала разметка части материала ЗКРЯ с точки зрения отклонений от нормативной речи на всех лингвистических уровнях. Оказалось, что отклонения в русской речи являются весьма обыденным явлением, по частотности сопоставимым с употреблением имен существительных и гораздо более частотным, чем, например, употребление прилагательных [9].

Статья [2] посвящена анализу зависимости способа передачи чужой речи от уровня речевой компетенции говорящего (материал MED).

[6] См. описание в [12].

В работе [8] приводятся результаты наблюдений над речевыми особенностями проявления агрессивности в спонтанной речи (материал ORD), в [4] анализируется лингвистическая структура высказывания с точки зрения проявления в ней психологических характеристик говорящего (ORD).

Проведено исследование специфики пересказа как вида речевой деятельности (на материале блока JUR). Анализ показал, что на порождение репродуктива оказывает влияние ряд факторов как лингвистического (характер первичного текста), так и экстралингвистического (социальные или психологические характеристики говорящих) толка [5].

В работе [4] показано, как профессия говорящего проявляется в спонтанной речи на лексическом уровне (блок MED).

Материалы Звукового корпуса по мере их обработки передаются в Национальный корпус русского языка.

Литература

1. *Богданова Н. В., Бродт И. С., Куканова В. В., Павлова О. В., Сапунова Е. М., Филиппова Н. С.* О «корпусе» текстов живой речи: принципы формирования и возможности описания // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7 (14). П материалам ежегодной международной конференции «Диалог» (2008) / Гл. ред. А. Е. Кибрик. М., 2008. С. 57–61.
2. *Богданова Н. В., Бродт И. С.* О способах передачи чужой речи (на материале звукового корпуса русского языка) // Материалы XXXVII международной филологической конференции. Выпуск 21. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 10–15 марта 2008 года / Отв. ред. А. С. Асиновский, науч. ред. Н. В. Богданова. СПб., 2008. С. 3–16.
3. *Иванова О. А.* К характеристике внутриязыкового контакта между литературной и профессиональной речью носителя русского языка // Материалы XXXVII международной филологической конференции. Выпуск 21. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 10–15 марта 2008 года / Отв. ред. А. С. Асиновский, науч. ред. Н. В. Богданова. СПб., 2008. С. 25–35.
4. *Королева И. В.* Индивидуальные состояния и свойства языковой личности: влияние на лингвистическую структуру высказываний // Материалы XXXVII международной филологической конференции. Выпуск 21. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 10–15 марта 2008 года / Отв. ред. А. С. Асиновский, науч. ред. Н. В. Богданова. СПб., 2008. С. 36–46.
5. *Куканова В. В.* Специфика пересказа как вида речевой деятельности: эндо- и эзоединицы звучащего спонтанного монолога // Вестник Санкт-Петербургского университета. Филология. Востоковедение. Журналистика. Серия 9. Вып. 4. Часть 2. СПб., 2008. С. 135–145.
6. *Ларин Б. А.* История русского языка и общее языкознание. М., 1977.
7. *Кибрик А. Е.* О «невыполненных обещаниях» лингвистики 50–60-х годов // Московский лингвистический альманах. Спорное в лингвистике. Вып. 1, М., 1996.
8. *Маркасова Е. В.* Риторическая энантиосемия в корпусе русского языка повседневного общения «Один речевой день» // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7 (14). По материалам ежегодной международной конференции «Диалог» (2008) / Гл. ред. А. Е. Кибрик. М., 2008. С. 352–355.
9. *Русакова М. В.* Сбои при порождении словоформы в устной речи как результат спонтанного взаимодействия стратегий и механизмов // Материалы XXXVI международной филологической конференции. Выпуск 20. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 12–17 марта 2007 года / Отв. ред. А. С. Асиновский, Н. В. Богданова. СПб., 2007. С. 59–71.
10. *Рыко А. И., Степанова С. Б.* Многоуровневая лингвистическая разметка звукового корпуса русского языка // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции «Диалог» (2008). М.: 2008. С. 460–465.
11. *Соссюр Ф. де.* Курс общей лингвистики. М., 1933.
12. *Степанова С. Б., Асиновский А. С., Богданова Н. В., Русакова М. В., Шерстинова Т. Ю.* Звуковой корпус русского языка повседневного общения «Один речевой день»: концепция и состояние функционирования // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7 (14). По материалам ежегодной международной конференции «Диалог» (2008) / Гл. ред. А. Е. Кибрик. М., 2008. С. 488–494.
13. *Фонетика* спонтанной речи / Под ред. Н. Д. Светозаровой. Л., 1988.
14. *Шерстинова Т. Ю.* «Один речевой день» на временно́й шкале: о перспективах исследования динамических процессов на материале звукового корпуса // Филология. Востоковедение. Журналистика. Серия 9. СПб., 2009 (в печати).
15. *Щерба Л. В.* Избранные работы по русскому языку. М., 1957. С. 11–20.