Е.А. Гришина

МУЛЬТИМЕДИЙНЫЙ РУССКИЙ КОРПУС: СОВРЕМЕННОЕ СОСТОЯНИЕ И ПЕРСПЕКТИВЫ РАЗВИТИЯ[1]

1. Введение

Мультимедийный русский корпус (МУРКО) был открыт для общего доступа в конце декабря 2010 г. На сегодняшний день его объем – 82 фильма, несколько некинематографических текстов, всего чуть более 0,5 млн словоупотреблений (кино), 10 тыс. словоупотреблений (некинематографический материал). В течение 2011 г. МУРКО будет существенно пополнен как кинематографическими, так и некинематографическими текстами. В данной статье мы не предполагаем подробно описывать структуру, состав и поисковые возможности МУРКО, поскольку полагаем, что это достаточно обстоятельно было сделано нами раньше[2]. Мы хотели бы описать направления, по которым, с нашей точки зрения, может развиваться МУРКО в будущем.

[1] Исследование проведено при поддержке РФФИ, гранты 10-06-00151-а и 11-06-00030-а, и программ РАН РФ «Генезис и взаимодействие социальных, культурных и языковых общностей» и «Корпусная лингвистика».

[2] См., среди прочего: *Гришина Е.А.* Национальный корпус русского языка как источник сведений об устной речи // Речевые технологии. 2008. № 3. С. 50–62; *Она же.* Мультимедийный русский корпус (МУРКО): проблемы аннотации // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. 2009. С. 17–214; *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/index.html**, а также инструкцию на сайте корпуса **http://www.ruscorpora.ru/search-murco.html**

2. Пополнение корпуса

Уже сегодня очевидно, что мы не можем ограничиваться включением в МУРКО только кинематографического материала. В частности, исследование лингвистической роли морганий в русской устной речи показало, что кинематографический материал имеет ряд ограничений: тот факт, что в кинематографе говорящий хорошо знает, что именно он скажет дальше, влияет как на частоту морганий, так и на их лингвистическую значимость. Следовательно, для целого класса (психо)лингвистических исследований кинематографического материала явно недостаточно.

Как следствие, одной из основных задач на ближайшие годы является пополнение МУРКО «естественными» устными текстами, как публичными, так и частными. Для этих целей предполагается привлечение материалов из семейных архивов, а также целенаправленный *сбор* разнообразных видеоматериалов.

Помимо простого расширения корпусных данных мы намереваемся ввести в корпус еще один вид текстов, которые до сих пор не фигурировали ни в МУРКО, ни в устной зоне Национального корпуса русского языка в целом. Этот тип текстов получил условное название «Авторское чтение». Предполагается, что в этот подкорпус войдет авторское воспроизведение прозаических и поэтических текстов. Априори кажется, что такого рода материалы не представляют особого интереса для изучения устной речи, поскольку являются простым дублированием соответствующих письменных текстов. Однако более пристальный анализ показывает, что это не так. Прежде всего, авторское чтение показывает, как именно *сам автор* трактует свой текст, в частности, с точки зрения его интонирования и коммуникативного членения. Очевидно, что при достаточно большом объеме зона «Авторское чтение» позволит проводить систематическое изучение параллелей между письменным и устным бытованием одного и того же типа коммуникативного членения фразы или интонационного контура. Среди прочего, материал такого рода позволит анализировать интонационные и коммуникативные паттерны в авторской

речи и в речи персонажей; появится возможность исследовать интонационные и иные устные способы передачи чужой речи, и т.д. Кроме того, авторское чтение, как выяснилось, характеризуется значительным числом отклонений устного варианта от письменного (это, как ни странно, касается не только прозаических, но и поэтических текстов), что дает возможность систематически и на достаточно большом материале исследовать, например, возможности синонимических замен, анализировать «зоны напряжения» в художественном тексте, которые заставляют самого автора при устном исполнении своего текста изменять его.

3. Мультимедийный параллельный корпус (МультиПАРК)

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 стандартизованных социальных реакций фиксированной формы, 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 тех или иных носителях, 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 может колебаться от говорящего к говорящему. Единственным способом приблизиться хотя бы вчерне к ответу на этот вопрос – предоставить говорящему возможность произнести данную фразу в одной и той же ситуации. Понятно, что в реальной жизни – за рамками возможных искусственно организованных экспериментов – таких возможностей чрезвычайно мало. Такую возможность, однако, 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) мы предполагаем разработать новый проект в рамках МУРКО, который получил название «Мультимедийный параллельный корпус», или МультиПАРК.

Предполагается, что МультиПАРК будет включать в себя три основные зоны.

1. Художественное чтение

Эта зона будет логическим продолжением зоны «Авторское чтение», описанной в предыдущем разделе. Предполагается, что сюда войдут авторское, актерское и непрофессиональное исполнение одного и того же поэтического и прозаического текста. Например, эта зона будет включать в себя стихотворение Ахматовой «Один идет прямым путем…» в авторском исполнении, в исполнении Светланы Крючковой, Аллы Демидовой и некоторых других актрис, а также в исполнении тех, кто не является профессиональным актером, а читает это стихотворение «для себя». Сопоставительные корпуса такого рода достаточно представительного объема нам на данный момент неизвестны.

2. Театральные постановки и экранизаци.

Аналогичным образом для МультаПАРКа можно использовать экранизации и театральные постановки одной и той же пьесы. Так, например, наличие двух экранизаций, около десяти театральных постановок и нескольких радиопостановок пьесы Н.В. Гоголя «Ревизор» дает нам уникальную возможность выровнять между собой и сопоставить порядка 20 вариантов произнесения фразы «Я пригласил вас, господа, с тем, чтобы сообщить вам 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 экранизированной или инсценированной хотя бы два раза, б) быть записанной на электронные носители, позволяющие анализировать 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 данного участка текста являются обязательными, т.е. воспроизводимыми всеми говорящими, какие – уникальными, свойственными только данному говорящему. Естественно, на результаты исследования здесь 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.

3. Разноязычный материал

Для специалистов по жестикуляции довольно давно является весьма актуальным вопрос, как проводить межкультурные и межъязыковые жестикуляционные исследования. Функционирование МУРКО показало, что анализ русской жестикуляции на материале русского кинематографа – чрезвычайно продуктивное занятие, которое расширяет как наши сведения о жестикуляции, так и наше понимание мультимодальной природы устной речи[1].

[1] См. *Гришина Е.А.* К вопросу о соотношении слова и жеста (вокальный жест *О* в устной речи) // Компьютерная лингвистика и интеллектуальные технологии (по матер. ежегодной Междунар. конф. «Диалог–2009»). 2009. Вып. 8 (15). С. 80–90; *Она же*. Вокальный жест *А* в устной речи // Компьютерная лингвистика и интеллектуальные технологии (по матер. ежегодной Междунар. конф. «Диалог–2010»). 2010. Вып. 9 (16). С. 102–112; *Она же.* О мультимодальных кластерах в устной речи // Компьютерная лингвистика и интеллектуальные технологии (по матер. Ежегодной Междунар. конф. «Диалог–2011»). 2011. Вып. 10 (17) (в печати).

Естественно, исследование жестикуляции на материале кино имеет свои ограничения (см. выше о морганиях). Но если мы говорим о том, что русская жестикуляция достаточно успешно может быть исследована на материале русского кинематографа, то следующим шагом может быть предположение, что сопоставительные жестикуляционные исследования вполне осмысленно проводить на материале сопоставительных кинематографических исследований. Именно эта идея положена в основу уже не сопоставительной, а собственно параллельной зоны МультиПАРКа.

Предполагается, что в состав параллельной зоны войдут разноязычные экранизации одного и того же произведения (например, американская, английская и русская версии «Войны и мира», русская и французская версии «Анны Карениной», английская и русская экранизации рассказов А. Конан-Дойля о Шерлоке Холмсе, и др.). В каждой из этих экранизаций могут быть вычленены ситуационно и текстуально (с поправкой на разные языки) идентичные эпизоды, которые должны быть выровнены между собой. Очевидно, что при достижении определенного объема такого параллельного корпуса возникнет возможность проводить сопоставительные жестикуляционные (как, впрочем, и интонационные, а также некоторые психо- и социолингвистические исследования). Ввиду абсолютной новизны постановки этого вопроса и отсутствия аналогов в мировой практике трудно предугадать степень полезности и показательности полученного материала, однако попытку вовлечь такого рода сопоставительные данные в лингвистический обиход следует признать вполне осмысленной.

4. Дополнительная разметка

На данный момент МУРКО имеет следующие типы разметки.

1. Стандартная разметка НКРЯ (морфологическая, семантическая, метатекстовая).
2. Стандартная разметка устных текстов (социологическая, акцентологическая).

3. Стандартная разметка МУРКО (орфоэпическая, разметка вокалической структуры).
4. Разметка глубоко аннотированного МУРКО (разметка речевых актов, повторов, междометий и вокальных жестов, манеры говорения и др.; разметка жестов).

Уже эти типы разметки позволяют ставить и решать задачи довольно высокого уровня сложности и получать из корпуса материал, который в отсутствие МУРКО чрезвычайно труднодоступен. Безусловной необходимостью, однако, следует считать разметку в устном материале интонационных контуров и пауз, чтобы иметь возможность быстро отбирать из корпуса однотипный аудиоматериал. Естественно, не может быть и речи о том, чтобы разметку такого плана осуществлять вручную. Таким образом, на сегодняшний день остро стоит задача автоматической разметки интонационных контуров и пауз во включенных в МУРКО клипах, с тем чтобы на основании этой разметки построить соответствующий поисковый интерфейс.