Е. А. Гришина
Институт русского языка им. В.В. Виноградова РАН
(Россия, Москва)
rudi2007@yandex.ru

МУЛЬТИМОДАЛЬНЫЙ МОДУЛЬ В СОСТАВЕ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА[1]

В статье описывается состав и структура Мультимедийного корпуса русского языка (МУРКО), сложившаяся за годы его существования (2010–2015), основные способы получения информации из корпуса, а также в общих чертах обрисовываются перспективы его развития — количественное пополнение состава, создание программного обеспечения для обработки мультимедийного контента, разработка новых модулей в составе МУРКО — мультимедийных параллельных корпусов.

Ключевые слова: Мультимедийный русский корпус, Национальный корпус русского языка, мультимедийный параллельный корпус, аннотация, инструменты поиска.

1. Введение

В 2010 году в составе Национального корпуса русского языка был открыт пилотный вариант Мультимедийного русского корпуса (МУРКО). Общая идеология будущего МУРКО была изложена в работах [Гришина, Савчук 2008], [Гришина 2008], [Гришина 2009 a], [Гришина, Кудинов 2009], [Grishina 2009a, b]. После открытия пилотного МУРКО корпус рос и развивался; разные этапы его развития были зафиксированы в статьях [Grishina, Savchuk, Sichinava 2010], [Grishina 2010], [Гришина 2011]. В настоящей статье плани-

[1] Работа выполнена при поддержке РФФИ (гранты № 14-06-00245 и № 15-06-04334) и РГНФ (грант № 15-04-12018).

руется описать состав и структуру МУРКО — так, как она сложилась на данный момент (2014 год), — основные способы получения информации из корпуса, а также в общих чертах обрисовать перспективы его развития.

В течение всех лет над пополнением и развитием МУРКО работали С. О. Савчук, С. Б. Иванютин, А. А. Курсакова, Л. Д. Алексеевский, А. А. Махова, М. С. Кудинов, автор статьи и другие коллеги. Огромную благодарность команда МУРКО выражает Алексею Зобнину, Семену Давыдову и Игорю Шалыминову, программистам Яндекса, которые очень творчески, с большим интересом и ответственностью подошли к непростой задаче создания и усовершенствования МУРКО, являющегося, насколько нам известно, на сегодня одним из самых больших открытых мультимедийных корпусов в мире (обзор мультимедийных корпусов см. в статье В. П. Захарова в настоящем сборнике).

2. Состав корпуса

Единицей выдачи в МУРКО является пара «клип+текст» (кликст) или, в случае звуковой дорожки, — пара «фрагмент звукового файла + текст». В ходе подготовки материала для размещения в корпусе мы разрезаем звуковой и видеофайл на небольшие фрагменты, затем разрезаем на такие же фрагменты файлы расшифровок, — и выравниваем между собой мультимедийные и текстовые файлы посредством присвоения им одних и тех же имен. Решение разрезать мультимедийные исходники на фрагменты, а не выкладывать их целиком было принято в связи с тем, что большие мультимедийные файлы, во-первых, медленно загружаются на компьютеры пользователей, т. е. качество использования корпуса начинает в значительной степени зависеть от трафика, а во-вторых, — в больших мультимедийных файлах очень трудно ориентироваться, и, таким образом, пользователь, которого обычно интересует не просмотр исходного мультимедийного файла целиком, а лишь один конкретный, иногда очень короткий, практически точечный фрагмент, вынужден тратить слишком много времени на «перемотку» исходного файла в поисках нужного эпизода. Разумеется, принятое решение имеет ряд недостатков, прежде всего — «обрубленные», иногда весьма неудачно, начало и конец клипа, а во-вторых, — вырванные из более широкого контекста

отдельные высказывания, что для исследования устной речи, понимание которой весьма плотно связано с широкой конситуацией, может оказаться весьма неприятным обстоятельством. Для преодоления этих двух недостатков в корпусе предусмотрен переход к предыдущему и последующему эпизоду, о чем см. в § 4.

По состоянию на апрель 2015 года объем МУРКО составлял 4,3 млн словоупотреблений, или более 187 тыс. кликстов.

ОСНОВНЫЕ ТИПЫ ТЕКСТОВ В СОСТАВЕ КОРПУСА.

1. Речь кино. На сегодняшний день это самая обширная часть корпуса — около 3,4 млн словоупотреблений, 664 кинофильмов/мультфильмов. МУРКО начинал формироваться именно с кинематографической речи — ввиду ее доступности, а также в связи с тем фактом, что полные и точные расшифровки кинофильмов уже были получены в ходе работы над устным и акцентологическим корпусами в составе Национального корпуса. Сейчас перед создателями корпуса стоит задача довести остальные зоны МУРКО до сопоставимых объемов.

2. Устная публичная речь. Объем — чуть более 812 тыс. словоупотреблений. Подкорпус включает в себя телепередачи, документальные фильмы, доклады, лекции и другие образцы устной неподготовленной или слабоподготовленной публичной речи. Подкорпус не включает в себя чтение вслух написанного текста (т. е. тексты типа written-to-be-spoken).

3. Устная непубличная речь. Объем — чуть более 12 тыс. словоупотреблений, включает в себя частные беседы между людьми в разных ситуациях — рассказы, краткие беседы и диалоги. Этот тип текстов — наиболее «дефицитный», поскольку связан с ограничениями экстралингвистического порядка (в частности, нежеланием людей делать свою речь широкодоступной). Создатели корпуса планируют и в дальнейшем пополнять эту зону подкорпуса, однако рассчитывать на ее принципиальное увеличение, по-видимому, не приходится.

4. Авторское чтение. Объем — чуть более 22 тыс. словоупотреблений. Включает в себя авторское чтение художественных и иных произведений (свои рассказы и воспоминания читают Шаламов, Пришвин, Довлатов, Чуковский и др.).

5. Художественное чтение. Объем — примерно 15,4 тыс. сло-

воупотреблений. Включает в себя актерское чтение художественных произведений, в т. ч. и по ролям.

6. Театральная речь. Запись радио- и телеспектаклей, объем около 55,7 тыс. словоупотреблений.

Последние три типа текстов представляют собой в высшей степени интересный источник для орфоэпических, фонетических, акцентологических и интонационных штудий.

В общей сложности материалы МУРКО покрывают временной интервал с 1930 по 2013 год. Естественно, при относительно небольшом объеме корпуса разные периоды представлены достаточно неравномерно. Однако и нынешний объем корпуса позволяет по ряду параметров ставить и решать диахронические задачи.

3. Типы поисковых запросов

МУРКО сохраняет все типы запросов, которые приняты в основном корпусе (лексический, морфологический, семантический поиск и их комбинации), а также в устном и акцентологическом модуле. Здесь мы не будем повторяться — типы соответствующих запросов описаны как в данном, так и в предыдущих сборниках (см. [Гришина, Савчук 2009], [Гришина 2009 б]), — сосредоточимся на тех запросах, которые характерны именно для МУРКО.

3.1. ОРФОЭПИЧЕСКАЯ СТРУКТУРА СЛОВА

Наличие звуковой дорожки в выдаче позволяет пользователю прослушать, как говорящие произносят те или иные сочетания звуков. Для этого в форме поиска существует форма запроса орфоэпической структуры слова (см. рис. 1)

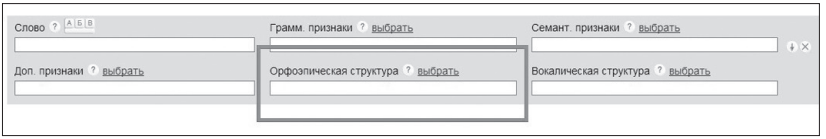

Рис. 1

Основная идея орфоэпического поиска базируется на том, что русская орфография имеет фонематический характер, что позволяет практически в один шаг переходить при корпусном поиске от орфографии слова к

его орфоэпии. Предположим, нам хотелось бы получить данные о том, как произносятся согласные [сз] перед мягкими согласными [тднрл] (очевидно, что поиск такого материала вручную чрезвычайно трудоемок). Нажав ссылку *выбрать* (см. рис. 1), мы получаем диалоговое окно, где формулируем наш запрос (см. рис. 2).

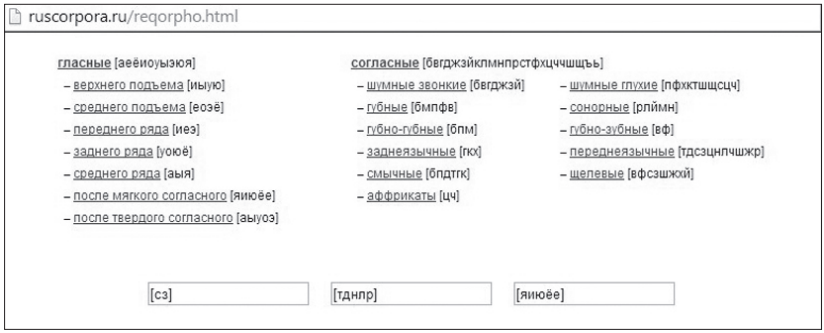

Рис. 2

После нажатия ОК и *Искать* на странице поиска мы получаем на выдаче большое количество клипов (порядка 90 тыс. вхождений), где встречаются искомые звукосочетания (например, *смыСЛе*, *еСЛи*, *поСЛедняя*, *чаСТи*, *иССЛедователей*, *миСТификатор*, *иЗДеваться*, *жиЗНи*, *ЗДесь*, *двеСТи*, *СДелать*, *чаСТичная*, *раЗНица* и многие другие). Этот материал с помощью сортировок можно упорядочить хронологически (и тогда мы получаем данные начиная с 1930 и заканчивая 2013 годом), а также выбрать подкорпус мужской и женской речи, или речи одного специального говорящего, если известно его имя. Кроме того, с помощью указания жанра можно сузить область поиска и искать материал, например, только в некинематографической речи.

Еще один пример. Предположим, нас интересует произнесение геминаты [жж] перед гласными. Орфографически гемината [жж] может быть реализована сочетанием букв *зж* и *жж*, поэтому запрос выглядит следующим образом (см. рис. 3).

Рис. 3

На выдаче мы получаем около полутысячи вхождений (*позже*, *жжение*, *жженка*, *жжет*, *жужжать*, *заезжий*, *изжелта*, *изжить*, *изъезженный*, *можжевельник*, *изжеванный*, *необъезженный*, *обезжирить*, *обезживотеть*, *попозже*, *приезжий* и др.— естественно, не только в исходных, но и в косвенных формах; материал — с 1934 по 2013 год).

Возможен поиск на границе слова. Например, если нас интересует произнесение начального *и-* после слов, заканчивающихся шумной согласной, то запрос будет выглядеть следующим образом (см. рис. 4, где _ обозначает границу слова).

Рис. 4

На выдаче получим порядка 20 тыс. вхождений в интервале с 1930 по 2013 год.

На границе слов можно запросить и сочетание групп, например, запрос на сочетание заднеязычная согласная — граница слова — заднеязычная согласная будет выглядеть следующим образом (см. рис. 5).

Рис. 5

На выдаче получаем порядка 12 тыс. примеров в интервале с 1930 по 2013 год.

3.2. ВОКАЛИЧЕСКАЯ СТРУКТУРА СЛОВА

Специфической особенностью МУРКО является возможность поиска по вокалической структуре слова. Чтобы исследовать редукцию русских гласных в разных позициях относительно ударного слога, фонетистам нужен материал, который представляет собой набор словоформ одной и той же вокалической структуры — вне зависимости от их морфологических и семантических особенностей. Понятно, что без соответствующих поисковых средств отбор такого материала весьма трудозатратен. Именно для осуществления тако-

го отбора предусмотрена опция ***Вокалическая структура слова*** (см. рис. 6).

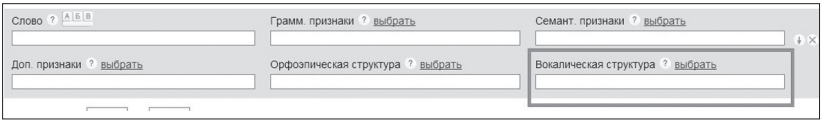

Рис. 6

Предположим, мы хотели бы подобрать фонетический материал следующей структуры: общее количество слогов в словоформе не задано, ударный слог по счету — третий от начала слова, ударная гласная *о*, первый предударный — фонема *о*, первый заударный — фонема *о*. Тогда запрос в диалоговом окне будет выглядеть следующим образом (см. рис. 7).

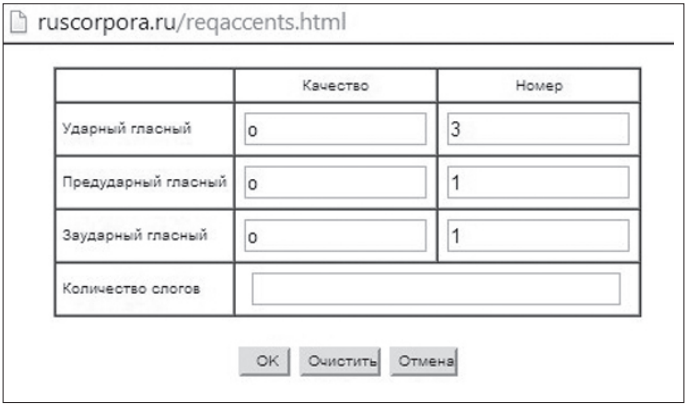

Рис. 7

На выдаче мы получим словоформы *разговорного*, *невозможно*, *паровозом*, *небольшого*, *мирового*, *руководство*, *Циолковского*, *оборотом*, *огромное*, *неспособность* и практически неограниченное, т. е. заведомо превосходящее любые исследовательские потребности число иных словоформ такой же вокалической структуры.

В том случае, если нам требуются данные не по одному, а по нескольким предударным / заударным слогам, запрос формулируется в два приема. Предположим, мы хотели бы получить мате-

риал по запросу: общее количество слогов не определено, ударный слог — пятый от начала, качество ударного слога не определено, все предударные слоги содержат фонему *о*. Сначала заполняем диалоговое окно только для первого предударного слога (см. рис. 8).

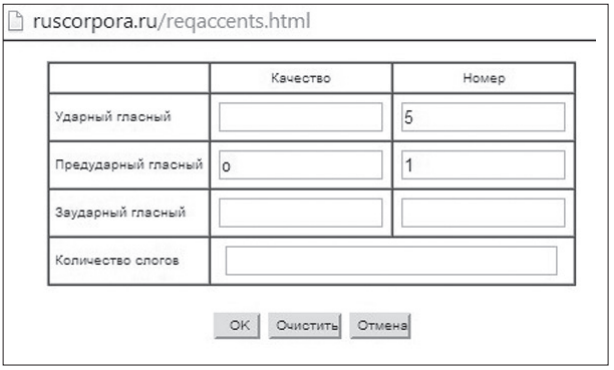

Рис. 8

После нажатия кнопки ОК в поисковой строке видим следующее (см. рис. 9).

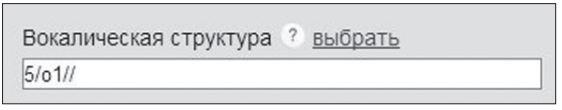

Рис. 9

После этого дописываем руками недостающие слоги прямо в строку поиска (см. рис. 10).

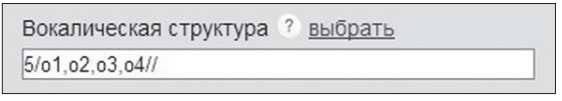

Рис. 10

На выдаче получаем 31 вхождение: *проголосовал*, *проголосовали*, *проголосовать*, *опростоволоситься*, *основоположницей*, *глоттохронологии*, *оплодотворению*, *оплодотворение*, *оплодотворяя*, *оплодотворения*, *оплодотворяются*, *многоговорящие*, *подмногообразии*, *подмногообразия*, *многокомпонентной*, *коротковолновым*,

основоположники, *основоположника*, *поосвобожусь*, *опростоволосилась*, *опростоволосились* (некоторые — по нескольку раз).

3.3. РЕЧЕВЫЕ ДЕЙСТВИЯ И ЖЕСТИКУЛЯЦИЯ

Как мы уже писали ранее, очень небольшая часть МУРКО (глубоко аннотированный МУРКО) имеет расширенную аннотацию, а именно — в нем размечены типы речевых действий и жестикуляция. Эта разметка позволяет отбирать материал и по этим параметрам.

Например, можно выбрать клипы, в которых говорящий осуществляет тот или иной речевой акт. Предположим, мы хотели бы выбрать все клипы, где говорящий благодарит кого-либо или задает общий вопрос. Для этого в разделе Речевые действия на странице поиска мы выбираем *Этикетные высказывания* и далее *Благодарность / Вопросы* и далее *Общие*. Если нам нужно узнать, какие вообще речевые действия размечены и в какую группу они помещены, мы нажимаем на вопросительный знак рядом с закладкой ***Речевые действия*** (см. рис. 11) и получаем полный список речевых актов с разбивкой по группам.

Рис. 11

С помощью разметки речевых действий можно целенаправленно отбирать выказывания определенной семантики, типа иллокуции, перлокутивных функций – для изучения интонационных контуров и иных фонетических особенностей устного высказывания (темп речи, уровень громкости, типы речевого подчеркивания – парцелляция, скандирование – и условия их функционирования, типы междометий и вокальных жестов, типы повторов).

Аналогичным образом устроен жестикуляционный модуль глубоко аннотированного МУРКО. Здесь можно отобрать жесты по их субъективным характеристикам (типу и значению), по объективным характеристикам (активному и пассивному органу, ориентации в пространстве, направлению движения и проч.). Выбрав соответствующие характеристики, пользователь получает

клипы, в которых встречаются жесты заданного типа. Таким образом, МУРКО предоставляет три основных способа получения из корпуса жестикуляционного материала:

1) обращение к глубоко аннотированному МУРКО и запрос жестов с конкретными характеристиками, вне зависимости от того, какой лингвистический ряд их сопровождает, и есть ли такой лингвистический ряд (последний случай — это жесты в режиме пантомимы или жесты слушающего);

2) обращение к тем или иным лингвистическим конструкциям, к отдельным лексемам или классам лексем — с тем, чтобы проанализировать, какие жесты их сопровождают (в частности, в нашей статье в данном сборнике, «О русском жестикуляционном отрицании» был использован именно этот тип запроса,— из МУРКО целенаправленно отбирались фрагменты, включающие в себя разные типы отрицательных конструкций).

Третий способ получения жестикуляционной информации будет нами описан в следующем разделе.

4. Страница выдачи

После формулировки запроса говорящий получает из МУРКО страницу выдачи. Каждый элемент выдачи представляет собой пару «текст расшифровки + соответствующий клип» (в относительно редких случаях, когда клип содержит только жестикуляцию, расшифровка клипа отсутствует; в случае, если элемент выдачи входит в глубоко аннотированный МУРКО, присутствует также таблица жестов, зафиксированных в данном клипе). На рис. 12 показана такая пара:

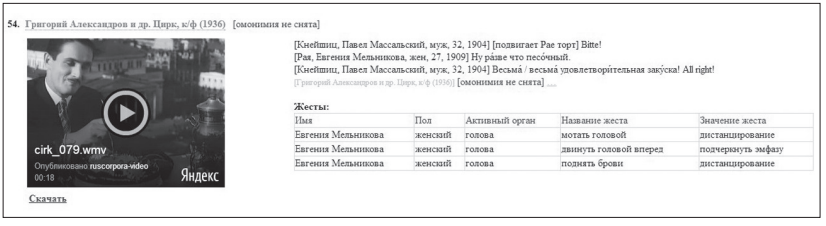

Рис. 12

Клип имеет несколько обозначений, существенных для поль-

зователя. Большая стрелка в кружочке, расположенная в центре, обозначает клавишу проигрыша клипа (при нажатии на нее клип начинает проигрываться). В левом нижнем углу обозначена длительность клипа (на рис. 12–18 сек.) и уникальное имя клипа в базе данных МУРКО (на рис. 12 — cirk_079). В случае, если пользователь захочет сохранить клип на своем компьютере для дальнейшего использования, а не просто просмотреть / прослушать его в режиме он-лайн, ему следует воспользоваться опцией *Скачать* слева под клипом — клип будет сохранен на пользовательском компьютере под своим уникальным именем.

Чрезвычайно важной является функция Расширение контекста (троеточие, которое на рис. 13 взято в кружок).

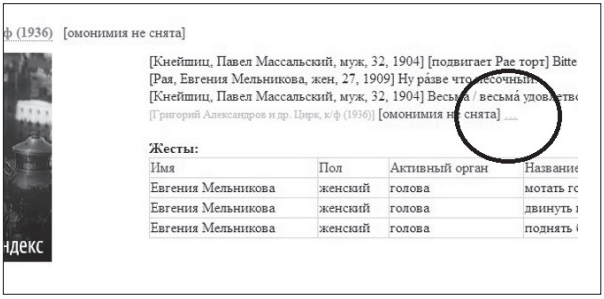

Рис. 13

При нажатии на эту гиперссылку открывается страница, на которой присутствует не набор клипов, а именно этот, выбранный пользователем единичный клип (см. рис. 14).

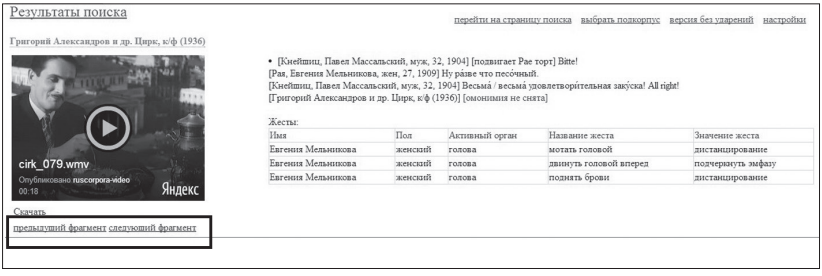

Рис. 14

Такая индивидуальная подача клипа, во-первых, предоставляет возможность получить уникальный адрес данной пары «клип+текст», а во-вторых, просмотреть предшествующий и последующий клипы — в ситуации, когда целый текст при подготовке корпуса расчленялся на небольшие фрагменты, эта проблема является весьма актуальной, поскольку без обращения к пред- и послетексту содержание данного клипа может оказаться абсолютно неясным. Для этого следует обратиться к ссылкам *предыдущий фрагмент* и *следующий фрагмент*, расположенным сразу под клипом (на рис. 14 — в прямоугольнике). Эта же опция предоставляет пользователю возможность просмотреть материал целиком — от клипа к клипу — и тем самым, среди прочего, осуществить третью возможность получения жестикуляционной информации из МУРКО, а именно, — с помощью сплошного просмотра видеоматериала.

5. Развитие МУРКО

5.1. КОЛИЧЕСТВЕННОЕ РАЗВИТИЕ

Мультимедийный модуль должен, прежде всего, развиваться количественно. Ситуация, когда бóльшая часть текстов относится к кинематографической речи, является для ряда исследований существенным ограничителем. Поэтому в перспективе некинематографическая часть МУРКО должна по крайней мере не уступать по объему речи кино. За последние два года в составе корпуса существенно увеличился объем научной речи. Это направление пополнения корпуса следует в обязательном порядке продолжать, но не менее насущным является и добавление других типов дискурса. Прежде всего это касается политической и бытовой речи. При этом при нынешнем уровне развития интернета добавление политического дискурса в МУРКО не представляет практически никаких сложностей и ограничено лишь недостатком ресурсов (людских и материальных) для обработки уже имеющегося в наличии материала. Что же касается бытовой речи, то ее помещение в корпус ограничено, так сказать, нематериально — создатели корпуса не имеют ни юридического, ни морального права помещать «личное» видео в открытый доступ без согласия участников и владельцев этого видео, даже если исходные файлы расположены на открытых площадках, например, на YouTube. Как следствие, принципиальные

изменения в пополнении подкорпуса непубличной речи в МУРКО возможны, по-видимому, только если удастся задействовать краудфандинг, т. е. добровольную передачу в МУРКО видео из личных архивов их владельцами.

Минимальным рабочим объемом МУРКО, который позволит решать бóльшую часть исследовательских задач, на наш взгляд, можно считать объем в 10 млн словоупотреблений, при том что как минимум половина из этого объема должна представлять собой неподготовленную или слабоподготовленную речь максимально широкой тематики.

5.2. КАЧЕСТВЕННОЕ РАЗВИТИЕ

5.2.1. Программное обеспечение

Такой корпус, как МУРКО, — открытый, т. е. бесплатный, и достаточно большой, — на наш взгляд, должен иметь свою программу первичной обработки и описания видео- и аудиоматериала. Разумеется, для таких целей уже созданы и широко используются мощнейшие и при этом бесплатные программы, прежде всего, ELAN, ANVIL, PRAAT, Speech Analyzer. При безусловном высочайшем качестве этих программ они имеют один существенный недостаток (это не касается Speech Analyzer): из-за огромного числа выполняемых функций программы эти слишком сложны и требуют специального обучения. Исследователю же очень часто требуется элементарная проверка собранного из корпуса материала на предмет того, — есть ли какие-то интересные для него явления в отобранном материале или нет, и если нет — то как следует поменять стратегию поиска и структуру запросов.

Предварительное техническое задание для программы мультимедийной разметки

1) Предполагается, что программа будет расположена на сайте Национального корпуса русского языка (точнее — в подкорпусе МУРКО) и будет лицензирована для свободного использования.

2) Задача — создать максимально простое и дружественное средство для разметки аудио- и видеоряда. **На входе** — видеоклип + расшифровка текста в формате xml или txt (если клип не имеет речевого фрагмента — только клип). **На выходе** — таблица Excel.

3) Программа включает в себя окно видео. Предполагается,

что все видео- и аудиоформаты, работающие на компьютере пользователя и на странице МУРКО, опознаются программой, т.е. переконвертации мультимедиа не требуется.

4) Должна быть предусмотрена опция *Кадрировать*, т.е. из видео- аудиофайла большого размера пользователь должен иметь возможность вырезать необходимый ему фрагмент прямо внутри программы. Если пользователю необходимо последовательно работать с разными фрагментами одного и того же клипа, то он должен иметь возможность, не выходя из программы, перезагружать исходный клип, чтобы его снова кадрировать.

5) При проигрывании клипа должно быть предусмотрено удобное замедление прогона (градуальное, с помощью движка).

6) Пользователь должен иметь возможность размечать клип в четырех ипостасях:

— видео

— waveform

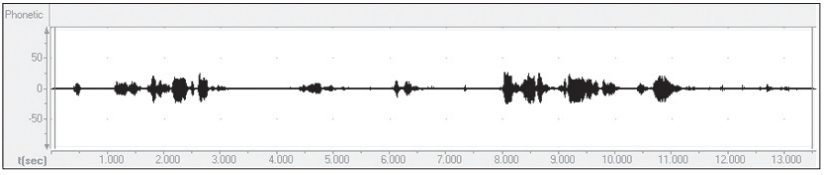

Рис. 15

– pitch (уровень тона)

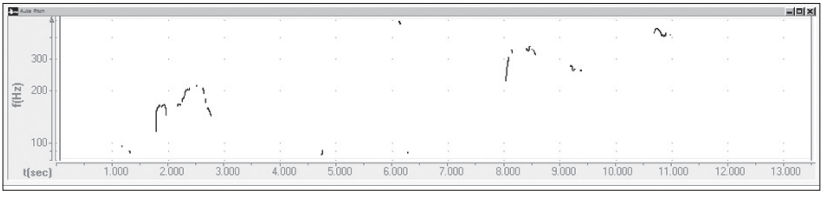

Рис. 16

— текстовом: если на входе в программу подается текст в виде xml, то xml-разметка должна сохраняться внутри программы, но быть невидимой и не доступной редактированию, чтобы при экспорте разметки из программы в результирующую таблицу Excel она

могла перейти туда в первоначальном виде. Кроме того, при экспорте данных из программы в результирующую таблицу Excel у пользователя должна быть возможность снять разметку xml и получить на выходе текст в формате txt.

7) Разметка текстового, видео- и аудиоряда происходит посредством привязки с помощью градуального движка к шкале времени, размеченной в миллисекундах. Таким образом, после проведенной разметки каждое слово (слог, часть слога, пауза), каждый элемент жеста, каждый элемент интонационного контура оказывается привязанным к одной и той же, общей для всех компонентов шкале времени (при этом жестикуляция разбивается на несколько линий — направление взгляда, закрытие/открытие глаз, движение головы, направление движение правой руки, левой руки, обеих рук, конфигурация ладони, ориентация, степень напряжения руки, движение корпуса; можно выбрать отдельную линию или две линии одновременно). Эта привязка к одной и той же шкале позволяет при выгрузке комбинировать данные: текст может автоматически накладываться на waveform, на уровень тона, на жестикуляционные линии, waveform и уровень тона могут комбинироваться с разметкой жестов.

8) Выгрузка разметки.

8.1) Основная идея состоит в том, что размеченный клип/фрагмент клипа должен в один клик выгружаться в таблицу Excel и формировать тем самым заготовку для базы данных, с которой затем пользователь будет работать. Пользователь выбирает, какая пара данных его интересует (например, waveform и текст, или направление движения руки, уровень тона и текст, или текст и направление взгляда, и т. д.), программа соединяет соответствующие данные и для пар, включающих в себя waveform или уровень тона, формирует графический объект (неподвижный или анимированный) под именем исходного видео- или аудиофайла с соответствующим уточнением в расширении. Для пар, включающих в себя текст и разметку жестов, формируется текстовый файл с обозначением жестового ряда, например, в верхнем регистре и/или в скобках — также под именем исходного видео-аудиофайла с соответствующим поясняющим расширением. При этом у пользователя должна быть возможность сохранить или убрать исходную xml-разметку, если она была в загружаемом файле.

8.2) Далее формируется таблица Excel в качестве заготовки для будущей базы данных: первая ячейка – гиперссылка на исходный видео- аудиофайл, вторая и далее ячейки – гиперссылки на файлы из п. 8.1, последняя ячейка – текстовый ряд файла в формате txt

9) Пожелание. Инструкция к программе не должна занимать более пяти страниц.

5.2.2. Мультимедийный параллельный русский корпус (МультиПАРК)

В конце 2014 г. открыт новый модуль в составе МУРКО, ориентированный на сопоставительные исследования.

Как известно, одним из самых серьезных ограничений при изучении устной речи является ее невоспроизводимость, а именно, невозможность получить одно и то же высказывание в одной и той же ситуации от разных говорящих. Это ограничение не соблюдается только для этикетных формул и иных фиксированных стандартизованных социальных реакций, которые в одной и той же ситуации у разных говорящих в значительной степени сходны. Все устные высказывания более сложной структуры уникальны в том смысле, что осуществляются только один раз и, не будучи зафиксированы на те или иные носители, не могут быть воспроизведены вместе с той ситуацией, которая их породила.

При этом чрезвычайно интересным и непростым является вопрос — что в данном устном высказывании является обязательным для любого говорящего, а что колеблется от говорящего к говорящему. Единственным способом приблизиться хотя бы вчерне к ответу на этот вопрос является создание для говорящего возможности произнести данную фразу в одной и той же ситуации. Понятно, что в реальной жизни — за рамками возможных искусственно организованных экспериментов — таких возможностей чрезвычайно мало (если мы, опять же, выходим за пределы этикетных формул). Однако такую возможность нам предоставляет сфера искусства.

Для исследования способов осуществления одного и того же высказывания разными говорящими в одной и той же ситуации (пусть не естественной, а смоделированной) и предназначен МультиПАРК. Предполагается, что МультиПАРК будет состоять из двух зон, направленных на выполнение двух разных, хотя и идеологически сходных задач.

5.2.2.1. Русскоязычный МультиПАРК.

Для проведения сопоставительного анализа русской устной речи предполагается использовать экранизации и театральные постановки одной и той же пьесы. Так, например, наличие трех экранизаций, порядка семи (легко доступных) театральных постановок и нескольких радиопостановок пьесы Н. В. Гоголя «Ревизор» дает нам уникальную возможность выровнять между собой и сопоставить более десятка вариантов произнесения начальной реплики пьесы «Господа, я собрал вас, чтобы сообщить пренеприятное известие: к нам едет ревизор». И таким образом может быть «размножена» не одна только эта фраза, а вся пьеса Гоголя. И не только Гоголя, но и, например, Чехова, Вампилова, Розова, Островского, и вообще любого другого автора, драматургия которого достаточно популярна для того, чтобы а) быть воспроизведенной хотя бы два раза, б) быть записанной на электронные носители, позволяющие воспроизводить данную постановку современными электронными средствами. Сопоставление разных произнесений одной и той же фразы в одной и той же ситуации профессиональными актерами, перед которыми поставлена задача произнести данную фразу максимально естественно, предоставляет нам возможность определить, какие интонационные, темповые, структурные (паузы), фонетические, жестовые особенности этой фразы являются обязательными, т. е. воспроизводимыми всеми говорящими, какие — уникальными или случайными. Кроме того, такое сопоставление дает нам возможность анализировать устный синтаксис с совершенно уникальных позиций, поскольку появляется реальный шанс определить, какие регулярные изменения в порядке слов происходят при переходе от письменного текста исходного драматического произведения к реальному звучанию на сцене или на экране. Естественно, на возможные выводы накладываются ограничения, связанные с искусственностью говорения в театральных и кинематографических условиях, но, тем не менее, определенные выводы, существенные как для понимания устной русской речи, так и русского языка в целом, мы сделать сможем.

5.2.2.2. Англо-русский МультиПАРК.

За три года существования Мультимедийного русского корпуса по его материалам было проведено много исследований, вполне

новаторских, результаты которых были доложены в интернациональных аудиториях. Практически всегда возникал вопрос — зафиксировано ли данное явление, характерное для устной русской речи (фонетическое или жестикуляционное), в других языках, прежде всего в английском? До сих пор никакого обоснованного ответа на вопросы такого рода дать было невозможно, поскольку не существовало корпусов, способных предоставить искомый материал в необходимом количестве.

Как представляется, ответом на данный запрос научного сообщества может стать англо-русский подкорпус МультиПАРКа. Как известно, параллельные подкорпуса и методика их подготовки уже хорошо освоены в Национальном корпусе русского языка. Однако все они ориентированы на письменную речь. Представляется, что МультиПАРК предоставит нам возможность проводить сопоставительные исследования и на устном материале.

Общая архитектоника англо-русского МультиПАРКа. В корпус будут включены расшифровки 1) сериалов и фильмов на английском языке, 2) расшифровки соответствующего русского дубляжа. Каждая видеодорожка (английская и русская) будет разрезана на небольшие фрагменты, одинаковые по содержанию. Каждый видеофрагмент будет выровнен с соответствующим фрагментом английского или русского транскрипта. Текстовые транскрипты при этом будут размечены по методике, принятой в НКРЯ и МУРКО (морфологическая и семантическая разметка для английской и русской зоны, акцентологическая, орфоэпическая разметка и разметка вокалической структуры слова — для русской зоны). Таким образом, за запрос пользователя будут выдаваться две пары «клип+текст» (на английском и русском языках), в которых будут выровнены между собой как видео-, так и текстовый ряд.

Такая подача материала позволит вести сопоставительные исследования в области интонации и фонетики, в области лексики, синтаксиса и семантики (в частности, как известно, английский язык является языком с твердым порядком слов,— даже в устной речи,— в отличие от русского, кроме того, имеет гораздо менее разветвленную систему частиц, которые в устной русской речи играют колоссальную роль; сопоставление устных текстов, полностью совпадающих по содержанию и конситуации, позволит проводить чрезвычайно плодотворные сопоставления столь разных языковых

систем). Несколько более сложным, однако, вполне реальным, будет проведение сопоставительных жестикуляционных исследований на данном корпусе. Во-первых, через поисковые формы корпуса исследователь получит прямой доступ к тем или иным лексемам, морфологическим характеристикам и синтаксическим конструкциям в английском тексте, которые (как и в русской жестикуляции) могут быть связаны с теми или иными жестами. Кроме того, получение жестикуляционных данных для английской жестикуляции будет возможно не напрямую, а через русский параллельный текст. Например, русская указательная частица *вот*, одно из самых распространенных слов в устной русской речи, 1) не имеет хорошего однословного перевода на английский язык, 2) регулярно сопровождается в русской жестикуляции указательными жестами строго определенного типа. Тем самым, получение английского жестикуляционного материала через обращение к русским контекстам с *вот* позволит определить, каким именно образом семантика вот передается в английском тексте и какие именно указательные жесты сопровождают соответствующий семантический компонент.

Литература

Grishina E. Multimodal Russian Corpus (MURCO): general structure and user interface. // NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference, Smolenice, Slovakia, 25–27 November 2009. Proceedings. Tribun, 119–131

Grishina E. Multimodal Russian Corpus (MURCO): types of annotation and annotator's workbenches // Corpus Linguistics Conference CL2009, Universuty of Liverpool, UK, 20-23 July 2009b, http://ucrel.lancs.ac.uk/publications/cl2009/#papers,%20#165

Grishina E. Multimodal Russian Corpus (MURCO): First Steps // 7th Conference on Language Resources and Evaluation LREC'2010, Valetta, Malta/ http://www.lrec-conf.org/proceedings/lrec2010/pdf/143_Paper.pdf

Grishina E., Savchuk S., Sichinava D. Multimodal Russian Corpus (MURCO): Studying Emotions // 7th Conference on Language Resources and Evaluation LREC'2010, Valetta, Malta. Workshop Best Practice for Speech Corpora in Linguistic research/ http://www.corpora.uni-hamburg.de/lrec2012/Proceedings_Complete.pdf

Гришина Е. А. Национальный корпус русского языка как источник сведений об устной речи // Речевые технологии, № 3, 2008, с. 50–62

Гришина Е. А., Савчук С. О. Корпус звучащей русской речи в составе Национального корпуса русского языка // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7(14). По материалам международной конференции Диалог'2008. М., 2008. С. 125–132

Гришина Е. А., Савчук С. О. Корпус устных текстов в Национальном корпусе русского языка: состав и структура // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб., 2009. С. 129–149

Гришина Е. А. Мультимедийный русский корпус (МУРКО): проблемы аннотации // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб., 2009а. С. 150–174

Гришина Е. А. Корпус «История русского ударения» // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб., 2009б. С. 175–214

Гришина Е. А., Кудинов М. С. Инструменты полуавтоматической разметки для Мультимедийного русского корпуса (МУРКО) // Компьютерная лингвистика и интеллектуальные технологии (по материалам ежегодной Международной конференции «Диалог 2009»). Вып. 8 (15), М., 2009. С. 248–261

Гришина Е. А. Мультимедийный русский корпус (МУРКО): современное состояние и перспективы развития // Труды международной конференции «Корпусная лингвистика – 2011», СПб., 27–30 июня 2011. С. 138-144

Захаров В. П. Корпуса русского языка // Труды Института русского языка им. В. В. Виноградова. Вып. 6. М., 2015. С. 20–64.

Elena A. Grishina
Vinogradov Russian Language Institute
of the Russian Academy of Sciences
(Russia, Moscow)
rudi2007@yandex.ru

THE MULTIMODAL MODULE AS PART OF THE RUSSIAN NATIONAL CORPUS

The article describes the Multimodal Russian Corpus (MURCO) as the new project in the framework of the Russian National Corpus (RNC). The pilot version of the MURCO was opened for general access in October, 2010 and since then MURCO has increased its volume up to 4.3 mln tokens and developed its structure. The corpus presently consists of the following subcorpora: 1) Movie speech, 2) Public Spoken Russian, 3) Private Spoken Russian 4) Theater speech 5) Written-to-be-Spoken Russian.

The MURCO is organized as a collection of clixts. A clixt is a pair of an audio- or video clip and the corresponding fragment of the text transcript. A user has the opportunity to download not only the text component of a clixt, but also its sound and video component, so after downloading a user may use any program to analyze it.

MURCO is marked up with different types of annotation. Some of them are standard for the RNC (metatextual, morphological, semantic annotation), some types are special for the spoken component of the RNC (sociological and accentological annotation), and some of the mark-up dimensions are specific only for the MURCO (orthoepic, annotation of vocal structure, speech act and gesture annotation). The speech act and gesture annotation is used in the smaller part of the MURCO, which is called the deeply annotated MURCO.

The article describes the main ways of retrieving information from the corpus and types of search queries. The prospects of further development of MURCO are as follows: increasing the number of texts, developing software for processing multimedia content, creating new modules of MURCO – multimodal parallel corpora.

Key words: Multimodal Russian Corpus, Russian National Corpus, Multimodal Parallel Corpus, annotation, searching instruments.

References

Grishina E. Multimodal Russian Corpus (MURCO): general structure and user interface. *NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference,* Smolenice, Slovakia, 25–27 November 2009. Proceedings. Tribun, 119–131.

Grishina E. Multimodal Russian Corpus (MURCO): types of annotation and annotator's workbenches. *Corpus Linguistics Conference CL2009,* Universuty of Liverpool, UK, 20–23 July 2009. Available at URL: http://ucrel.lancs.ac.uk/publications/cl2009/#papers,%20#165 [accessed 3.05.2015].

Grishina E. Multimodal Russian Corpus (MURCO): First Steps. *7th Conference on Language Resources and Evaluation LREC'2010,* Valetta, Malta. Available at URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/143_Paper.pdf [accessed 3.05.2015].

Grishina E. Multimodal Russian Corpus (MURCO): Studying Emotions. Third International Workshop on emotion: Corpora for research on emotion and affect. *7th International conference on language resources and evaluation (LREC 2010)*. Valletta Malta. Available at: http://www.academia.edu/224975/Multimodal_Russian_Corpus_MURCO_Studying_Emotions (accessed 3.05.2015)

Grishina E., Savchuk S., Sichinava D. Multimodal Parallel Russian Corpus (MultiPARC): Main Tasks and General Structure. *8th Conference on Language Resources and Evaluation LREC'2012*, Istanbul, Turkey. Workshop Best Practices for Speech Corpora in Linguistic research. Available at URL: http://www.lrec-conf.org/proceedings/lrec2012/workshops/03.Speech%20Corpora%20Proceedings.pdf (accessed 3.05.2015).

Grishina E. A. [National corpus of the Russian language as a source of information on oral speech]. *Rechevye tehnologii*, 2008, no. 3, pp. 50–62. (In Russ.)

Grishina E. A., Savchuk S. O. [Speech Corpus within the Russian National Corpus]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Vypusk 7(14). Po materialam mezhdunarodnoi konferentsii Dialog'2008* [Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2008)]. Issue 7(14). Moscow, RSUH Publ., 2008, pp. 125–132. (In Russ.)

Grishina E. A., Savchuk S. O. [Corpus of spoken texts in the Russian National Corpus: the composition and structure]. *Natsional'nyi korpus*

russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy [The National Corpus of the Russian Language: 2006–2008. New results and perspectives]. St. Petersburg, Nestor-Istoriya Publ., 2009, pp. 129–149. (In Russ.)

Grishina E.A. [Multimodal Russian Corpus (MURCO): Problems of Annotation]. *Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy* [The National Corpus of the Russian Language: 2006-2008. New results and perspectives]. St. Petersburg, Nestor-Istoriya Publ., 2009, pp. 150–174. (In Russ.)

Grishina E. A. [Corpus "The History of Russian accent"]. *Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy* [The National Corpus of the Russian Language: 2006–2008. New results and perspectives]. St. Petersburg, Nestor-Istoriya Publ., 2009, pp. 175–214. (In Russ.)

Grishina E. A., Kudinov M. S. [Tools of semi-automatic markup for Multimodal Russian corpus (MURCO)]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii (po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog 2009»* [Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference «Dialogue» (2009)]. Issue. 8 (15). Moscow, RSUH Publ., 2009, pp. 248–261. (In Russ.)

Grishina E. A. [Multimodal Russian Corpus (MURCO): current state and perspectives]. *Trudy mezhdunarodnoi konferentsii* "*Korpusnaya lingvistika — 2011*" [Proceedings of International Conference "Corpus Linguistics — 2011"]. St. Petersburg, St. Petersburg Univ. Publ., pp. 138–144. (In Russ.)

Zakharov V. P. [Corpora of the Russian language]. *Trudy instituta russkogo yazyka im. V. V. Vinogradova* [Proceedings of V. V. Vinogradov Russian language Institute], no. 6. Moscow, 2015, pp. 20–64. (In Russ.)