Мультимедийный корпус русского языка: опыт создания и использования[1]

Е.А. Гришина, С.О. Савчук
Институт русского языка им. В.В. Виноградова РАН
Москва, Россия
rudi2007@yandex.ru, savsvetlana@gmail.com

Национальный корпус русского языка, мультимедийный корпус русского языка, использование текстовых корпусов в лингвистических исследованиях и преподавании языка.

The paper introduces the Multimodal Russian Corpus (MURCO), which has been created in the framework of the Russian National Corpus (RNC). The MURCO provides the users with the great amount of phonetic, orthoepic, intonational information related to Russian. Moreover, the deeply annotated part of the MURCO contains the data concerning Russian gesticulation, speech act system, types of vocal gestures and interjections in Russian, and so on. The total structure of MURCO, the types of annotation and the possibilities of its usage in studying and teaching Russian are described.

Мультимедийный корпус русского языка – это электронный ресурс, предназначенный для изучения звучащей речи, «погруженной» в обстоятельства ее произнесения. Основу корпуса составляют видео- и аудиозаписи текстов, выровненные с их расшифровками, что позволяет исследовать не только языковые единицы, но и речевые действия говорящего в различных ситуациях общения, и его неречевое поведение (мимику, жесты, позы). Подобное представление звучащей речи в виде корпусов для русского

[1] Работа выполнена при поддержке Программы ОИФН РАН «Генезис и взаимодействие социальных, культурных и языковых общностей» и РФФИ (грант 10-06-00151-а, 08-06-00371-а).

языка только начинается [см.: Степанова 2008, Гришина 2008], а в виде большого общедоступного корпуса производится впервые [Гришина 2010].

В настоящее время основу корпуса составляют видеоматериалы из отечественных фильмов и аудиозаписи публичной и непубличной устной речи. Технология подготовки материалов для корпуса предполагает расшифровку видео и аудиоматериалов, произведенную с высокой степенью подробности (т.е. включая не только собственно слова, но и междометия, возгласы, а также оговорки); фрагментирование видео и аудио материалов на относительно самостоятельные отрезки (длительностью от 10 до 20 секунд); фрагментирование текстовых расшифровок, или транскриптов; выравнивание мультимедийных и текстовых фрагментов между собой.

Таким образом, единицей выдачи в мультимедийном корпусе служат единицы двух типов: 1) текстовый фрагмент, соединенный гиперссылкой с соответствующим звуковым или мультимедийным фрагментом, эта единица условно называется клипотекстом, или кликстом и 2) мультимедийный фрагмент, содержащий некоторый жестовый материал, но не содержащий текст, т.е. клип из кинофильма.

Клипотексты снабжены принятой в Национальном корпусе русского языка аннотацией – морфологической, семантической, социологической, акцентологической. Поскольку разметка клипотектов стандартная, то по ним возможен обычный для Национального корпуса поиск – по морфологическим,

семантическим категориям, по социологическим параметрам и по их комбинации.

Наряду с метатекстовой разметкой, которая относится к тексту как целому, каждый клипотекст или клип считается отдельным текстом и описывается как отдельный текст с точки зрения его автора, названия, даты создания, жанра, хронотопа и некоторых других.

Однако, кроме того, добавляются и другие, дополнительные характеристики, которые и раскрывают в полной мере своеобразие мультимедийного корпуса по сравнению, например, со стандартным устным подкорпусом. Прежде всего, предлагается некоторая система параметров, характеризующая речевую составляющую клипотекста.

- тип ситуации
- тип речевых действий (вопрос, просьба, извинение, совет и пр.)
- полнота речевого действия (полное, незаконченное, прерванное)
- манера говорения (нормальная речь, шепот, крик, диктовка)
- наличие и типы повторов (однократный, многократный, переспрос, цитирование, передразнивание);
- наличие и типы междометий и вокальных жестов (причмокивание, цоканье, присвистывание, подзывание и под.);
- характеристика говорящих (количество, пол, язык, на котором говорят)

Кроме того, разработана система параметров, по которой описывается жестовая составляющая клипа и клипотекста.

- орган, осуществляющий жест (рука, голова, туловище, нога)
- активный орган (рука, голова, кисть, подбородок, глаза и т.д.);
- пассивный орган рука, голова, грудь
- адаптор (необходимая составляющая жеста, не являющаяся частью тела жестикулирующего, например, *одежда* при жесте «поправить пиджак»);
- направление движения активного органа (вверх, вперед, назад, вбок, вниз, сверху, спереди, сзади, сбоку, снизу, по кругу и пр.);
- кратность жеста (однократный/многократный);
- тип жеста (внутреннего состояния, дейктический, декоративный);
- название жеста (*покачать головой, закатить глаза, махнуть рукой*);
- тип коммуникативного действия/тип внутреннего состояния (приветствие, извинение, прощание, согласие, отрицание, угроза, утешение и под.; нежность, удивление, радость, догадка и под.);
- наличие удлинителя (напр., «головной убор» при жесте «прижать руку к груди», если жестикулирующий держит шляпу в руке);
- наличие спойлера (объекта, мешающего осуществлению жеста в полной мере);
- полнота жеста (полный, прерванный, трансформированный и т.д.);

- аутентичность жеста (притворный, зеркальный, передразнивание и т.д.).

Уникальный материал и система разметки делают мультимедийный корпус мощным исследовательским и обучающим ресурсом. В чем его особенность по сравнению с существующими электронными ресурсами, которые давно используются в преподавании, особенно в преподавании иностранных языков, – аудиокурсами, мультимедийными пособиями, учебными фильмами? Прежде всего в том, что методика использования существующих пособий ограничивается учебными задачами и предполагает в основном имитационные или имитационно-аналитические упражнения. Что касается корпуса, то в нем аудио- и видеоматериалы, выровненные с текстом, снабжены сложной лингвистической и металингвистической разметкой и снабжены инструментом поиска – и это расширяет возможности их использования. Корпус благодаря этому можно использовать не только в учебных курсах, но и в научно-исследовательских целях. Сфера учебного использования корпуса также расширяется – он будет полезен не только в практическом, но и в теоретическом изучении языка – в школьных и вузовских курсах фонетики, орфоэпии, стилистики и культуры речи.

Другим недостатком готовых мультимедийных пособий, так же как и печатных учебных пособий, является то, что в них заложена определенная методика, подбор материала отражает вкус и предпочтения автора. Случаи,

когда такие пособия используются в обучении целиком, без изменений, крайне редки. Обычно преподаватель компилирует разные методические материалы при формировании своего курса. Корпус лишен этого недостатка готовых пособий. Он предоставляет огромный выбор материала и разные способы его извлечения и комбинации, которые можно использовать для составления заданий и упражнений практически любого типа – от имитационных до творческих. Примеры таких заданий с применением мультимедийного корпуса будут приведены в докладе.

Литература

Богданова Н.В. О корпусе текстов живой речи: новые поступления и первые результаты исследования // Труды международной конференции «Диалог 2010» http://www.dialog-21.ru/dialog2010/materials/html/7.htm

Гришина Е.А. Мультимедийный русский корпус (МУРКО): проблемы аннотации // Национальный корпус русского языка: 2006--2008. Новые результаты и перспективы. - СПб., 2009. С. 175-214 http://ruslang.academia.edu/ElenaGrishina/ Papers

Гришина Е.А., Савчук С.О. Корпус звучащей русской речи в составе Национального корпуса русского языка. Проект. // Труды международной конференции «Диалог 2008» http://www.dialog-21.ru/dialog2008/materials/html/19.htm

Степанова С.Б., Асиновский А.С., Богданова Н.В. и др. Звуковой корпус русского языка повседневного общения «Один речевой день»: концепция и состояние формирования // Труды международной конференции «Диалог 2008» http://www.dialog-21.ru/dialog2008/materials/html/76.htm

Grishina, E. Multimodal Russian Corpus (MURCO): First Steps At: http://www.lrec-conf.org/proceedings/lrec2010/pdf/143_Paper.pdf