МУЛЬТИМЕДИЙНЫЙ ПАРАЛЛЕЛЬНЫЙ КОРПУС: ПЕРСПЕКТИВЫ РАЗВИТИЯ[1]

С.О. Савчук
corpusruslan@yandex.ru
ИРЯ им. В.В. Виноградова РАН
(Москва)

Мультимедийный параллельный корпус (МультиПарк) – это самостоятельный корпус внутри мультимедийного модуля НКРЯ. Он предназначен для проведения сопоставительных исследований в самом широком смысле слова. Отбор и организация текстов в корпусе позволяют изучать особенности звучащей речи, произнесенной в одной и той же ситуации разными говорящими, в том числе и говорящими на разных языках. Работы начаты под руководством Е.А. Гришиной в 2014 году [Гришина 2015, Гришина 2016], в настоящее время доступен пилотный корпус и готовится его пополнение. Ниже представлены основные направления развития корпуса.

МультиПарк состоит из двух разделов, каждый из которых имеет свою технологию подготовки и организации материала. Общим в этой технологии является фрагментирование аудио- или видеозаписей и их текстовых расшифровок. Как показал многолетний опыт подготовки и использования мультимедийного корпуса в целом, способ деления материала на фрагменты оказался оптимальным решением и для выравнивания фрагментов при формировании корпуса, и для обеспечения скорости и надежности поиска по корпусу, и для удобства сохранения результатов.

Русский МультиПарк

В настоящее время этот раздел включает 9 постановок пьесы Н.В. Гоголя «Ревизор» в виде видеозаписей театральных спектаклей и фильмов и аудиозаписей радиопостановок. Все расшифровки и видео-/аудио записи фрагментированы и выровнены между собой и сопоставлены с каноном – исходным печатным текстом пьесы. По окончании процедуры подготовки в тексте расшифровки остаются только те фрагменты, которые присутствуют и в каноне. Таким образом, выдача по запросу пользователя обязательно включает фрагмент канона и некоторое количество (от 1 до 9) соответствующих фрагментов из разных постановок, которые сопровождаются клипами.

[1] Работа выполнена при поддержке Программы фундаментальных исследований Президиума РАН «Памятники материальной и духовной культуры в современной информационной среде».

Русский МультиПарк дает возможность сопоставительного изучения одной и той же реплики, произнесенной разными говорящими в одинаковых обстоятельствах. В результате таких исследований могут быть установлены пределы варьирования различных аспектов звучащей речи и ее жестового сопровождения в зависимости от факторов, связанных с личностью актера, временем и стилем постановки, замыслом режиссера и т.д.

В планах развития этого раздела – расширение существующего «театрального» корпуса и создание новых коллекций. В ближайшие два года планируется подготовка нескольких постановок пьесы А.П. Чехова «Вишневый сад». Близкой к «театральному» подкорпусу является готовящаяся коллекция записей чтения одного и того же произведения разными чтецами, среди которых может быть автор текста, профессиональные чтецы, актеры и любители. Другая коллекция представляет собой разные записи одного и того же устного рассказа в исполнении одного рассказчика, что позволит изучать жанр многократно воспроизводимых текстов, или так называемых «рассказов-пластинок». Будут подготовлены рассказы И.Л. Андроникова в записи для радио и телевидения. Корпус таких текстов уже нельзя назвать параллельным в строгом смысле, он будет относиться к разряду сопоставимых корпусов. Сопоставление таких текстов даст богатый материал для изучения вариативности и синонимии в самом широком смысле – не только на уровне лексических и синтаксических единиц, но на уровне целых синтагм, фраз, эпизодов.

Англо-русский МультиПарк

В настоящее время англо-русский МультиПарк содержит фрагменты трех англоязычных сериалов с закадровым переводом и фильмов на английском языке с русским дубляжем. Каждый фильм (оригинал и перевод) разрезан небольшие клипы, английские и русские расшифровки этих клипов также разрезаны на соответствующие фрагменты. Два клипа (английский и русский) и две расшифровки (английская и русская) выровнены между собой. Нумерация клипов и текстовых фрагментов совпадает в английском и русском варианте. В качестве выдачи на запрос пользователь получает выровненные между собой английский и русский тексты в сопровождении соответствующих клипов. Корпус дает возможность изучать вербальные и невербальные компоненты звучащей английской речи в сопоставлении с русской, а также жестикуляцию в англоязычном дискурсе. Сопоставительные жестикуляционные исследования можно проводить путем сравнения полученных данных с данными Мультимедийного корпуса (МУРКО).

Развитие корпуса планируется по двум направлениям – пополнение

существующего подкорпуса новыми фильмами и создание экспериментальных коллекций. Одним из таких экспериментов является подготовка подкорпуса, в котором будут представлены две постановки пьесы А.П. Чехова – на русском языке и в англоязычной интерпретации. Выбирая пьесу, а не экранизацию романа или рассказа, мы исходили из предположения, что интерпретации драматургических произведений окажутся ближе к исходному тексту, чем экранизации прозы, и у нас будет больше совпадающих фрагментов при выравнивании материала. Такой корпус также будет не совсем параллельным, а скорее сопоставимым, поскольку его текстовая составляющая будет представлять собой не пару «оригинал – перевод», а два сценария на разных языках, восходящих к одному источнику. Соответственно усложняется и технология подготовки материалов для корпуса: перед фрагментированием необходимо не только выровнять русский и английский транскрипты видеозаписей, но и сопоставить их с каноном (текстом пьесы и ее переводом). После фрагментирования приводятся в соответствие оставшиеся после выравнивания текстовые и видеофрагменты на двух языках. Материал такого типа будет впервые представлен в мультимедийном модуле НКРЯ. Он даст возможность сравнивать и изучать речевое поведение людей, относящихся к разным культурам, говорящим на разных языках, при этом оказавшихся в сходных ситуациях. Хотя игра на сцене – это не вполне естественное речевое поведение (необходимо учитывать влияние литературного контекста, замысла режиссера и др.), однако положительный опыт использования кинематографического материала в мультиканальных исследованиях дает надежду на то, что результат использования такого корпуса будет стоить затраченных на его подготовку усилий.

Литература

Гришина Е.А. Мультимодальный модуль в составе Национального корпуса русского языка // Труды Института русского языка им. В.В. Виноградова. Вып. 6. – М., 2015. С. 65–87.

Гришина Е.А. Мультимедийный параллельный корпус (МультиПАРК): новый тип корпуса для сопоставительных исследований // Седьмая международная конференция по когнитивной науке: Тезисы докладов. Светлогорск, 20–24 июня 2016. / Отв. ред. Ю.И. Александров, К.В. Анохин. М.: Ин-т психологии РАН, 2016. С. 672–673.