Светлана Олеговна Савчук,
Александра Александровна Махова
(Институт русского языка им. В. В. Виноградова РАН)

МУЛЬТИМЕДИЙНЫЙ МОДУЛЬ В СОСТАВЕ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА: НАПРАВЛЕНИЯ РАЗВИТИЯ[1]

Мультимедийный модуль в составе НКРЯ — это инструмент для изучения устного дискурса. В корпусах, формирующих этот модуль, звучащая речь представлена наиболее объемно и доступна для многоаспектного изучения. Благодаря тому, что в мультимедийных корпусах текстовая фиксация звучащего фрагмента сопровождается аудио- или видеозаписью, исследователь получает возможность изучать как вербальные, так и невербальные компоненты высказывания.

В настоящее время мультимедийный модуль НКРЯ включает три корпуса. Все три корпуса автономны, их объединяет сходство в способах представления материала, но отличают способы его организации.

Мультимедийный корпус (МУРКО)

Основные принципы и этапы работы над корпусом подробно описаны в работах [1–6] и др. Разработка и размещение в открытом доступе пилотной версии корпуса (которая первоначально представляла собой кинокорпус) относится к 2009–2010 гг., в дальнейшем осуществлялось пополнение корпуса и включение в него

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

образцов звучащей речи, функционирующей в различных речевых сферах. В настоящее время объем корпуса приближается к 4,5 млн словоупотреблений. В состав МУРКО входят следующие разделы (подкорпусы).

1. Речь кино включает советские и российские кинофильмы 1930-х —2000-х годов. Общий объем составляет 3,4 млн словоупотреблений.

2. Устная публичная речь — динамично развивающийся подкорпус, объем достигает 815 тыс. словоупотреблений. Наиболее полно представлена устная научная речь: доклады и дискуссии на конференциях, популярные лекции на ТВ и в дискуссионных клубах (проект Academia, Polit.ru), передачи и ток-шоу на радио и ТВ (программа «Гордон» и др.).

3. Устная непубличная речь включает тексты повседневного бытового общения — диалоги и микродиалоги, разговоры в дружеском и семейном кругу, телефонные разговоры и мн. др. Объем этого подкорпуса пока невелик (13 тыс. словоупотреблений).

4. Театральная речь представлена аудиозаписями театральных постановок на сцене и на радио, объем составляет 40 тыс. словоупотреблений.

5. Авторское и художественное чтение — эти два раздела представляют озвученную письменную речь (written-to-be-spoken), которая интересна в плане изучения фонетических особенностей звучащего текста, орфоэпии и акцентологии, интерпретации текста. В разделе собраны записи прозаических текстов в авторском исполнении (С. Д. Довлатов, Ф. А. Искандер, М. М. Зощенко, М. М. Пришвин, К. И. Чуковский и др.) и в исполнении мастеров художественного слова (О. Н. Абдулов, М. И. Бабанова, И. В. Ильинский, А. Г. Коонен, Б. Н. Ливанов, Р. Я. Плятт, Ф. Г. Раневская и др.). Объем этого блока превышает 40 тыс. словоупотреблений.

Звучащий текст в Мультимедийном корпусе представлен в виде аудио- и видеофайла, разрезанного на небольшие фрагменты (клипы) длительностью 10–30 сек., каждому из которых поставлен в соответствие фрагмент текстовой расшифровки. Пара клип + текст (или кликст, по терминологии Е. А. Гришиной) представляет, как правило, относительно законченный в смысловом отношении коммуникативный фрагмент.

Каждый текстовый фрагмент размечен в соответствии со стандартами МУРКО и содержит метатекстовую, морфологическую, семантическую, акцентологическую и социологическую аннота-

цию, по которым возможен онлайновый поиск на сайте. Кроме того, в форме поиска предусмотрена возможность запроса орфоэпической структуры слова и поиска по вокалической структуре слова.

В составе МУРКО выделяется глубоко аннотированная часть, в которой размечены типы речевых действий и жестикуляция (разметка выполнена Е. А. Гришиной). В настоящий момент эта часть включает около 6 фильмов. С помощью разметки речевых действий можно целенаправленно отбирать выказывания определенной семантики (вопросы, императивы, модальные высказывания, этикетные высказывания и мн. др.), типы речевого подчеркивания (парцелляция, скандирование и др.), типы междометий и вокальных жестов, типы повторов. Разметка жестов позволяет отобрать жесты по их субъективным (типу и значению) и объективным характеристикам (активному и пассивному органу, ориентации в пространстве, направлению движения и проч.). Выбрав соответствующие характеристики, пользователь получает клипы, в которых встречаются речевые действия и жесты заданного типа.

Мультимедийный параллельный корпус (МультиПАРК)

Мультимедийный параллельный корпус можно рассматривать как одно из направлений развития МУРКО. Это последний проект, который успела организовать и довести до материального воплощения Е. А. Гришина. Концепция и технология разработки корпуса описана в работах [3, 4, 7]. Корпус состоит из двух независимых зон, которые отличаются как характером материала, так и способом его организации. МультиПАРК сочетает в себе свойства мультимедийного и параллельного корпусов и предназначен для сопоставительных исследований.

Русскоязычный МультиПарК дает возможность сопоставить разные кино-, теле-, радио- и театральные постановки одной и той же пьесы на русском языке. В настоящее время пьеса Н. В. Гоголя «Ревизор» представлена в корпусе тремя экранизациями (1952, 1977, 1996 г.), одним радиоспектаклем и тремя театральными постановками (1982, 1985, 2002 г.).

Технология подготовки корпуса довольно сложна и похожа на подготовку мультиязычного параллельного корпуса письменных переводов одного и того же текста. Поскольку нужно выровнять

несколько текстов, их необходимо привязать к какому-то каноническому, «якорному» тексту, с которым они будут сопоставляться. В мультиязычном корпусе в качестве канона выступает оригинал, в МультиПАРКе — опубликованный текст пьесы. Текст пьесы разрезан на фрагменты, в соответствии с которыми фрагментируется аудио- или видеозапись постановки, а затем каждый аудио- или видеофрагмент выравнивается с его транскриптом. Нумерация фрагментов канона и нумерация соответствующих кликстов в каждой постановке совпадают, что обеспечивает правильную выдачу по запросам. Результаты поиска выдаются в виде кластеров: в каждый кластер входит контекст из канона, содержащий запрашиваемый элемент, и выровненные с ним фрагменты из всех постановок, сопровождаемые соответствующими клипами.

Русскоязычный МультиПАРК предназначен для сопоставительного исследования одной и той же реплики, произнесенной в одних и тех же обстоятельствах разными говорящими. Сопоставление разных произнесений одной и той же фразы предоставляет нам возможность определить, какие интонационные, структурные, фонетические, жестовые особенности этой фразы являются обязательными, воспроизводимыми всеми говорящими, а какие — уникальными или случайными. Кроме того, такое сопоставление дает нам возможность определить, какие регулярные изменения происходят при переходе от письменного текста исходного драматического произведения к реальному звучанию на сцене или на экране. Даже учитывая искусственность ситуации говорения в театральных и кинематографических условиях, можно ожидать, что будут получены результаты, существенные для понимания устной русской речи.

Англо-русский МультиПАРК позволяет сопоставить фильмы на английском языке с их дублированными версиями. В него включены фрагменты англоязычных сериалов с закадровым переводом («Друзья», «Скорая помощь», «Убойный отдел»), а также англоязычные фильмы на английском языке и с русским дубляжем («Как украсть миллион», «Тутси», «Роман с камнем»). Каждый фильм (оригинал и перевод) разрезан на небольшие фрагменты (клипы). На соответствующие фрагменты разрезаны английские и русские расшифровки этих фрагментов. После этого два клипа (английский и русский) и две расшифровки (английская и русская) выравниваются между собой. Нумерация клипов и текстовых фрагментов совпадает в английском и русском варианте.

Каждый текстовый фрагмент размечен в соответствии со стандартами МУРКО и параллельного корпуса НКРЯ и содержит метатекстовую, морфологическую аннотацию (размечены оригинал и перевод), семантическую аннотацию (русский перевод), акцентологическую аннотацию (русский перевод), социологическую аннотацию (сведения об актере — исполнителе роли и актере дубляжа). На запрос пользователя выдаются две пары клип + текст (на английском и русском языках), в которых выровнены между собой видео- и текстовый ряд. Такая подача материала позволяет вести сопоставительные исследования в области интонации и фонетики, лексики и семантики, фразеологии, синтаксиса, анализировать жестикуляцию в англоязычном дискурсе и с помощью сопоставления полученных данных с данными МУРКО проводить сопоставительные жестикуляционные исследования. Кроме того, этот корпус дает образцы особого вида речевой деятельности на русском языке — перевода аудиовизуальных текстов, который рассматривается как самостоятельный вид переводческой деятельности.

Расширение, или количественный рост, корпусов предполагает включение в МУРКО текстов, представляющих разные сферы устной коммуникации, а также текстов, интересных в сопоставительном плане. В частности, в настоящее время ведутся работы по развитию зоны публичной речи МУРКО: концу 2017 г. будет сформирован значительный по объему корпус устной научной и политической речи. Актуальной представляется задача увеличения зоны непубличной коммуникации, при осуществлении которой приходится сталкиваться с проблемами юридического и психологического характера. Подготовлены новые материалы для разделов авторского и художественного чтения, которые будут пополнены в ближайшее время. В более отдаленной перспективе — включение в корпус таких сфер публичной коммуникации, как юридический, религиозный, рекламный дискурс.

В русскоязычный МультиПАРК предполагается включить несколько постановок «Ревизора», а также пьес А. П. Чехова, А. Н. Островского и др. Хороший материал для исследования представляют собой так называемые рассказы-пластинки — несколько версий одной истории, рассказанной одним и тем же человеком или разными рассказчиками. Интерес для изучения фонетических и просодических явлений представляют также образцы чтения одного и того же текста разными чтецами. Обсуждается вопрос о создании раздела поэзии — как выяснилось, возможность

сравнивать разные образцы звучания одного и того же поэтического текста заинтересовала стиховедов. Предполагается, что в этот раздел будут включены записи чтения стихов в исполнении разных чтецов, как профессионалов, так и любителей. Англо-русский МультиПАРК предполагается пополнить новыми фильмами и фрагментами сериалов. Возможность представления в корпусе разных экранизаций (английской и русской) одного литературного источника кажется более сложной задачей и требует предварительного изучения.

Развитие глубоко аннотированного корпуса — другое направление, в котором будет совершенствоваться мультимедийный модуль. В ближайших планах — разметка речевых действий в части научных и политических текстов. Такая разметка расширит возможности изучения научного и политического дискурса, в том числе и в сопоставительном плане. Обсуждается возможность выборочной жестовой разметки. Кроме того, качественное развитие корпуса помимо усложнения аннотации предполагает совершенствование инструментов его анализа. Думается, что работа с корпусами, входящими в мультимедийный модуль, знакомство со всем спектром их возможностей, использование в научных исследованиях и преподавании поставит перед разработчиками новые интересные задачи и определит пути дальнейшего развития.

Список использованной литературы

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

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

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

4. *Гришина Е. А.* Мультимодальный модуль в составе Национального корпуса русского языка // Труды Института русского языка им. В. В. Виноградова. 2015. № 6 (6). С. 65–88.

5. *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»). М., 2009. С. 248–261.

6. *Гришина Е. А., Савчук С. О.* Корпус звучащей русской речи в составе Национального корпуса русского языка // Компьютерная лингвистика и интеллектуальные технологии (по материалам ежегодной Международной конференции «Диалог 2008»). М., 2008. С. 125–132.

7. *Grishina E., Savchuk S., Sichinava D.* Multimodal Russian Corpus (MURCO): Studying Emotions // Proceedings of LREC 2010, Workshop Best Practice for Speech Corpora in Linguistic research.