Светлана Олеговна Савчук,
Ольга Николаевна Ляшевская
(Институт русского языка им. В. В. Виноградова РАН,
Высшая школа экономики)

УСТНАЯ РАЗГОВОРНАЯ РЕЧЬ И СПОСОБЫ ЕЕ ПРЕДСТАВЛЕНИЯ В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА[1]

Под устной разговорной речью понимается бытовая повседневная речь, которая отражает непубличную сферу общения. В терминологии НКРЯ тексты, относящиеся к этой сфере, кодируются как «устная непубличная речь» и входят в состав модуля устной речи.

При разработке проекта устного модуля десять лет назад мы учитывали опыт представления устной речи в больших корпусах (в 2000-х годах это были прежде всего Британский, Чешский, Словацкий, Польский, Американский). Следуя определенным стандартам представления устных текстов, в каждом национальном корпусе предлагаются особенные решения в отборе и организации материала. В НКРЯ устройство устного компонента основано на следующих принципах [1, 2].

1) Устный подкорпус выделен из основного корпуса письменных текстов в самостоятельный модуль, поскольку в нем используются такие виды разметки, которые значимы только для устной речи.

2) Мы отказались от идеи создания универсального корпуса устной речи, с помощью которого можно изучать все ее лингвистические аспекты — от фонетики до анализа дискурса (эту роль выполняет ряд специализированных корпусов звучащей речи, таких как [3, 4]

[1] Работа выполнена при поддержке РФФИ (грант № 15-06-04334).

и др.). Поэтому в составе устного модуля НКРЯ не один, а три корпуса, при этом каждый нацелен на решение специфических задач: а) устный корпус; б) акцентологический корпус; в) мультимедийный корпус.

Ниже приведена краткая характеристика функционирующих корпусов по следующим параметрам: объем, характер материала и форма его представления, виды лингвистической разметки, состав текстов, назначение корпуса, способы оптимального использования. Отдельно описана устная непубличная речь в каждом из этих корпусов — состав и источники текстов, качество материала, проблемы его сбора и обработки.

1. Устный подкорпус в настоящее время составляет более 11,3 млн словоупотреблений. Материал представлен в виде транскриптов, соответствующий звучащий текст недоступен. Ценность этого корпуса для исследователей устной речи заключается прежде всего в его большом объеме, большом временном диапазоне, отраженном в записях, функциональном разнообразии текстов. Доля непубличной речи составляет более 1,3 млн словоупотреблений, или 11,5 %. В корпусе используется стандартная для НКРЯ метатекстовая, морфологическая, семантическая разметка, а также специфическая для устного корпуса социологическая аннотация. Каждой реплике приписаны сведения о говорящем (если они известны): пол, возраст или год рождения, род занятий. Корпус сбалансирован скорее по составу текстов, чем по составу говорящих, однако благодаря наличию метатекстовой и социологической разметки пользователь может отобрать для изучения свой подкорпус более или менее однородных текстов по интересующим его признакам. Например, подкорпус устных научных текстов, спортивных комментариев, подкорпус реплик только мужских или женских, представителей разных возрастных групп и пр.

2. Акцентологический корпус дает сведения о таком аспекте устной речи, как словесное ударение. Его объем составляет около 11 млн словоупотреблений. Корпус состоит из двух зон. 1) Поэтическая зона содержит тексты, в которых размечены слоги, на которые может падать ударение, путем пересчета по специальным правилам мы в большинстве случаев можем получить точные сведения о месте ударения в том или ином слове. 2) Прозаическая зона составляет 6 млн словоупотреблений и включает транскрипты записей устной речи, в которых расставлены ударения в соответствии с реальным произношением. В корпусе богато представлены разные сферы

бытования устного слова: устная научная речь, устная официальная речь, теле- и радиопублицистика, радио- и телереклама, церковная проповедь, речь кино и театра, устное художественное слово. Наряду с текстами, рассчитанными на публичную аудиторию, есть записи, отражающие общение в небольших группах (тренинги, семинары, экскурсии, заседания, совещания, беседы и под.), а также личное, непубличное общение. Доля непубличной речи сравнительно невелика и составляет около 250 тыс. словоупотреблений, или 4,2 %. Все транскрипты выверены по звуковому оригиналу, при этом записи могут быть невысокого качества. Аудиофайлы находятся в архиве и могут быть доступны для проверки.

3. Мультимедийный корпус (МУРКО) дает наиболее полное представление об устной коммуникации и позволяет изучать фонетическую сторону речи, поскольку для онлайн-поиска доступен и транскрипт, и сопровождающий его аудио- или видеоклип. Объем корпуса приближается к 4,5 млн словоупотреблений. Записи непубличной речи составляют пока незначительную часть — около 12 тыс. словоупотреблений (около 0,3 %), поскольку подготовка требует значительных материальных и трудозатрат. Кроме того, приходится преодолевать трудности юридического и психологического характера, связанные с публикацией в открытом ресурсе, каким является корпус, материалов личного характера. Однако корпус находится в процессе становления, и в планах его развития предусмотрено увеличение доли непубличной коммуникации.

Обзор текущего состояния устного модуля НКРЯ с точки зрения отражения в нем естественной устной разговорной речи приводит нас к выводу о том, что надо увеличивать зону непубличной коммуникации в НКРЯ, особенно в мультимедийном представлении. Кроме того, предстоит решить ряд насущных задач: исследовать состав говорящих и отобрать сбалансированный по говорящим подкорпус, провести корректировку метаразметки транскриптов, в частности унифицировать комментарии к неречевым событиям и пр.

4. Развитие устного модуля. Основной способ пополнения подкорпуса устной разговорной речи включает сбор материалов и их обработку по технологии МУРКО. Этот способ трудоемок, но он обеспечивает высокую степень соответствия транскрипта звуковому оригиналу, точность передачи акцентуации, некоторых особенностей произношения и пр. Таким способом было собрано и обработано несколько коллекций.

1) Нижегородская коллекция подготовлена студентами Нижегородского филиала ВШЭ под руководством В. Сибирцевой, М. Фокиной и В. Паниной. Она включает небольшие записи (3–5 мин.) разнообразного тематического содержания и жанрового состава. Записи выполнялись среди родственников и друзей с их согласия, тем самым снимались вопросы юридического характера. Примерный объем коллекции — более 30 тыс. словоупотреблений.

2) Московская коллекция собрана студентами московских вузов и отражает в основном общение в молодежной аудитории. В ее состав входят записи телефонных разговоров, беседы с друзьями и родственниками, впечатления о поездках, рассказы о ярких случаях из жизни, пересказы интересных сюжетов и под. Объем материалов превышает 20 тыс. словоупотреблений.

3) Региональные коллекции — одно из приоритетных направлений, которое мы намерены развивать в рамках Мультимедийного корпуса. Эти коллекции создаются при активном участии коллег из региональных исследовательских центров (Дальневосточный федеральный университет, Казахстанский филиал МГУ им. Ломоносова, Саратовский университет и др.). Материалы будут размещены на сайте по мере достижения некоторой полноты и сбалансированности состава.

5. Большие (но несовершенные) данные. Подготовка материалов традиционным способом, отличающегося трудоемкостью, обеспечивает медленный прирост объема корпуса. Новый подход, который позволит многократно увеличить объем данных, доступных для исследователей, состоит в том, чтобы организовать подкорпус аудио- и видеоматериалов, собранных нелингвистами — историками, социологами, политологами, журналистами (см. проекты «Прожито», «Та сторона», «Устная история» и др.). Эти материалы снабжены «несовершенными» расшифровками, которые не соответствуют стандартам подготовки транскриптов для устных корпусов (например, в них отсутствует разметка хезитаций и самоперебивов). Вместе с тем такие расшифровки способны стать ключом к обширному устному материалу, на котором можно решать ряд важных лингвистических задач, например, изучать синтаксис устной речи.

Другой источник «несовершенных» расшифровок — результат работы систем автоматического распознавания речи [5]. Целесообразность создания такого подкорпуса, способы представления данных, их лингвистическая оценка, область их использования требуют предварительного тщательного анализа и обсуждения.

Список использованной литературы

1. *Гришина Е. А.* Устная речь в Национальном корпусе русского языка // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М.: Индрик, 2005. С. 94–110.

2. *Гришина Е. А., Савчук С. О.* Корпус устных текстов в Национальном корпусе русского языка: состав и структура // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб., 2009. С. 129–149.

3. Звуковой корпус как материал для анализа русской речи. Коллективная монография / отв. ред. Н. В. Богданова-Бегларян СПб., 2013.

4. Рассказы о сновидениях и другие корпуса звучащей речи / Кибрик А. А., Подлесская В. И., Коротаев Н. А. и др. Электронный ресурс. *http://spokencorpora.ru.*

5. *Ляшевская О. Н.* Параллельный корпус автоматических и ручных расшифровок устной русской речи // Труды международной научной конференции «Корпусная лингвистика — 2015». СПб., 2015. С. 315–324.