А.А. Кибрик, О.В. Федорова, В.И. Подлесская

Мультиканальные корпуса: вчера, сегодня, завтра

Введение. Мультиканальная лингвистика

Согласно традиционному подходу XX в., который до сих пор преобладает в современной лингвистике, человеческий язык сводится к набору иерархически организованных единиц – фонем, морфем, слов, словосочетаний и предложений. В рамках такого представления языковая форма отождествляется с вербальной, т. е. с сегментным материалом, относящимся к вокальной (слуховой) модальности[1]. Однако, как известно, помимо звукового сигнала существуют и другие компоненты естественной коммуникации, в первую очередь те, которые связаны с кинетической (зрительной) модальностью.

В последние годы, однако, традиционная точка зрения XX в. постепенно уступает место новой *мультиканальной*[2] перспективе (Scollon 2006; Knight 2011; Abuczki, Esfandiari 2013; Müller et al. eds. 2014; Кибрик в печ.), согласно которой для успешной языковой коммуникации важны и, следовательно, заслуживают внимания все каналы передачи информации: вербальные единицы, просодия, жесты, мимика, направление взора и т. д. При нашем мультиканальном подходе (Кибрик 2010; Кибрик в печ.) выделяются два вокальных (слуховых) канала – вербальный и просодический, а также группа кинетических (зрительных) каналов. Под *вербальным* каналом мы понимаем весь речевой материал, который в конечном счете сводится к последовательности

Работа выполнена при финансовой поддержке РНФ, грант №14-18-03819 «Язык как он есть: русский мультимодальный дискурс».

фонем. К *просодическому* каналу относятся несегментные аспекты звука – интонация, дискурсивные акценты, громкость, тембр и т. д. (Кодзасов 2009; Кибрик, Подлесская ред. 2009). *Кинетическая* составляющая, или язык тела, имеет сложную организацию, которая включает как динамический компонент – различного рода движения и жесты (в том числе мимические жесты, жесты головы, рук и тела), так и статический – выражение лица, позы, проксемику.

Одно из наиболее важных и перспективных направлений развития мультиканальной лингвистики – это разработка и создание *мультиканальных корпусов*. В отличие от моноканальных и мономодальных корпусов, уже имеющих свою историю и традицию, мультиканальные корпуса в настоящий момент находятся на этапе становления. Данная статья посвящена описанию (1) методологии создания мультиканальных корпусов, (2) нескольких наиболее известных западных и отечественных мультиканальных корпусов, а также (3) нашего текущего проекта «Язык как он есть», осуществляемого в Институте языкознания РАН.

1. Мультиканальные корпуса: принципы создания и типология

Термин «мультимодальная коммуникация» появился в конце 1980-х годов (см., в частности, Taylor 1989). Согласно определению, мультимодальный (в нашей терминологии «мультиканальный») корпус – это «аннотированное собрание скоординированной информации из разных коммуникативных каналов, включая речь, направление взора, мануальные жесты и язык тела, которое обычно создается на материале записей человеческого поведения» (Foster, Oberlander 2007: 307–308)[3]. В отличие от моноканальных (письменных, основанных на вербальном канале) и мономодальных (речевых, основанных на вербальном и просодическом каналах, принадлежащих к вокальной модальности) корпусов параметры, по которым можно классифицировать

мультимодальные корпуса, еще только вырабатываются. Ниже мы перечислим четыре из них: объем, характер общения, количество собеседников и среда общения.

Если сравнивать мультимодальные корпуса с мономодальными, они все проигрывают им с точки зрения своего *объема* – многомиллионных (в словах) мультимодальных корпусов не существует; согласно официально приводимым данным, самый большой заявленный объем подобного корпуса, а именно, англоязычного корпуса *AMI Meeting Corpus*, составляет 100 часов (Ashby et al. 2005; Carletta 2006), однако бóльшая часть информации представлена в нем в виде неразмеченных файлов; подробнее об этом корпусе см. раздел 2). Практически все лингвисты, работающие в области корпусной лингвистики, разделяют мнение о том, что «в некоторых областях лингвистики, например, в синтаксисе, должны использоваться корпуса объемом в несколько миллионов слов, в то время как в области просодии, в которой аннотирование производится вручную, несколько часов размеченной записи считается большим объемом» (Blache et al. 2008: 110); то же относится и к разметке компонентов кинетических каналов. Соглашаясь с этим высказыванием относительно объема мультиканальных корпусов, мы полагаем, однако, что на основе анализа устной речи должна быть подвергнута ревизии и область синтаксиса, и, шире, – грамматики естественного языка, так как традиционная грамматика основана на письменных литературных текстах, представляющих собой вторичный и весьма своеобразный вид использования языка.

Характер общения собеседников удобно изображать в виде шкалы от контролируемых экспериментов на левом краю до ничем не ограниченного общения на правом. На самом левом краю находится *Czech Audio-Visual Speech corpus* (Železsny et al. 2006), созданный для тестирования системы распознавания речи и включающий 25 часов записи 65 испытуемых, читающих вслух по 200 предложений. Правее расположен *Fruit Carts Corpus* (Aist et al. 2012), в котором записано 240 видеороликов продолжительностью

4–8 мин каждый. Испытуемые выполняли стандартное задание – инструктор давал раскладчику инструкции по раскладыванию карточек с нарисованными на них фруктами; данный корпус был создан в первую очередь для изучения временной дискурсивной неоднозначности и референциальных стратегий. Еще правее на шкале естественности располагается англоязычный корпус *D64*, в который вошли записи коммуникации между 5 испытуемыми, сделанные по 4 часа каждая с интервалом в два дня, т. е. в общей сложности 8 часов записи. Этот корпус, созданный для изучения бытового социального общения, включает спонтанные диалоги, записанные в домашней обстановке (Campbell 2009). В комнате, в которой происходила запись, было установлено 7 видеокамер, 10 микрофонов, а также установки для регистрации движений головы, рук и тела. Корпус *InSight Interaction* (Brône, Oben 2015), включающий 15 диалогов по 20 мин. каждый, находится еще немного правее (подробнее см. раздел 2). На правом краю находятся корпуса, созданные в традиции анализа бытового диалога (Conversation Analysis, см. Mondada 2014). Наконец, на самом правом краю находится корпус, который может быть создан по методологии корпуса «Один речевой день» (Богданова и др. 2010), т. е. в ничем не ограниченных условиях коммуникации.

Кроме объема и характера общения выделяются также такие параметры, как *количество собеседников (2 vs. 3+)* и *среда общения* (специально созданные условия для проведения записей vs. неподготовленная среда). Три последних параметра важно оценивать с точки зрения *естественности* коммуникации. Наиболее естественные данные собираются в ходе бытового общения трех и более собеседников в неподготовленной среде.

Наконец, важно подчеркнуть, что согласно мнению авторитетного исследователя мультимодальности Д. Найт, большинство существующих корпусов создаются в узких исследовательских целях и дают ответы только на частные вопросы, а стандартная процедура сбора, аннотирования и проведения исследований еще не разработана (Knight 2011:

403). Кроме того, Найт отмечает, что на сегодняшний день ни одного большого мультимодального корпуса нет в свободном доступе.

2. Мультиканальные корпуса: показательные примеры

Корпус AMI. Данный корпус (объемом около 100 часов) был создан в Эдинбурге для изучения социального общения небольших групп людей и состоит из записей разнообразных по тематике рабочих встреч (https://www.idiap.ch/dataset/ami), в которых участвовали три и более человек; для большинства участников коммуникации английский язык не был родным. Запись велась при помощи нескольких видеокамер и диктофонов, расположенных на разном расстоянии от собеседников, а также интерактивных экранов для презентаций (Ashby et al. 2005; Carletta 2006).

Транскрибирование размеченной части корпуса осуществлялось в полуавтоматическом режиме в несколько проходов; ручная часть аннотирования была выполнена при помощи специально разработанной транскрипции, включающей в том числе неправильное произнесение слова, речевые сбои, смех, невербальные вокализации, заполненные паузы хезитации и некоторые другие. Что касается кинетической разметки, в корпусе в автоматическом режиме фиксировались движения рук, головы и ног, а также движения глаз. Размеченная часть данного корпуса имеет подробную аннотацию социального и речевого взаимодействия (в том числе разметка речевых актов, деление на топики, выделение резюме, фокус внимания, перемещения людей, их эмоции), минимальную аннотацию кинетических каналов и дискурсивную транскрипцию.

Корпус InSight Interaction. В ходе данного проекта авторы собрали 15 записей диалогов на нидерландском языке по 20 минут каждый; каждая запись состояла из двух частей:

решения совместной когнитивной задачи и свободного общения (Brône, Oben 2015).

На каждого из двух собеседников был надеты очки-айтрекер фирмы Arrington Gig-E60 с частотой 30 к/с и разрешением 320x240; кроме того, общий план фиксировался видеокамерой Sony HDRFX1000E с частотой 25 к/с и разрешением 720x576.

Полученные записи были аннотированы в программе Praat на вокальном уровне и в программе ELAN на кинетическом уровне. С точки зрения просодии авторы делили речь на интонационные единицы и аннотировали главный акцент, интонационные контуры, удлинения и различного рода паузы. Для аннотирования мануальных жестов ими была разработана аннотационная схема, включающая как формальные признаки жеста (сегментация на unit / phrase / phase, рукость, форма руки, направление и место производства жеста), так и их функцию (иконические vs. дейктические vs. символические жесты).

Англоязычный корпус, описанный в работе Holler, Kendrick 2015. В рамках сбора данного корпуса записей естественной коммуникации было записано десять пар и десять троек коммуникантов, длительность каждой записи была примерно 20 мин.

Речь каждого собеседника фиксировалась на индивидуальный микрофон, также велась общая аудиозапись. Три видеокамеры Canon Legria 293 HFG10 с частотой 25 к/с снимали каждого из коммуникантов. Кроме того, каждый коммуникант был в очках-айтрекерах фирмы SMI, частота 30 к/с.

Собранные аудио- и видеофайлы были синхронизированы в программе Adobe Premier Pro. Полученные записи были аннотированы в программе Praat на вокальном уровне и в программе ELAN на кинетическом уровне. Данный корпус максимально похож на наш собственный (см. раздел 3).

Мультимедийный русский корпус (МУРКО, созданный под руководством Е.А. Гришиной). Корпус, включающий

фрагменты кинофильмов 1930–2000-х годов, был создан в конце 2010 г. в составе НКРЯ. На январь 2014 г. этот корпус насчитывал 4 млн словоупотреблений, т. е. является самым большим в мире мультимодальным корпусом. Данные представлены в виде параллельных видеоряда, аудиоряда и текстовой расшифровки звучащей речи, а также наблюдаемых в кадре жестов. В корпусе возможен поиск не только по произносимому тексту, но и по жестам (кивание головой, похлопывание по плечу и т. п.) и типу речевого действия (согласие, ирония и т. п.) (Гришина 2017).

С точки зрения сбалансированности данный корпус обладает подробной жестовой аннотацией, однако его вокальная составляющая (т. е. вербальный и просодический каналы) ограничена орфографической записью (за исключением вокалической структуры слова с 4 значениями: ударный слог; предударный слог; заударный слог; количество слогов в слове).

Русскоязычный эмоциональный корпус (REC). Корпус состоит из 295 видеозаписей устных университетских экзаменов и 510 случаев общения с клиентами в службе одного окна ГУ ИС г. Москвы (http://www.harpia.ru/rec/, Котов 2014). В корпусе размечаются речь, мимика и жесты участников диалога; данный корпус является узкоспециальным, так как «принципы разметки корпуса состоят в том, чтобы выделить те особенности, которые отличают эмоциональное поведение людей от некоторого “воображаемого” нейтрального поведения» (http://www.harpia.ru/rec/). Разметка осуществляется в программе ELAN по следующим слоям: текст 1 (орфографическая запись), фазы речи (хезитации, междометия, поправки), текст 2 (текст оппонента), текст 3 (текст еще одного участника), голова, глаза (11 значений, в том числе ‘взгляд вверх’, ‘хмурит брови’), рот (17 значений, в том числе ‘смеется’, ‘трубочка’, ‘кусает губу’), руки – активный орган, руки – пассивный орган, способ, траектория, поза, комментарий, остроты (смех, ирония), микросостояния, фазы микросостояний.

Как можно видеть, аннотация данного корпуса почти полностью сконцентрирована на кинетической составляющей при отсутствии просодической разметки и с минимальной разметкой вербального канала.

3. Мультиканальный корпус «Язык как он есть»

В нашем текущем проекте, осуществляемом в ИЯ РАН (сайт проекта multidiscourse.ru), существенное место отводится всем основным мультиканальным составляющим, в том числе просодической (Кибрик и Подлесская ред. 2009), жестовой (Литвиненко и др. 2017) и окуломоторной (Федорова и др. 2016). Цель данного проекта двоякая: во-первых, создается ресурс нового типа; во-вторых, на основе этого ресурса реализуется научная программа изучения реальной мультиканальной коммуникации.

В качестве *стимульного материала* при сборе корпуса был использован известный «Фильм о грушах» У. Чейфа; коллективная монография под его редакцией «Рассказы о грушах: Когнитивные, культурные и языковые аспекты порождения повествования» является одной из самых известных работ в области анализа дискурса (Chafe ed., 1980). Изданная в 1980 г. по итогам пятилетней работы большого коллектива авторов, она во многом задала направление дискурсивных исследований конца XX – начала XXI в. Специально созданный для этих целей «Фильм о грушах» не содержит звучащей речи, а показанные в нем события в целом понятны жителям практически любого уголка земного шара. Кроме того, видеоряд был подобран таким образом, чтобы стимулировать испытуемых порождать описания пейзажа, причинно-следственных отношений, а также мыслей и эмоций героев повествования.

Использовав этот фильм в качестве стимульного материала, мы разработали новую *методику* проведения исследования. В каждой записи принимали участие четыре человека с заранее распределенными ролями.

Три участника – Рассказчик, Комментатор и Пересказчик – участвовали в основной части записи, а четвертый – Слушатель – присоединялся в конце. Сначала Рассказчик и Комментатор смотрели каждый на своем ноутбуке шестиминутный «Фильм о грушах» и старались как можно лучше его запомнить. Затем к ним присоединялся третий участник – Пересказчик – и начиналась основная часть записи. Задача Рассказчика состояла в том, чтобы рассказать сюжет просмотренного фильма Пересказчику, который этот фильм не смотрел; это был этап *рассказа* в режиме монолога. На следующем этапе – *разговора* – Комментатор дополнял рассказ Рассказчика и при необходимости исправлял его, а Пересказчик уточнял у Рассказчика и Комментатора необходимые детали. Наконец, на этапе *пересказа* Пересказчик пересказывал сюжет Слушателю, опять в режиме монолога. После этого Слушатель письменно фиксировал на бумаге услышанный пересказ. Таким образом, основная задача каждого участника была максимально подробно и понятно донести до других полученную им информацию.

Важной особенностью корпуса было использование высокоточного *оборудования*. Речь испытуемых фиксировалась на *шестиканальном рекордере* ZOOM H6 с параметрами записи 96 kHz / 24 bit; речь каждого из трех говорящих записывалась на индивидуальный петличный микрофон SONY ECM-88B; кроме того, отдельно велась общая стереозапись с микрофона диктофона. Три *промышленные видеокамеры* JAI GO-5000M-USB с частотой 100 к/с и разрешением 1392x1000 записывали крупным планом каждого из трех основных участников; эти камеры позволяют получить запись в формате mjpeg; данный формат выгодно отличается от остальных отсутствием межкадрового сжатия, что является необходимым условием для дальнейшего покадрового аннотирования; кроме того, камера GoPro Hero 4 с частотой 50 к/с и разрешением 2700×1500 записывала общий план. Для регистрации движений глаз были использованы две пары *очков-айтрекеров* Tobii Glasses II Eye Tracker с частотой 50 Hz и разрешением видеокамеры 1920×1080.

Один из двух айтрекеров был надет на Рассказчика, причем запись также велась и во время просмотра им «Фильма о грушах»; второй айтрекер был надет на Пересказчика. Данная модель айтрекеров выпускается с декабря 2014 г. и активно используется в маркетинговых и спортивных исследованиях, а также в исследованиях безопасности вождения автомобилей. Насколько нам известно, подобные айтрекеры еще не были использованы в когнитивных исследованиях мультимодальной коммуникации.

Собранный корпус «Рассказы и разговоры о грушах» включает 24 записи (суммарной длительностью 9 часов). Вокальная составляющая корпуса была аннотирована в программе Praat, кинетическая – в программе ELAN. К числу наших *актуальных задач* относится качественное развитие мультиканального ресурса, а также согласование подходов в области анализа вокального и кинетического компонентов. Проект должен послужить развитию теории коммуникации и теории дискурса и, в частности, помочь уточнить и переинтерпретировать ряд базовых понятий, таких как разграничение адресанта и адресата, чередование реплик, паузация, дискурсивная единица. В рамках проекта ставится также широкий круг конкретных задач – уточнение роли просодии как интерфейса между речью и кинетическим поведением, развитие аннотации кинетического поведения, исследование плавности речи и жестикуляции, изучение координации речевых и жестовых единиц.

Литература

Богданова Н.В., *Асиновский А.С.*, *Маркасова Е.В.*, *Степанова С.Б.*, *Супрунова А.В.*, *Шерстинова Т.Ю.* Звуковой корпус русского языка «Один речевой день»: пути пополнения и первые результаты исследования / А.Е. Кибрик и др. (ред.) // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». М.: РГГУ, 2010. С. 41–46.

Гришина Е.А. Русская жестикуляция с лингвистической точки зрения (корпусные исследования). М., 2017.

Кибрик А.А. Мультимодальная лингвистика // Когнитивные исследования. Вып. IV. М., 2010. С. 134–152.

Кибрик А.А. Русский мультиканальный дискурс как перспективный объект исследования // Психологический журнал. (в печ.)

Кибрик А.А., Подлесская В.И. (ред.) Рассказы о сновидениях: корпусное исследование устного русского дискурса. М.: ЯСК, 2009.

Кодзасов С.В. Исследования в области русской просодии. М.: ЯСК, 2009.

Котов А.А. Коммуникативное поведение при ответе на сложный вопрос в эмоциональном диалоге / О.В. Федорова, А.А. Кибрик (ред.) // Мультимодальная коммуникация: теоретические и эмпирические исследования: Сборник статей. 2014. С. 74–85.

Литвиненко А.О., Николаева Ю.В., Кибрик А.А. Аннотирование русских мануальных жестов: теоретические и практические вопросы // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2017». М.: РГГУ, 2017.

Федорова О.В., Кибрик А.А., Коротаев Н.А., Литвиненко А.О., Николаева Ю.В. Временная координация между жестовыми и речевыми единицами в мультимодальной коммуникации // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2016». М.: РГГУ, 2016. С. 159–170.

Abuczki A., Esfandiari B.G. An overview of multimodal corpora, annotation tools and schemes // Argumentum. 2013. 9. P. 86–98.

Aist G., Campana E., Allen J., Swift M., Tanenhaus M.K. Fruit Carts: A Domain and Corpus for Research in Dialogue Systems and Psycholinguistics // Computational Linguistics. 38 (3). 2012. P. 469–478.

Ashby S., Bourban S., Carletta J., Flynn M., Guillemot M., Hain T., Kadlec J., Karaiskos V., Kraaij W., Kronenthal M., Lathoud G., Lincoln M., Lisowska A., McCowan I., Post W., Reidsma D., Wellner P. The AMI Meeting Corpus. Proceedings of Measuring Behavior 2005. Wageningen, 2005. P. 4–8.

Bavelas J. B., *Chovil N.*, *Lawrie D.*, *Wade A.* Interactive gestures // Discourse Processes. 15 (4). 1992. P. 469–489.

Blache P., *Bertrand R.*, *Ferré G.* Creating and exploiting multimodal annotated corpora. Proceedings of Sixth International Conference on Language Resources and Evaluation (LREC) 2008 [online]. P. 110–115.

Brône G., *Oben B.* InSight Interaction. A multimodal and multifocal dialogue corpus // Language Resources and Evaluation. 2015. 49(1). P. 195–214.

Campbell N. Tools and Resources for Visualising Conversational-Speech Interaction / M. Kipp et al. (eds.) // Multimodal Corpora: From Models of Natural Interaction to Systems and Applications. Springer. Heidelberg, 2009.

Carletta J. Announcing the AMI Meeting Corpus // The ELRA Newsletter, 11(1), January-March, 2006. P. 3–5.

Chafe W. (ed.) The pear stories: Cognitive, cultural, and linguistic aspects of narrative production. Norwood, Ablex, 1980.

Foster M.E., *Oberlander J.* Corpus-based generation of head and eyebrow motion for an embodied conversational agent // Language Resources and Evaluation. 41 (3/4). 2007. P. 305–323.

Holler J., *Kendrick K.H.* Unaddressed participants' gaze in multi-person interaction: Optimizing recipiency // Frontiers in Psychology. 2015. 6. P. 98.

Kendon A. Gesture. Visible action as utterance. Cambridge, 2004.

Knight D. Multimodality and active listenership: A corpus approach. London: Bloomsbury, 2011.

McNeill D. Gesture and thought. Chicago, 2005.

Mondada L. Bodies in action // Language and Dialogue. 4 (3). 2014. P. 357–403.

Müller C., *Fricke E.*, *Cienki A.*, *McNeill D.* (eds.) Body – Language – Communication. Mouton de Gruyter. Berlin, 2014.

Scollon R. Multimodality and the language of politics / K. Brown (ed.) Encyclopedia of language and linguistics. Elsevier, 2006. Vol. 9. P. 386–387.

Taylor M. The Structure of Multimodal Dialogue. Amsterdam: Elsevier, 1989.

Železný M., Krňoul Z., Cisař P., Matoušek J. Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis // Signal Processing. 2006. 83 (12). P. 3657–3673.

Примечания

[1] В психологии и нейрофизиологии модальность определяется как принадлежность сигнала к определенной сенсорной системе человека.

[2] В настоящее время больше распространен термин «мультимодальный», однако корректнее говорить именно о «мультиканальной», или «бимодальной», коммуникации, так как в данной области пока изучаются только две модальности – вокальная (слуховая) и кинетическая (зрительная), а остальные модальности, например, обоняние или осязание, остаются за пределами рассмотрения.

[3] Если более подробно расшифровать в этом определении компонент «речь», можно выделить два типа мультимодальных корпусов: мультимодальные корпуса в строгом смысле слова, для которых обязательным является аннотирование всех коммуникативных каналов, включая просодический (т. е. понимание под «речью» не только вербальной составляющей, но и просодической), и мультимодальные корпуса в нестрогом смысле слова, для которых просодическое аннотирование необязательно. Все мультимодальные корпуса, описанные в настоящем разделе, являются корпусами в нестрогом смысле слова; более того, большинство подобных корпусов ограничиваются аннотацией мануальной составляющей жестикуляции, т. е. жестами рук (Kendon 2004; McNeill 2005).