

М. П. КОНЦЕВОЙ

Брест, УО «БрГУ имени А. С. Пушкина»

КОРПУСНЫЕ ТЕХНОЛОГИИ ОНТОЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ

Онтолингвистика (лингвистика детской речи) является научным фундаментом системы языкового образования в современной начальной школе. Онтолингвистические исследования направлены на особенности детской языковой системы как таковой и сам процесс овладения языком в детском возрасте [1]. Конструктивистская парадигма онтолингвистических исследований предполагает, что ребенок сам конструирует свой язык, опираясь на речевую продукцию взрослых, которую подвергает бессознательному анализу. Постепенно он овладевает языком, каждодневно совершая очень сложную работу, состоящую в категоризации языковых фактов, установлении разного рода аналогий и ассоциаций. Такой подход к научному осмыслению развития речевых способностей ребенка выдвигает на первый план корпусные технологии как в самих онтолингвистических исследованиях, так и в педагогической практике языкового образования детей.

Лингвистический корпус – репрезентативная (представительная – соответствующая той области функционирования языка, которую будет отражать) совокупность речевых данных (письменных, устных, мультимодальных текстов), собранных в соответствии с определенными принципами (соответствующими конкретной исследовательской задаче), размеченных (снабженных аннотациями) по определенному стандарту и обеспеченных специализированной поисковой системой. Корпус звучащей речи (речевая база данных) – структурированная совокупность речевых фрагментов, которая обеспечена программными средствами поиска и доступа к самим фрагментам и метаданным об этих фрагментах. Речевой фрагмент – оцифрованный фрагмент речевого сигнала с дополнительной ассоциированной метаинформацией. Использование представительных речевых корпусов открывает недоступную ранее возможность для проведения крупномасштабных и статистически достоверных исследований детской речи на аутентичном и репрезентативном материале. Лингвистика детской речи первоначально фундировалась лонгитюдными наблюдениями (case-study – одна и та же группа детей наблюдается на протяжении длительного времени). Фундаментальную роль имели дневниковые записи Н. Гвоздева, в которых он фиксировал речь своего сына, сопровождая их лингвистическими комментариями. Важную роль в исследовании детской речи и развитии корпусных методов сыграла книга В. К. Харченко с

публикацией корпуса детских высказываний в хронологической последовательности прозвучавших реплик на протяжении семи лет наблюдения [2].

Современные информационно-технические системы и средства позволили существенно повысить уровень репрезентативности таких данных на основе расшифровки стенограмм в компьютерные файлы, которые можно было легко скопировать, отредактировать и обработать с помощью стандартных методов обработки данных. В 1981 г. был задуман проект CHILDES (Children Language Database Exchange System, <http://childes.psy.cmu.edu>) – международная система обмена данными по детской речи, разработанная американскими учеными К. Сноу и Б. Мак-Винни в Питтсбургском университете и применяемая для анализа разговорной речи, спонтанной речи детей, а также для исследования усвоения второго языка. На сегодняшний день база данных CHILDES представляет собой лингвистический корпус, включающий языки различного типологического строя, объемную библиографию по психолингвистике, лингвистике, теории усвоения первого и второго языков, а также правила ввода материала и пакет программ для его анализа. Каждый исследователь может воспользоваться необходимыми ему данными, а также разместить в CHILDES свои материалы. Данный корпус позволяет не только исследовать детскую речь, но и составить представление об информанте, установить автора записи, реализуя несомненные преимущества корпусных технологий в исследовании детской речи: доступность, полиаспектность, единство формата записей, разнообразие материалов и их достоверность [3].

Создание качественного репрезентативного речевого корпуса представляет собой сложную технологическую задачу, требующую значительных финансовых и кадровых вложений. Анализировать устную речь, когда она существует только в форме звука, практически невозможно – звук слишком многоаспектен, чтобы человеческий взгляд и сознание могли одновременно удержать какой-то его значимый фрагмент [4]. Для объективной фиксации устной речи используется переложение ее в графический вид, т. е. создание транскрипции. Таким образом, база данных корпуса устной речи должна состоять по меньшей мере из двух компонентов – аудиозаписей речи и соответствующих этим записям транскриптов. Представление устной речи в письменном виде является сложной научной задачей. Разрабатывая транскрипцию для конкретной задачи, исследователь может выбирать, какие фонетические и просодические признаки в нее включать, а какие нет. При транскрибировании устной речи следует основываться на верности реальной аудиозаписи, как бы она ни противоречила нашим априорным представлениям о том, «как надо говорить по-русски», какой должен быть

порядок слов, какие допустимы синтаксические конструкции и т. д. Необходимо избегать искушения подгонки под привычные шаблоны письменного языка. То, чего не может быть в письменном языке (или мы думаем, что его там не может быть), вполне может встречаться в устном. В особенности это касается детской речи [5].

Для стандартизации сбора, хранения, распространения корпусных баз (в том числе речевых) созданы специальные координационные центры:

- Linguistic Data Consorciium (<http://www ldc.upenn.edu>);
- Center for Spoken Language Understanding (<http://www CSLU.ogi.edu>);
- European Language Resources Association (<http://www.elra.info>).

Для расшифровки полученных аудиозаписей детской речи разработан стандартный формат CHAT CHILDES. Морфологическое аннотирование транскрипций записей может быть реализовано с помощью специализированного программного обеспечения, например MORCOMM и CHILDES CLAN [5]. Вместе с тем актуальные речевые корпуса имеют свои ограничения, прежде всего в плане речевой мультимодальности. Например, в CHILDES некоторые записи содержат аудио и видеофайлы, однако они никак не аннотированы и не проанализированы с точки зрения невербальных компонентов общения.

Первый представительный речевой корпус для русского языка с разметкой речевых фрагментов на звуковые единицы ISABASE создан еще в конце 90-х гг. в Институте системного анализа РАН при участии специалистов речевой группы филологического факультета МГУ, а представительный речевой корпус RuSpeech разработан в 2000–2001 гг. в ИСА РАН для разработки систем распознавания русской речи. Помимо самих речевых баз, важным результатом данных проектов явились отлаженная технология создания речевых корпусов и комплекс программных средств для обеспечения этой технологии, отладка автоматического транскриптора русской речи, создание программы для подготовки текстового материала с нужными фонетическими и статистическими характеристиками и др.

Активно используются в исследованиях такие современные корпусные проекты для русской звучащей речи, как МУРКО (<http://ruscorpog a.ru/search-murco.html>), Русскоязычный эмоциональный корпус (<http://www.harpia.ru/rec/>), «Один речевой день» (<http://model.org.spbu.ru/>), «Рассказы о сновидениях» (<http://spokencorpora.ru/>).

Разрабатываются и применяются в онтолингвистических исследованиях и корпусные проекты русской детской звучащей речи. CHILDES Project (https://www.hse.ru/neuroling/cla_project_childes/) направлен на изучение процесса усвоения уровней языка детьми от 1 года до 3 лет (на материале видеозаписей общения русскоговорящих детей в

семейном кругу). Семьи, которые участвуют в проекте, раз в две недели записывают на видео обычное взаимодействие ребенка со взрослым. CHILDES Project имеет стандартизированный процесс аннотации материалов для их последующей интеграции в корпус детской речи The Child Language Data Exchange System (CHILDES) и адаптированные правила транскрибирования и аннотирования русскоязычных материалов для этого корпуса (опираясь на англоязычное транскрибирование в CLAN).

Корпус INFANT.RU содержит вокализации и речь 187 детей от 0 до 3 лет жизни, корпус CHILD.RU – образцы спонтанной и читаемой речи детей 4–7 лет, а корпусная база данных Emo.Child.Ru – записи спонтанной эмоциональной речи детей 4–7 лет [6]. Собранный в данных корпусах речевой материал уже используется при проведении междисциплинарных исследований по изучению различных аспектов становления речи и их связи с когнитивным и эмоциональным развитием ребенка. Однако данные корпуса не являются общедоступными, что обусловлено необходимостью защиты персональных данных и вытекающими отсюда правовыми и этическими ограничениями в публичном использовании фрагментов детской речи: при проведении исследований с участием детей ни аудио, ни видеофайлы, содержащие запись голоса или представляющие лицо ребенка, не могут быть представлены в сети Интернет в открытом доступе.

Особое значение имеет «Конduit» (Корпус Неподготовленных Детских Устных Извлеченных Текстов, <http://konduitcorpus.ru/accounts/login/?next=/>), содержащий сотни устных текстов русскоязычных детей в возрасте от 2 лет 7 месяцев до 6 лет 7 месяцев, представленных в виде аннотированных орфографических записей, а также в виде аннотированных аудио и видеофайлов полученных рассказов. Корпус «Конduit» на данный момент является единственным русскоязычным корпусом устных детских текстов, содержащим мультимодальные данные в открытом доступе. В «Конduit» на основе субтитрования демонстрировавшегося детям фрагмента мультфильма все полученные тексты были покadroво соотнесены с описываемой ребенком в данный момент времени ситуацией, а при помощи программы аудио- и видеообработки ELAN проведена аннотация речевого сигнала, его просодики и жестикюляции. Полученные форматы (исходный мультфильм с наложенным в виде субтитров рассказом ребенка и сам рассказ в аннотированном виде) синхронизированы, что позволяет одновременно наблюдать как процесс восприятия (как быстро ребенок реагирует на различные действия героев, какие из действий персонажей оказываются достаточно важными, чтобы найти свое отражение в устном рассказе и т. д.), так и процесс порождения (какие вербальные и невербальные средства использует для описания данного действия ребенок) [7].

Развитие корпусных технологий и реализация конкретный проектов в системной совокупности с решением правовых и этических вопросов использования баз данных детской речи позволит на качественно более высоком уровне научной репрезентативности осмыслить процесс формирования речевой личности, что, в свою очередь, откроет новые возможности для языкового развития, обучения и воспитания.

Список использованной литературы

1. Цейтлин, С. Н. Язык и ребенок: Лингвистика детской речи : учеб. пособие / С. Н. Цейтлин. – М. : ВЛАДОС, 2000. – 240 с.
2. Харченко, В. К. Корпус детских высказываний / В. К. Харченко. – М. : Лит. ин-т им. А. М. Горького, 2012. – 520 с.
3. Зырянова, Е. В. Система CHILDES как метод сбора материалов и изучения детской речи [Электронный ресурс] / Е. В. Зырянова. – Режим доступа: https://lib.herzen.spb.ru/text/zyryanova_35_76_2_113_118.pdf. – Дата доступа: 29.10.2019.
4. Кривнова О. Ф. Речевые корпуса (опыт разработки и использование) [Электронный ресурс] / О. Ф. Кривнова, Л. М. Захаров, Г. С. Строкин // Диалог: Компьютерная лингвистика и интеллектуальные технологии : сб. ст. – Режим доступа: <http://www.dialog-21.ru/digest/2001/articles/krivnova/>. – Дата доступа: 29.10.2019.
5. Кибрик А. А. К созданию корпусов устной русской речи: принципы транскрибирования [Электронный ресурс] / А. А. Кибрик, В. И. Подлесская. – Режим доступа: https://iling-ran.ru/kibrik/Corpora_speech_transcription@S&I_2003.pdf. – Дата доступа: 29.10.2019.
6. Риехакайнен, Е. И. Методика создания корпуса для изучения редуцированных реализаций в детской речи / Е. И. Риехакайнен // Корпусная лингвистика–2019 : тр. междунар. конф., 24–28 июня 2019 г., Санкт-Петербург. – СПб. : С.-Пб. ун-т, 2019. – С. 349–355.
7. Эйсмонт, П. М. Мультиmodalность в корпусе устных детских текстов «КОНДУИТ» / П. М. Эйсмонт // Корпусная лингвистика–2019: тр. междунар. конф., 24–28 июня 2019 г., Санкт-Петербург. – СПб. : С.-Пб. ун-т, 2019. – С. 373–379.