КОРПУС ЗВУЧАЩЕЙ РУССКОЙ РЕЧИ В СОСТАВЕ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА. ПРОЕКТ*

CORPUS OF ORAL RUSSIAN IN THE FRAMEWORK OF RUSSIAN NATIONAL CORPUS. CONSTRUCTION PROJECT

Гришина Е.А. *(rudi2007@yandex.ru),* ***Савчук С.О.*** *(savsvetlana@mail.ru)*
Институт русского языка им. В.В.Виноградова РАН

Статья содержит описание проекта «Корпуса звучащей русской речи», который может быть создан на материале подкорпуса «Речь кино» в составе Национального корпуса русского языка. В статье предлагаются предварительные решения, касающиеся структуры корпуса, типов разметки, формата выдачи материала, типов пользовательских запросов и разновидностей задач, которые могут быть решены посредством данного корпуса.

1. Постановка проблемы

Как известно, Национальный корпус русского языка включает в себя подкорпус устной речи (ср. Гришина 2005, Grishina 2006, Савчук 2008). В данный момент этот подкорпус выведен из состава основного корпуса и функционирует как самостоятельный модуль в рамках НКРЯ (наряду с поэтическим, диалектологическим, параллельным, образовательным и акцентологическим (в недалеком будущем) подкорпусами). Устный подкорпус включает в себя расшифровки реальной устной речи (публичной и частной, общий объем 4,4 млн словоупотреблений), а также кинотранскрипты (объем 1,6 млн словоупотреблений)[1].

Естественно, с самого начала функционирования устного подкорпуса возникал вопрос – и у самих создателей НКРЯ, и у его пользователей – о возможности создании корпуса не просто устной, но звучащей речи. И в силу ряда причин, в основном субъективного, но также и объективного характера этот вопрос регулярно получал отрицательный ответ[2].

Данная статья рассматривает следующую проблему: предположим, было бы принято решение начать создание корпуса звучащей русской речи (КЗРР) в рамках НКРЯ, – какой в этом случае могла бы быть технология создания этого корпуса? Каковы те задачи, которые было бы разумно ставить перед собой на начальных этапах, в какой последовательности и какими средствами и способами их решать?[3]

2. Источники КЗРР

Очевидным источником КЗРР являются расшифровки устных текстов, уже включенные в состав НКРЯ, для которых имеются соответствующие звуковые файлы, – казалось бы, достаточно разбить на небольшие блоки

* Статья написана при поддержке грантов РФФИ 06-06-80133а и 08-06-00371а.

[1] Разработка корпусов устной речи ведется по двум направлениям, в рамках которых решаются различные научные и практические задачи, что обусловливает и существенные различия в принципах и технологии создания корпусов: 1) фонетические (акустические, речевые) базы данных; 2) корпуса устных текстов (подробнее см. Савчук 2007). В существующих корпусах устных текстов на русском языке используются варианты графической фиксации звучащей речи – либо специально разработанная дискурсивная транскрипция, как в корпусе детских рассказов о сновидениях (Кибрик, Подлесская 2003; Подлесская, Кибрик 2004), либо упрощенная фонетическая транскрипция на базе русской орфографии, как в диалектном корпусе в составе НКРЯ, либо орфографическая запись с передачей отдельных фонетических особенностей устного дискурса, как в подкорпусе устной речи НКРЯ.

[2] Отметим, что за рубежом подходы к созданию аналогичных корпусов уже осуществляются, в основном в рамках психолингвистических исследований эмоций, ср. Clavel et al 2006, Devillers et al. 2006, исследований детской речи, ср. широко известный проект CHILDES, а также в образовательных проектах, ср. Braun et al. 2007.

[3] Предлагаемые в статье решения, безусловно, носят субъективный и вполне дискуссионный характер, но практика работы над Национальным корпусом показывает, что максимальные результаты достигаются только при наличии личной заинтересованности исполнителя в проекте, а неизбежной платой за это является определенная субъективность принятых решений.

скрипты и звуковые файлы, выровнять их между собой, добавить метаразметку, морфологическую и семантическую разметку, и корпус звучащей речи готов.

Однако, по зрелом размышлении, это очевидное решение следует признать если не неверным, то недостаточным. Мы считаем, что начинать строительство КЗРР следует с кинотранскриптов, включенных в состав НКРЯ, и соответствующих фильмов[4]. Аргументы здесь таковы.

Во-первых, кино содержит, помимо звукового, зрительный ряд, что позволяет при создании КЗРР поставить вопрос об одновременном создании корпуса русских жестов – то есть одна и та же операция (нарезка мультимедийного файла на блоки и их выравнивание с текстом) в случае работы с кинотранскриптами дает пользователю в два раза больше возможностей получения информации, чем в случае работы с доступными в настоящий момент создателям НКРЯ аудиофайлами[5].

Во-вторых, никакая, сколь угодно большая коллекция аудиофайлов не даст такого разнообразия устной речи, как кинотранскрипты, – хотя бы потому, что кино в большом количестве включает в себя ситуации, в которых в реальной жизни никакая аудиозапись в нормальном случае невозможна, а если и возможна, то результаты такой аудиозаписи для нас недоступны[6].

Очевидным недостатком принятого решения является тот факт, что в КЗРР включается «ненатуральная», «неестественная» устная речь. Здесь трудно что-либо возразить, кроме, пожалуй, следующего. Ненатуральность киноречи – миф, если не полностью, то в значительной степени (см., в частности, работу Капанадзе 1986 об ориентации языка кино на реальную разговорную речь). Нам уже приходилось писать о том, что речь кино – это результат совершенно особого процесса усвоения чужого текста актером и превращения этого текста в его, актера, собственный текст (см. Grishina 2007). И можно даже утверждать, что успех фильма в значительной степени зависит от того, насколько удачно и полно пройдет этот процесс присвоения[7].

С другой стороны, естественное повседневное общение протекает в стереотипных формах, усвоенных и воспроизводимых в стереотипных ситуациях. Не случайно в научных описаниях речевого взаимодействия, активно используются театральные метафоры: «Повседневное, бытовое по преимуществу, общение (персональный дискурс) обслуживает *сценарии* социального взаимодействия, которые можно уподобить принципам *комедии dell'arte*, где при достаточно четкой определенности *характеров действующих лиц актерам* предоставляется значительная свобода в содержании *реплик*» (Седов 2007: 21)[8].

Безусловным «проигрышем» речи кино по сравнению с реальной устной речью является повышенная связность устного текста в кинематографе. Но здесь речь кино проигрывает в основном частной, приватной устной речи, в особенности бытовой, где разговаривают хорошо знакомые люди, для которых нет необходимости проговаривать все до конца. Если мы сравним киноречь с публичной устной речью, то здесь различия по связности нивелируются, а что касается естественности человеческого речевого поведения (обрыв, наложение реплик, перебивы, запинки, смены стратегий и проч.), то по этому параметру в ряде случаев киноречь даст фору некоторым образцам публичной устной речи.

Как бы то ни было, нам представляется, что плюсы в случае работы с кинотранскриптами существенно мощнее, чем минусы.

---

[4] Безусловно, это не значит, что в качестве второго эшелона скрипты реальных устных текстов не должны быть включены в КЗРР, – речь в данный момент идет именно о начальном этапе работы.

[5] Таким образом, проект нацелен на реализацию многократно высказывавшейся, начиная с пионерских работ Л.П. Якубинского (1923) и В.Н. Волошинова (1930), идеи о том, что устную речь необходимо изучать в связи с ситуацией общения, в которой она порождается, поскольку сама внеязыковая ситуация становится частью акта коммуникации, а в передаче смысла сообщения участвуют не только вербальные, но и невербальные средства – жесты, мимика, интонация и под.; лингвистические описания русской разговорной речи с опорой на ситуацию находим, в частности, в работах Земская и др. 1981, Китайгородская, Розанова 2005. Корпус звучащей речи в этом плане резко расширяет возможности исследователя. В последнее время аудио- и видеофиксация материалов с целью создания корпусов и баз данных начинает использоваться и в других областях филологии: при документировании малых языков (Кибрик А.Е. и др. 2006), в фольклористике (Кляус 2004, Мороз 2003).

[6] Именно богатство и разнообразие жизненных ситуаций, разыгранных в кинематографе, дает нам возможность максимально полно исчислить те параметры, которые следует предусмотреть для разметки звуковых файлов, – стандартные записи устной речи производятся обычно в слишком «тепличных» условиях. Таким образом, совокупность параметров, предусмотренных для описания киноклипов, заведомо будет включать в себя параметры, необходимые для описания реальной устной речи, но не наоборот.

[7] Следствием этого процесса присвоения чужого текста является совершенно однозначное поведение в речи кино т. н. маркеров устной речи (см. Гришина 2007) – по этому параметру речь кино и реальная устная речь практически не отличаются. Свидетельством того, что устная речь в кино не подвергалась и не подвергается никакой особенной редактуре, служит существенное число ошибок, встречающихся в киноречи, – лексических, фактических и стилистических.

[8] Ср. также широко распространенные терминологические сочетания: роли говорящих, речевой сценарий, репертуар речевых жанров и пр.

3. Структура КЗРР

3.1. Единица выдачи и основные типы запросов

Предполагается, что единицей выдачи в КЗРР являются объекты двух типов:

1) морфологически и семантический размеченный текст (в соответствии с параметрами и методикой, принятыми при разметке текстов в НКРЯ) длиной от одного до 3-4-х предложений[9], соединенный посредством гиперссылки с клипом – эпизодом фильма, содержащим соответствующий звучащий текст; эту единицу выдачи в дальнейшем мы называем *клипотекстом*;

Заметим, что текст, соответствующий клипу, будет даваться в стандартной русской орфографии, без использования какой-либо системы транскрибирования, что связано с а) базовыми характеристиками НКРЯ, в рамках которого осуществляется данный проект, – в НКРЯ все тексты выдаются в стандартной русской орфографии, что делает их доступными абсолютно всем пользователям; с б) отсутствием необходимости в транскрибировании звучащего текста, поскольку пользователю доступен сам первоисточник, звучащий текст – соответственно, любой пользователь может при необходимости затранскрибировать его сам, используя свою собственную транскрипцию и с той степенью подробности, которая удовлетворит его самого. Тем самым предложенная система выдачи звучащего текста снимает с создателей КЗРР одну из самых обременительных проблем – проблему разработки последовательной системы транскрибирования звучащего текста.

2) *клип* – эпизод фильма, не содержащий звучащей речи, но содержащий определенный жестовый материал.

Выдача примеров в КЗРР возможна по двум типам запросов.

Метатекстовый запрос. Пользователь может запросить клипы и клипотексты, которые обладают теми или иными метатекстовыми (т. е. не связанными с конкретным словом, морфемой, граммемой, значением) характеристиками (о предварительном наборе таких характеристик – ниже). Дальше стратегия работы пользователя с полученными при выдаче единицами строится в зависимости от его потребностей – он может работать с клипами и клипотекстами как со зрительными и звуковыми объектами, или может на базе полученного материала сформировать пользовательский подкорпус, чтобы на его основе осуществлять запросы второго типа, текстовые.

Текстовый запрос. Строится как стандартный для НКРЯ запрос по стандартным характеристикам: запрос от точной формы, от лексемы, от морфологической характеристики, от семантической характеристики, а также от комбинации этих параметров. В качестве ответа на такого рода запросы выступает набор клипотекстов, с которыми пользователь имеет возможность вести работу, – аналогично тому, как сейчас ведется работа с контекстами, полученными по стандартным запросам в НКРЯ.

Обратим внимание на то, что по текстовым запросам пользователь получает только клипотексты, а по метатекстовым – и клипотексты, и клипы.

3.1.1. Типы текстовых запросов

Типы текстовых запросов в КЗРР не отличаются от тех, которые разработаны для устного подкорпуса НКРЯ, т. е. это стандартные запросы (лексические, морфологические и семантические), принятые в НКРЯ в целом, в сочетании со специфическими возможностями, предусмотренными для устного подкорпуса. Среди последних следует прежде всего отметить возможность поиска по социологическим характеристикам. К последним относятся 1) гендерные характеристики[10] и 2) характеристики по году рождения говорящего. В устном подкорпусе уже сейчас каждому слову, наряду с морфологическими и семантическими, приписываются гендерные и возрастные характеристики (в случае, если они известны), соответственно, возможны запросы типа «использование определенной лексемы в речи мужчин», «использование определенной словообразовательной модели в речи женщин того или иного года рождения» и под. Кроме того, социологическая разметка в подкорпусе «Речь кино» позволяет строить лексические, морфологические и семантические запросы к речи определенного актера, к речи группы актеров, год рождения которых попадает в определенный период, и нек. др.

Все эти типы запросов, как предполагается, будут сохранены в КЗРР, просто, в отличие от устного подкорпуса НКРЯ, при этом полученные на выдаче контексты будут клипотекстами.

---

[9] Следует отметить, что чрезвычайно чувствительной является проблема разбиения фильма на отдельные блоки. На данном этапе мы принимаем решение считать отдельной единицей описания минимальный относительно законченный блок текста/видеоряда, при вычленении которого не приходится насильственно прерывать речь персонажей и параллельный жестовый ряд. Мы надеемся, что будущее устройство КЗРР позволит пользователю, в случае, если ему потребуется расширить контекст выдачи, обратиться к предыдущему и последующему блоку/эпизоду.

[10] Поскольку на начальном этапе работы материалом для КЗРР, как сказано выше, послужат кинотранскрипты, то в качестве говорящих выступают актеры, а гендерные характеристики принимают несколько своеобразный вид, а именно: помимо очевидных характеристик «мужской» и «женский», используются характеристики «мужской-женский» (актер-мужчина, играющий женскую роль) и, соответственно, «женский-мужской» (актриса, играющая мужскую роль).

3.1.2. Типы метатекстовых запросов

Каждый клип и клипотекст в составе КЗРР обладает рядом метатекстовых характеристики.

1) ***Метатекстовые характеристики всего фильма целиком*** (и, соответственно, каждого составляющего этот фильм клипа/клипотекста): авторы (режиссер, автор(ы) сценария, автор исходного текста – при экранизации), год рождения авторов, название, год создания фильма, место расположения киностудии, жанр (все эти характеристики и сейчас используются при описании кинотранскриптов в устном подкорпусе НКРЯ).

2) ***Метатекстовые характеристики клипа, имеющего звуковую, но не имеющего жестовой составляющей*** (чаще всего это касается клипов, содержащих т. н. «голос за кадром», но возможны и случаи, когда персонажи переговариваются в темноте, или в отдалении, так что жесты попросту недоступны для восприятия, в отличие от речи; впрочем, встречается некоторое количество эпизодов, в которых при полноценности речевой составляющей жесты максимально редуцированы).

В этом случае, как нам представляется, клип должен характеризоваться по следующим параметрам (очевидно, что если в клипе есть несколько фраз, то каждая из них получает свой набор характеристик, которые плюсуются при характеристике целого клипа):

– **тип речевых действий,** имеющихся в клипе (предварительно выделяются следующие речевые действия: *аргумент, баюканье, благодарность, брань, возражение, вопрос общий, вопрос частный, вопрос косвенный*[11]*, восклицание, жалоба, запрет, заявление, знакомство, зов, извинение, инструкция, клятва, команда, комментарий, комплимент, констатация, молитва, незнание, обвинение, обещание, обращение, объявление, объяснение, оскорбление, отказ, отрицание, пароль, перечисление, пересказ, подсказка, поздравление, порицание, поучение, похвала, похвальба, предложение, предостережение, предсказание, предупреждение, приветствие, приглашение, признание, призыв, приказ, проводы, проклятье, просьба, прощание, разрешение, рапорт, раскаяние, распоряжение, рассказ, реклама, сентенция, соболезнование, совет, согласие, сообщение, торг, торопить, тост, требование, уговор* (когда персонажи договариваются о чем-то)*, уговоры* (когда один персонаж уговаривает другого что-л. сделать)*, угроза, указание направления, упрек, успокаивать, утверждение, шутка*)[12]. Очевидно, что этот список достаточно велик, однако вполне возможно, что он редуцируется в процессе конкретной работы над разметкой материала, – некоторые из перечисленных речевых действий останутся обязательно, а некоторые, например, *торг, реклама* могут и не потребоваться в виду того, что описываемый текст может оказаться слишком небольшим для таких укрупненных характеристик. В любом случае, вопрос окончательной доводки данной позиции в метатекстовой характеристике может ставиться не априори, а только в процессе работы с конкретным материалом;

– **полнота осуществления речевого действия** (предварительно: действие может оцениваться как *полное*, как *незаконченное* – когда речевое действие добровольно прекращается говорящим, как *прерванное* – когда речевое действие прерывается внешним фактором, включая *автопрерывание*, как *продолженное* – когда слушающий продолжает реплику говорящего; кроме того, здесь же отмечаются случаи *наложения* реплик, а также *вопросы, оставшиеся без ответа*);

– **манера говорения** (здесь могут быть выделены *нормальная речь, крик, шепот, пение, речь с дефектами дикции, ненормально быстрая речь, диктовка, скандирование, декламация, голос за кадром, дубляж*);

– **наличие повторов** (здесь могут быть выделены *однократный–многократный* повтор, *однословный–неоднословный, повтор с интенсификатором, повтор с разной интонацией, переспрос, передразнивание, цитирование,* а также *эхо* – повтор со сменой говорящего и с сохранением типа речевого действия)[13];

– **наличие междометий и вокальных жестов**[14]

(имеются в виду не лексикализованные междометия типа *ах, ох, эх, ай, ой* и под. – включающие их клипотексты могут быть получены посредством обыкновенного лексического запроса через текстовый вход в КЗРР, – а междометия, которые не получили стандартного орфографического воплощения, т.е. разного рода эканья, меканья (заполнители пауз, по терминологии А. Д. Шмелева, см. Шмелев 2005), или маркеры хезитации (Подлесская, Хуршудян 2006), а также причмокивания, цоканье языком, физиологически немотивированные сплевывания (то, что в письменных текстах фиксируется как *тьфу*), свисты (недоуменный, имитирующий

[11] Ср. (Кустова 2007).

[12] При формировании списка использованы работы (Вежбицкая 2007), (Гольдин 2007), (Китайгородская, Розанова 2005), (Шмелева 2007), а также результаты предварительного анализа киноматериала.

[13] В принципе, в настоящий момент в НКРЯ в целом и в устном подкорпусе в том числе, благодаря работе программистов Н. Григорьева и А. Аброскина, возможен поиск редупликаций самого разного свойства, однако по очевидным причинам мы не можем запросить в НКРЯ «редупликацию вообще», без привязки к конкретной лексеме, граммеме и т. д. Что касается КЗРР, то единицей в нем является достаточно компактный объект, и информация о наличии/отсутствии в нем повторов и их типе может быть вынесена в метаописание целого клипотекста.

[14] О последних см. (Шаронов 2006).

быстрое движение (в письменных текстах иногда фиксируется как *фьють, фьюить*), подзывающий и др.) и т.д.);

– **количество говорящих** в клипе;
– **пол говорящих** (*мужской, женский, смешанный*);
– **язык** (*русский, с акцентом, иностранный, квазиязык, тайный язык*);

3) ***Метатекстовые характеристики клипа, имеющие жестовую, но не имеющие звуковой (речевой) составляющей*** (это касается очень частых в кинематографе эпизодов, отражающих неречевое поведение персонажей, а также случаи, когда речь, фактически сопровождая действия героев, реально зрителю недоступна, например, просто выключена).

В данном случае речь не идет о характеристике клипа как целого – напротив, метатекстовыми атрибутами в соответствии с перечисленными ниже параметрами снабжается каждый жест, который разметчик выделяет в клипе. Соответственно, по жестикуляционным параметрам каждый клип получает несколько метатекстовых описаний.

На данный момент логичными нам представляются следующие параметры описания жестов в КЗРР (предложенный ниже набор параметров полностью основан на работах Крейдлин 2004 и Григорьева и др. 2001 – идеи и разработки, изложенные в этих книгах, были лишь адаптированы нами к реальным обстоятельствам, обычно сопровождающим работу над созданием мало-мальски объемных корпусов):

– **социологические параметры** (имя, пол, возраст актера (возраст не точный, а приблизительный – *ребенок, подросток, молодой человек, взрослый, старый*[15]), пол и возраст персонажа);
– **орган, осуществляющий жест** (рука, голова, туловище, нога);
– **активный орган** (рука, голова, кисть, подбородок, глаза и т.д.);
– **пассивный орган** (орган тела говорящего, являющийся необходимой составной частью жеста, но не являющийся активным/движущимся органом, например, *грудь* при жесте «сложить руки на груди»);
– **адаптор** (необходимая составная часть жеста, не являющаяся частью тела жестикулирующего, т.е. своего рода отчужденный от тела жестикулирующего пассивный орган – например, *одежда* при жесте «поправить пиджак», *собеседник* (*голова*) при жесте «погладить по голове», *внешний объект* при жесте «показать пальцем» и т.д.);
– **направление движения** активного органа (обычно это касается таких органов, как голова и руки, – здесь направления задается с помощью наречий *вверх, вперед* (=вдоль направления взгляда жестикулирующего), *назад, вбок, вниз, сверху, спереди, сзади, сбоку, снизу, по кругу, горизонтально, вертикально* и их комбинаций);
– **кратность жеста** (однократный–многократный);
– **название жеста** (в случае, если жест не отрефлектирован в языке, используются условные названия посредством типичных речевых формул, сопровождающих этот жест, – типа жеста *«стоп»*, описанного в цитированных выше исследованиях по русским жестам, или жеста *«иди»* – распоряжения, сделанного с помощью однократного движения подбородка вперед, и др.)[16];
– **тип жеста** (здесь предварительно выделяются следующие типы: 1) *жесты внутреннего состояния* – эмоции, ментальные состояния, 2) *дейктические жесты* – включают в себя, по терминологии Г. Е. Крейдлина, дейктические эмблемы, дейктические иллюстраторы, 3) *изобразительные жесты* – с помощью которых жестикулирующий показывает форму, количество, направление, расстояние, иллюстрирует значение некоторых слов, 4) *регулирующие жесты* – регулируют ход общения и поведение участников коммуникации, 5) *этикетные* – приветствия, прощания, извинения и под., 6) *декоративные* – прихорашивания, оправление одежды, 7) *условные*, или *символические* – жест «на ять», жест ОК, 8) *корпоративные* – пионерский салют, молитвенные позы, бандитская распальцовка, 9) *жесты – речевые действия* – клятва, согласие, несогласие, 10) *физиологические жесты* – зевать, почесываться, 11) *поисковые* – искать кошелек в карманах, вспоминать слово, выбирать выражение, 12) *риторические* – подчеркивающие и иллюстрирующие ритм и отдельные компоненты содержания речи, 13) *пейоративные жесты*);
– **тип коммуникативного действия** (этот параметр существен а) для *этикетных жестов* – именно здесь указывается, в какой именно ситуации осуществляется данный этикетный жест (приветствие, извинение, знакомство, прощание и под.), б) для *жестов – речевых действий* – именно здесь указывается, какое именно речевое действие совершается данным жестом (согласие, отрицание, угроза, утешение, клятва и под.), в) для *пейора-*

[15] Напомним, что точный возраст актера может быть получен из базы данных корпуса на той же странице, где расположена поисковая форма для текстовых запросов

[16] Отметим, что в названии жеста, которое содержит глагол, должна соответствующим образом отражаться кратность жеста – для однократных жестов используется совершенный, для многократных – несовершенный вид (таким образом, например, различаются кивнуть и кивать)..

тивных жестов – указывается, какой именно тип бранного жеста (‘дурак’, ‘сумасшедший’ и под.) используется);
– **тип внутреннего состояния** (очевидным образом параметр имеет отношение только к *жестам внутреннего состояния* и описывает эмоциональные и ментальные состояния жестикулирующего – нежность, удивление, радость, задуматься, догадаться и под.);
– **наличие удлинителя** (наличие необязательного предмета, посредством которого осуществляется жест в данной ситуации, например, удлинитель «головной убор» при жесте «прижать руку к груди», если жестикулирующий держит шляпу в той руке, посредством которой осуществляется жест);
– **наличие спойлера,** или **редуктора** (наличие предмета, который мешает исполнить жест в классическом варианте, например, указание ногой в случае, если руки и голова заняты);
– указание на то, сопровождается жест **улыбкой, смехом, плачем** или нет;
– **полнота жеста** (*полный* жест имеет полный цикл осуществления, *прерванный жест,* в том числе и *автопрерывание*, – прерывается внешними обстоятельствами, *трансформированный жест* по ходу своего осуществления превращается в другой жест);
– **аутентичность жеста** (является ли жест *притворным*; является ли жест *пародией*, *передразниванием* или *зеркальным повторением* чужого жеста).

4) ***Метатекстовые характеристики клипов, имеющих как речевую, так и жестовую составляющую,*** естественным образом являются суммой пунктов 2) и 3).

4. Типы задач, которые могут решаться с помощью КЗРР
4.1. Задачи класса ‘текст звук‘

Естественно, мы не беремся перечислить в данной статье все задачи, которые можно будет решать с помощью КЗРР, – если бы такое перечисление было возможно, то за создание этого корпуса, как представляется, не стоило бы и браться. Хороший корпус, по-видимому, в значительной степени должен жить своей жизнью, которую невозможно было предугадать на стадии его проектирования. Однако некоторые типы задач обозначить все-таки можно.

Прежде всего, КЗРР позволит ставить акцентологические, фонетические и орфоэпические задачи.

1) Как известно, в настоящее время в рамках НКРЯ создается акцентологический подкорпус (или, официально, корпус «История русского ударения»). Этот подкорпус включает в себя русскую силлабо-тоническую поэзию, в которой размечены сильные доли, а также кинотранскрипты, в которых словесное ударение расставлено в соответствии с *реальным* произношением. С внутрисловными ударениями в кинотранскриптах особенных проблем нет, они расставляются достаточно однозначно. Не вызывают трудностей и случаи переноса ударения с полнозначного слова на проклитику. Однако в ряде случаев ситуация с ударениями достаточно неоднозначна – прежде всего это касается сочетания служебных частей речи и местоимений, которые в разных позициях, сочетаниях и смысловых вариантах могут быть то ударными, то безударными. Например, фразы типа *Где он? – Вот он! Что это?* могут произноситься как с отчетливо безударными *он, это* (вплоть до полной редукции гласных, в том числе и потенциально ударных), так и со слабым ударением. Естественно, человек, профессионально занимающийся ударением клитик в современном русском языке, будет испытывать настоятельную необходимость в такого рода точках обращаться к реальному звучанию.

2) Очевидным образом звучащий корпус будет неплохим подспорьем для исследователей и преподавателей фонетики и орфоэпии. Более того, поскольку корпус, как предполагается, будет достаточно сбалансирован с точки зрения хронологии и, кроме того, для значительного числа исполнителей будут указаны года рождения, то на корпусе можно будет ставить своего рода исторические задачи – рассматривать те или иные фонетические явления в их истории от 1930-х годов до сегодняшнего дня. Уже самое первое приближение к использованию звуковой составляющей кинематографического корпуса показывает, что, например, в использовании частицы *вот* в фонетическом варианте [от], т. е. без начального согласного, есть некоторые хронологические закономерности – в фильмах, снятых до 1961-го года, такой произносительный вариант встречается существенно чаще, чем в фильмах, снятых позже[17]. Безусловно, интересно было бы проследить, к примеру, историю произнесения сочетаний заднеязычных (*буХГалтер, К Кому* и под.), стяжений типа *када, тада* (‘когда’, ‘тогда’). И так далее. Фонетисты и специалисты по орфоэпии могут сделать этот ряд примеров практически бесконечным.

3) Для ряда конструкций возможно исследование эмфазы самого широкого свойства (рематической, смысловой и т. д.). В частности, например, интересно распределение логического, фразового и рематического ударения в такой сугубо разговорной конструкции, как *Что, Р, что ли?*, где *Р* может быть равно как одному слову, так и целой фразе.

[17] См. об этом (Гришина 2008).

4.2. Задачи ‘текст жест‘

В рамках КЗРР можно будет ставить, среди многих прочих, вопрос о связи тех или иных жестов с определенным лексемами, граммемами и семантическими множителями, например, исследовать риторические жесты, подчеркивающие указания, в их отношении к указательным и неуказательным единицам речи, использованным в соответствующей фразе/комплексе фраз. Можно исследовать степень обязательности связи того или иного жеста с определенной синтаксической конструкцией. Например, при употреблении конструкции *Вот + вопросительное местоимение* (*Вот как…, Вот где…, Вот почему… Вот о чем…* и т.д.) говорящий часто производит характерное движение подбородком вперед с одновременным поднятием бровей и иногда однократным закрытием глаз. Возникает вопрос, обязательно ли это мимическое движение и почему именно оно используется вместе с данной конструкцией? Представляется, что корпус окажется полезным для решения такого рода проблем.

4.3. Задачи ‘звук‘, ‘жест‘, ‘звук жест‘

Возможности постановки тех или иных задач, связанных с не с текстовыми запросами, а лишь со звуковыми или жестовыми, полностью определяются той разметкой, которая предусматривается для клипов и клипотекстов. Так, на систематическую основу можно будет поставить изучение русской интонации, связанной с разными типами речевых действий, изучение типов повторов, прерываний речи, поведение и семантику нелексикализованных междометий и проч. Что касается жестов, то можно было бы обозначить множество русских этикетных, риторических и т.д. жестов, способы и вариации их осуществления, и т. д. Не будучи специалистами в области жестовой коммуникации, мы не беремся продолжить этот ряд, но подозреваем, однако, что он достаточно длинен. Кроме того, возможно соотнести между собой в запросе разметку клипа как звукового файла и разметку этого же клипа как совокупности жестов. Тогда можно было бы, например, ставить проблему соотношения тех или иных типов речевых действий с теми или иными типами жестов.

Разметка внутреннего состояния говорящего и/или жестикулирующего делает корпус незаменимым инструментом в психолингвистических исследованиях, в частности в изучении способов жестового и речевого выражения эмоциональных и ментальных состояний (именно в сфере описания эмоций мультимедийные корпуса широко используются в зарубежной корпусной лингвистике).

Отдельного анализа требуют возможности использования КЗРР в процессе обучения русскому языку. Очевидно, что такие возможности есть, и они достаточно разнообразны, но этот аспект использования корпуса требует отдельного специализированного обсуждения.

***

На этом можно завершить предварительное описание проекта КЗРР (Корпус звучащей русской речи). Представляется, что нам удалось сформулировать этот проект достаточно законченно и целостно, чтобы, по крайней мере, начать его обсуждение. Добавим, что на настоящий момент нами уже практически разработана технологическая цепочка подготовки материалов для КЗРР – здесь не место ее излагать, однако уже сейчас ясно, что этот проект вполне реален и при благоприятных экстралингвистических обстоятельствах может быть воплощен в жизнь в ближайшие годы.

Список литературы

1. Вежбицкая 2007 – Вежбицка Анна. Речевые жанры (в свете теории элементарных смысловых единиц) // Антология речевых жанров: Повседневная коммуникация. М.: Лабиринт, 2007. С. 68-80.

2. Волошинов 1930 – Волошинов В.Н. Конструкция высказывания // Литературная учеба, 1930, № 3. С. 65-87.

3. Гольдин 2007 – Гольдин В.Е. Имена речевых событий, поступков и жанры русской речи // Антология речевых жанров: Повседневная коммуникация. М.: Лабиринт, 2007. С. 90-102.

4. Григорьева и др. 2001 – Григорьева С.А., Григорьев Н.В., Крейдлин Г.Е. Словарь языка русских жестов. М–Вена: 2001

5. Гришина 2005 – Гришина Е.А. Устная речь в Национальном корпусе русского языка // Национальный корпус русского языка: 2003–2005. М.: 2005

6. Гришина 2007 – Гришина Е.А. О маркерах разговорной речи (предварительное исследование подкорпуса кино в Национальном корпусе русского языка) // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2007» (Бекасово, 30 мая – 3 июня 2007 г.). С. 147-156

7. Гришина 2008 – Гришина Е.А. Частица *вот*: варианты, используемые в непринужденной речи // Что проис-

ходит в современном русском языке (в свете данных языковых корпусов). Slavica Helsingiensia (в печати)

8. Земская и др. 1981 – Земская Е.А., Китайгородская М.В., Ширяев Е.Н.. Русская разговорная речь. Общие вопросы. Словообразование. Синтаксис. М.: Наука, 1981.

9. Капанадзе 1986 – Капанадзе Л.А. Разговорная речь и киноязык // Л.А. Капанадзе. Голоса и смыслы. Избранные работы по русскому языку. М.: 2005. С. 228-231.

10. Кибрик, Подлесская 2003 – Кибрик, А.А., Подлесская, В.И. К созданию корпусов устной русской речи. // НТИ. Сер. 2. 2003, № 10. С. 5–12.

11. Кибрик и др. 2007 – Кибрик А. Е., Архипов А. В., Даниэль М. А., Кодзасов С. В., Майерс Т., Нахимовский А.Д. Технологии обработки языковых данных в документировании малых языков // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2007» (Бекасово, 30 мая – 3 июня 2007 г.). С. 231-235.

12. Китайгородская, Розанова 2005 – Китайгородская М.В., Розанова Н.Н. Речь москвичей: Коммуникативно-культурологический аспект. М.: Научный мир, 2005.

13. Кляус 2004 – Кляус В.Л. К методике видеофиксации фольклора. http://folk.pomorsu.ru/publicat/_complsobir/vlklaus.html

14. Кустова 2007 – Кустова Г.И. Косвенный речевой акт вопроса как средство речевой агрессии и негативной оценки в русской разговорной речи // Культура русской речи. I Международная научная конференция. 15–17 октября 2007 года.

15. Крейдлин 2004 – Крейдлин Г.Е. Невербальная семиотика. М.: 2004.

16. Мороз 2003 - Мороз А.Б. Из опыта работы над базой данных «Традиционная культура Русского Севера (Каргополье)»// Актуальные проблемы полевой фольклористики. Вып. 2. М . 2003. С. 85-99.

17. Подлесская, Кибрик 2004 – Подлесская, В.И., Кибрик, А.А. Транскрипция устного дискурса для нужд корпусных исследований. // Труды международного семинара «Диалог 2004» по компьютерной лингвистике и ее приложениям. Верхневолжский, 2–7 июня 2004 г. http://www.dialog-21.ru/Archive/2004/Podlesskaja.htm

18. Подлесская В.И., Хуршудян В.Г. О лексических маркерах хезитации в спонтанной речи: уроки армянского. // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.). С. 429-439.

19. Русская разговорная речь // М.: 1973 (РРР 1973).

20. Савчук 2007 – Российские разработки корпусов устной речи (Russische Phonokorpora-Modelle) // 1. Symposium „Die phonetisch-phonologischen, orthoepischen und orthographischen Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen“ Graz, 12.–14. April 2007 (в печати)

21. Савчук 2008 – Савчук С.О. Устный корпус: состав и структура. // Национальный корпус русского языка: 2006–2008 (в печати)

22. Седов 2007 – Седов К.Ф. Человек в жанровом пространстве повседневной коммуникации // Антология речевых жанров: Повседневная коммуникация. М.: Лабиринт, 2007. С. 7-38.

23. Шаронов 2006 – Шаронов И.А. Эмоциональные междометия и вокальные жесты // Русский язык сегодня. Вып. 4. Проблемы языковой нормы. 2006. С. 605-617.

24. Шмелев 2005 – Шмелев А.Д. «Заполнители пауз» как коммуникативные маркеры. // Язык. Личность. Текст. М.: 2005.

25. Шмелева 2007 – Шмелева Т.В. Модель речевого жанра // Человек в жанровом пространстве повседневной коммуникации // Антология речевых жанров: Повседневная коммуникация. М.: Лабиринт, 2007. С. 81-89.

26. Якубинский 1923 – Якубинский Л.П. О диалогической речи // Русская речь. Пг., 1923. С. 96-194.

27. Braun et al. 2007 – Sabine Braun, Ylva Berglund Prytz, Kurt Kohn and Pascual Pérez-Paredes. Multimedia Corpora for Applied Linguistic Contexts // Corpus Linguistics 2007. Book of Abstracts. Birmingham, 2007. P. 22.

28. Clavel et al. 2006 – Clavel Ch., et al. Fear-type emotions of the SAFE Corpus: annotation issues // 5th International Conference on Language Resources and Evaluation. 22-28 May 2006. Genoa, Italy. Conference Abstracts. P.76.

29. Devillers et al. 2006 – Devillers L., et al. Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches // 5th International Conference on Language Resources and Evaluation. 22-28 May 2006. 22-28 May 2006. Genoa, Italy. Conference Abstracts. P.76.

30. Grishina 2006 – Grishina Elena. Spoken Russian in the Russian National Corpus (RNC) // LREC2006 5th International Conference on Language Resources and Evaluation. 22-28 May 2006. Genoa Italy. Proceedings. P. 121-124.

31. Grishina 2007 – Grishina E. Text Navigators in Spoken Russian. / Proceedings of the workshop “Representation of Semantic Structure of Spoken Speech” (CAEPIA’2007, Spain, 2007, 12-16.11.07, Salamanca), Salamanca, 2007. P. 39-50.