Е. С. КРАСНОПЕРОВА
г. Екатеринбург, Россия
evgenia113@gmail.com

УДК 81'23
DOI 10.26170/2306-7462_2021_02_13

КОРПУСНЫЕ ИССЛЕДОВАНИЯ ДЕТСКОЙ РЕЧИ: ПЕРСПЕКТИВЫ РАЗВИТИЯ

Аннотация. В статье рассматриваются различные методы изучения детской речи, используемые в современной онтолингвистике. Исследуется возможность применения корпусных методов в области детской речи. Поднимается вопрос о необходимых свойствах корпусов, отличающих их от коллекций текстов, дневниковых записей и словарей. Представлен обзор корпусов детской речи на русском языке. Особое внимание уделяется трудностям составления корпусов детской речи. Описана возможность составления электронного корпуса детской речи с грамматической разметкой на материале издания «Корпус детских высказываний» В. К. Харченко. Представлена авторская разработка такого корпуса с использованием программы автоматизированной морфологической разметки MyStem. Практическая значимость исследования видится в создании инструментария для исследования языковой способности ребенка.

Ключевые слова: онтолингвистика; корпусные исследования; детская речь; методы исследования; развитие речи; языковая способность.

KRASNOPEROVA EVGENIYA S.
Ekaterinburg, Russia

CORPUS STUDIES OF CHILDREN'S SPEECH: DEVELOPMENT PROSPECTS

Abstract. The article examines various methods of studying children's speech used in modern ontolinguistics. The possibility of using corpus methods in the field of children's speech is being investigated. The question is raised about the necessary properties of corpora that distinguish them from collections of texts, diary entries and dictionaries. An overview of the corpora of children's speech in Russian is presented. Particular attention is paid to the difficulties of drawing up corpuses of children's speech. The

article describes the possibility of compiling an electronic corpus of children's speech with grammatical markup on the material of the publication "Corpus of children's utterances" by V. K. Kharchenko. The author's development of such a corpus is presented using the MyStem automated morphological marking program. The practical significance of the research is seen in the creation of tools for the study of the child's language ability.

Keywords: ontolinguistics; corpus research; children's speech; research methods; development of speech; language ability.

Изучение детской речи предполагает использование ряда методов: дневниковые записи, видео и аудиофиксация речевого поведения, психолингвистические эксперименты. Образцовое использование метода дневниковых записей представлено в работе А. Н. Гвоздева «От первых слов до первого класса. Дневник научных наблюдений» [Гвоздев 2005], представляющей систематическую запись фактов детской речи. На основе этих наблюдений А. Н. Гвоздев создал первое систематическое описание процесса усвоения языка ребенком [Гвоздев 2007]. Методы дневниковых записей и различных способов фиксации речевого поведения ребенка остаются основными и продолжают активно использоваться в современной онтолингвистике. В частности, в трудах санкт-петербургской научной школы онтолингвистики представлен большой материал для исследований детской речи [Речь русского ребенка. Звучащая хрестоматия 1994; От нуля до двух 1997; От двух до трех 1998; Две девочки: Соня и Надя: 2001]. На основе этого материала создаются различные словари детской речи [Словарь детских словообразовательных инноваций 2001], появляются работы, исследующие процесс освоения языка ребенком, его возрастную динамику, индивидуальные особенности речевого развития детей и т.д.

Использование психолингвистических экспериментов в изучении перцептивной и продуктивной речи ребенка привело в том числе к созданию ряда уникальных словарей, представляющих различные аспекты языковой способности. Например, «Лексикон младшего школьника (характеристика лексического компонента языковой компетенции)» И. Г. Овчинниковой, Н. И. Бересневой, Л. А. Дубровской, Е.Б. Пенягиной, «Возрастная психолингвистика: Толковый словарь русского языка гла-

зами детей» А. Д. Палкина, «Мотивационный словарь детской речи» К. В. Гарганеевой, «Объяснительный словарь детских инноваций» Т. А. Гридиной [Овчинникова, Береснева 2000; Гуц 2004; Палкин 2004; Гарганеева 2007; Гридина 2012].

Относительно новым, вошедшим в арсенал лингвистической науки с развитием информационных технологий и повсеместным внедрением математических и статистических методов, стал метод корпусных исследований, позволяющий получать наборы статистически значимых показателей, устанавливать семантические, грамматические, стилистические и другие закономерности на большом массиве данных. В онтолингвистике использование корпусных методов позволяет получить статистически достоверные знания о процессе освоения ребенком языка и в перспективе расширить набор словарей детской речи на основе материалов корпусов.

На сегодняшний день созданы корпусы детской речи на английском (американский и британский варианты), немецком, испанском, французском, шведском, датском, китайском языках, существует международная база данных CHILDES (Child Language Data Exchange System, https://childes.talkbank.org), где собраны записи детской речи для более чем 40 языков. Исследователи могут пользоваться размещенными материалами и пополнять базу данных. К сожалению, банк данных русскоязычных детей невелик и пополняется медленно из-за необходимости предварительной обработки материала по правилам этого ресурса.

Кроме указанных источников на русском языке существуют корпус вокализаций и речи детей первых трех лет жизни INFANT.RU, спонтанной и читаемой речи детей 4-7 лет жизни CHILD.RU, эмоциональной речи детей 4-7 лет Emo.Child.Ru, созданные группой по изучению детской речи СПбГУ [Ляксо, Фролова 2017]. Эти аудиокорпусы, как и корпус детской письменной речи StartWrit [Ахапкина, Сосновцева 2017], не находятся в открытом доступе.

Отсутствие в открытом доступе готовых корпусов детской речи усложняет внедрение методов корпусной лингвистики в область изучения детской речи. Нами предпринята попытка создания ана-

лога такого корпуса для исследования, в первую очередь, грамматических закономерностей процесса освоения языка.

Этапы создания корпуса

1. Отбор материала.

Необходимость представления в корпусе таких текстов, которые наиболее полно демонстрируют лингвистическое явление, делают задачу создания корпуса детской речи достаточно сложной. Детская речь отражает процесс становления языковой способности ребенка, процесс овладения языком. Этот процесс динамичен, индивидуален и изменчив, проходит в разных коммуникативных условиях, поэтому набор текстов, репрезентирующий детскую речь, во-первых, должен охватывать определенную часть жизни ребенка, во-вторых, фиксация высказываний должна производиться регулярно, в-третьих, условия коммуникации, в которых высказывания порождаются, не могут быть однотипными. Большинство дневниковых записей детской речи охватывает определенный период в жизни ребенка (от нескольких месяцев до двух-трех лет), не являются сплошными (не охватывают все коммуникативные ситуации) и не ведутся регулярно.

2. Определение необходимого объема текстов.

Объем текста должен быть достаточен для репрезентации явления. Такой объем записей речи детей в большинстве случаев просто отсутствует.

3. Создание корпуса в электронном виде.

Электронный формат корпуса – это необходимое условие для работы с большими массивами информации, для автоматизации процессов обработки материала и для простоты и наглядности представления результатов работы. Электронный характер корпуса предполагает необходимость перевода имеющихся записей детской речи в текстовый формат.

4. Создание системы разметок.

Разметка – это обязательная составляющая корпуса, которая позволяет осуществлять навигацию по тексту, проводить анализ лингвистических явлений, зафиксированных в записях детской речи, обнаруживать количественные и качественные закономерности. Создание разметки является важнейшим этапом корпус-

ного исследования. Наличие разметки отличает корпус от дневниковых записей и коллекций текстов.

5. Постобработка.

Создание корпуса детских высказываний требует особого внимания к этапу постпроверки и постобработки. Обработка текстов для корпуса, создание разметки, грамматической, в первую очередь, производится с помощью специализированных программных средств. Автоматизированная обработка дает хороший результат, если текст приближен к языковой норме, но точность работы программы снижается при обработке аномальных текстов, таких как детские высказывания с большим количеством грамматических, фонетических, семантических инноваций.

В качестве источника материала для проекта корпуса детской речи нами был выбран «Корпус детских высказываний», созданный В. К. Харченко [Харченко 2012]. В этой работе представлены высказывания двух детей, фиксировавшиеся на протяжении семи лет наблюдения. Записи ввелись регулярно, фиксировались детские высказывания трех типов: спонтанная монологическая речь, диалог между ребенком и взрослым, диалог между детьми. Жанровое обозначение «корпус», выбранное автором, соответствует объему и характеру зафиксированного материала [Харченко 2012: 9]. В данном корпусе используется различная разметка: дата записи (*27 апреля 2008 г.*); комментарий составителя к ситуации: *Я леплю и вижу, что что-то не похоже. (Лепит ихтиозавра)*; квалификация лингвистического явления: *Аля мама! (инверсия)*; интерпретация высказывания: *Я учусь по пятёркам (на пятёрки)*; ненормативное ударение: *ЖУка дай! Баб, а откуда ты знаешь, что коричневая – это гнИлая трава?* [Харченко 2012: 2019-220]. Однако указанная авторская разметка является выборочной и смешанной (лингвистической и экстралингвистической). С точки зрения корпусных исследований этот источник нельзя рассматривать как корпус, скорее, как материал для его создания.

Представим основные шаги по созданию корпуса детской речи с грамматической разметкой и информацию, которую можно извлечь из него на первом этапе нашего исследования.

1. Перевод материала в электронный вид, выделение детских высказываний интересующего нас ребенка из всего зафиксированного материала. Для целей нашего исследования были отобраны высказывания детей в возрасте от двух до пяти лет.

2. Обработка материала.

Обработка осуществляется с помощью программы AntConc, мультиплатформенного инструмента для проведения лингвистических исследований. Данная программа позволяет находить все контексты интересующего нас слова, искать кластеры и N-граммы по заданным условиям, подсчитывать все слова в корпусе и представлять их в виде упорядоченного списка, отмечать необычно часто (или редко) встречающиеся в корпусе слова.

Определение частотности слов, как элемент количественного анализа, снабжает исследователя информацией разного типа. Частотный список слов может стать основой для создания частотного словаря детской речи или определения лексического ядра словарного запаса конкретного ребенка. Сопоставление частот слов позволяет отметить преобладание местоимения «я» в речи ребенка (на первой позиции по частоте в возрасте от двух до трех лет и от трех до четырех лет). Среди существительных и глаголов в речи ребенка от двух до трех лет самыми частотными оказываются термины родства, имена членов семьи и названия любимых игрушек, глаголы волеизъявления: *баб/баба, машина, мама, Вера, дядя, Лёвка, хочу, давай/дай, буду/не буду.* Расширение материала корпуса и ранжирование слов по частоте позволит сделать и менее очевидные наблюдения.

Возможность выделять и выстраивать в отдельный список все случаи употребления конкретного слова позволяет отследить становление и изменение лексического значения или грамматической формы слова, его устойчивую сочетаемость и типовые контексты употребления. Ниже представлены контексты употребления слова *машина* ребенком в возрасте от двух до трех лет:

– *Непонятая какая-то машина!*
– *Да это ещё машина с маминого детства лежит!*
– *Эта машина гудит. А пока она не гудит.*

– *Где машина? Во-от! А я не успел догадаться, где, машина!*
– *Эта машина приехала от пожара, с пожара!*
– *Машина меня тронула. Машина хорошая?*
– *Машина нагревается. Всё! Она нагрелась.*
– *Нашла самолёта? Машина хочет поднять самолёта.*
– *Ехала машина, все спали, только шофёр не спал.*
– *Всё! Машина сухая.*

Можно сделать предварительный вывод об усвоении ребенком грамматической категории рода применительно к этому слову. Можно посмотреть на ситуативные контексты употребления слова *машина* и выделить те, что связаны с ситуацией игры и игрушкой, и те, что связаны с транспортным средством. Таким образом, характер информации, которую можно получить с помощью этого инструмента, весьма разнообразен.

3. Выполнение автоматизированной грамматической разметки.

Разметка выполнялась программой MyStem (https://yandex.ru/dev/mystem/), которая проводит морфологический разбор всех единиц корпуса.

Пример разбора формы существительного *машина*: *<w>машина<analex="машина" gr="S,жен,неод=им,ед" /></ w>; <w>буду<analex="быть" gr="V,нп=непрош,ед,изъяв,1-л" /></w>*.

Грамматическая разметка позволяет выявлять грамматические закономерности развития детской речи, отслеживать процесс формирования отдельных грамматических категорий, выявлять функциональные грамматические системы на каждом из разных возрастных периодов.

4. Постобработка.

Необходимость проверять результаты работы программы автоматизированной разметки связана с возможностью квалификации одной и той же словоформы по разным грамматическим признакам. Например, словоформа *большие* может быть квалифицирована как форма именительного или винительного падежа множественного числа прилагательного «большой» (<w>большие<analex="большой" gr="A= (вин,мн,полн,неод| им,мн,полн)" /></w>).

Кроме того, на материале детской речи программы морфологического анализа работают менее точно из-за большого количества детских грамматических инноваций: Хочу всё *разбрОсить*. Всё *разбрОсил*! Она *вЫгружила*, она *вЫгружила*... Нет, я *растЮ*! Такие инновации могут квалифицироваться верно, могут неверно и нуждаются в постпроверке вручную.

Изучение детских инноваций является важной частью онтолингвистических исследований, описывающих не только лингвистические механизмы усвоения языка, но и доминанты детского языкового сознания (см., «Объяснительный словарь детских инноваций», «Своя игра»: ребенок в мире языка» Т. А. Гридиной [Гридина 2012, 2015]). Отбор и определение программой некоторых грамматических инноваций как «неправильных», «неквалифицируемых» дает дополнительный материал для таких исследований.

Таким образом, как мы попытались показать в данной статье, корпус детской речи может стать важным инструментом исследования языковой способности ребенка, источником разнообразной информации о процессе овладения языком.

Литература

Ахапкина Я. Э., Сосновцева Е. Г. Корпус детской письменной речи: StartWrit // Проблемы онтолингвистики – 2017. Освоение и функционирование языка в ситуации многоязычия : Материалы ежегодной международной научной конференции, Санкт-Петербург, 26–28 июня 2017 года, СПб.: ЛИСТОС, 2017. URL: https://www.elibrary.ru/item.asp?id=30048197 (дата обращения: 09.04.2021).

Гарганеева К. В. Мотивационный словарь детской речи. Томск: Изд-во Томского ун-та, 2007.

Гвоздев А. Н. Вопросы изучения детской речи. СПб.: Детство-Пресс, 2007.

Гвоздев А. Н. От первых слов до первого класса. Дневник научных наблюдений. М.: КомКнига, 2005.

Гридина Т. А. Объяснительный словарь детских инноваций: монография и словарь / Урал. гос. пед. ун-т. Екатеринбург, 2012.

Гридина Т. А. «Своя игра»: ребенок в мире языка: Монография. Екатеринбург: ФГБОУ ВПО «Урал. гос. пед. ун-т», 2015.

Гуц Е. Н. Ассоциативный словарь подростка. Омск: Ом. гос. ун-т, 2004.

Две девочки: Соня и Надя: дневниковые записи / сост.: О. А. Юнтунен, О. Б. Сизова; отв. ред. С. Н. Цейтлин. СПб.: Тускарора, 2001.

Лаксо Е. Е., Фролова О. В., Григорьев А. С., Остроухов А. В. Корпуса детской речи "infant.Ru", "infant.MAVS", "child.Ru", "EmoChildRu" на материале русского языка и их использование в исследованиях речевого онтогенеза // Теоретическая и прикладная лингвистика. 2017. Т. 3. № 1. С. 28-58. URL: https://www.elibrary.ru /item.asp?id=29161390 (дата обращения: 09.04.2021).

От двух до трех: дневниковые записи / сост.: С. Н. Цейтлин, М. Б. Елисеева. СПб.: Изд-во РГПУ им. А. И. Герцена, 1998.

От нуля до двух: дневниковые записи / сост. С.Н. Цейтлин. СПб.: Изд-во РГПУ им. А.И. Герцена, 1997.

Овчинникова И. Г., Береснева Н. И., Дубровская Л. А., Пенягина Е. Б. Лексикон младшего школьника (характеристика лексического компонента языковой компетенции). Пермь, 2000.

Палкин А. Д. Возрастная психолингвистика: Толковый словарь русского языка глазами детей. М., 2004.

Речь русского ребенка: Звучащая хрестоматия / сост.: Т. В. Кузьмина, Э. И. Столярова, С. Н. Цейтлин. СПб.: Бохум, 1994.

Харченко В. К. Корпус детских высказываний. М.: Изд-во Литературного института им. А. М. Горького, 2012.

Цейтлин С. Н. Словарь детских словообразовательных инноваций. Munchen, 2001.

Цейтлин С. Н. Очерки по словообразованию и формообразованию в детской речи. М.: Знак, 2009.

REFERENCES

Akhapkina YA.E., Sosnovtseva YE.G. Korpus detskoy pis'mennoy rechi: StartWrit // Problemy ontolingvistiki – 2017. Osvoyeniye i funktsionirovaniye yazyka v situatsii mnogoyazychiya: Materialy yezhegodnoy mezhdunarodnoy nauchnoy konferentsii, Sankt-Peterburg, 26–28 iyunya 2017 goda, SPb.: LISTOS, 2017. URL: https://www.elibrary.ru /item. asp?id=30048197(data obrashcheniya: 09.04.2021).

Garganeyeva K.V. Motivatsionnyy slovar' detskoy rechi. Tomsk: Izd-vo Tomskogo un-ta, 2007.

Gvozdev A.N. Voprosy izucheniya detskoy rechi. SPb.: Detstvo-Press, 2007.

Gvozdev A.N. Ot pervykh slov do pervogo klassa. Dnevnik nauchnykh nablyudeniy. M.: KoMKniga, 2005.

Gridina T.A. Ob"yasnitel'nyy slovar' detskikh innovatsiy: monografiya i slovar' / Ural. gos. ped. un-t. Yekaterinburg, 2012.

Gridina T.A. «Svoya igrA»: rebenok v mire yazyka: Monografiya. Ye-katerinburg: FGBOU VPO «Ural. gos. ped. un-T», 2015.

Guts YE. N.Assotsiativnyy slovar' podrostka. Omsk: Om. gos. un-t, 2004.

Dve devochki: Sonya i Nadya: dnevnikov·yye zapisi / sost.: O.A. Yuntunen, O.B. Sizova; otv. red. S.N. Tseytlin. SPb.: Tuskarora, 2001.

Lyakso YE. YE., Frolova O. V., Grigor'yev A. S., Ostroukhov A. V. Korpusa detskoy rechi "infant.Ru", "infant.MAVS", "child.Ru", "EmoChildRu" na materiale russkogo yazyka i ikh ispol'zovaniye v issledovaniyakh rechevogo ontogeneza // Teoreticheskaya i prikladnaya lingvistika. 2017. T. 3. № 1. S. 28-58. URL: https://www.elibrary.ru/item.asp?id=29161390 (data obrashcheniya: 09.04.2021).

Ot dvukh do trekh: dnevnikov·yye zapisi / sost.: S.N. Tseytlin, M.B. Yeliseyeva. SPb.: Izd-vo RGPU im. A.I. Gertsena, 1998.

Ot nulya do dvukh: dnevnikov·yye zapisi / sost. S.N. Tseytlin. SPb.: Izd-vo RGPU im. A.I. Gertsena, 1997.

Ovchinnikova I.G., Beresneva N.I., Dubrovskaya L.A., Penyagina YE.B. Leksikon mladshego shkol'nika (kharakteristika leksicheskogo komponenta yazykovoy kompetentsii). Perm', 2000.

Palkin A.D. Vozrastnaya psikholingvistika: Tolkovyy slovar' russkogo yazyka glazami detey. M., 2004.

Rech' russkogo rebenka: Zvuchashchaya khrestomatiya / sost.: T.V. Kuz'mina, E.I. Stolyarova, S.N. Tseytlin. SPb.: Bokhum, 1994.

Kharchenko V.K. Korpus detskikh vyskazyvaniy. M.: Izd-vo Literaturnogo instituta im. A.M. Gor'kogo, 2012.

Tseytlin S. N. Slovar' detskikh slovoobrazovatel'nykh innovatsiy. Munchen, 2001.

Tseytlin S. N. Ocherki po slovoobrazovaniyu i formoobrazo-vaniyu v detskoy rechi. M.: Znak, 2009.

Красноперова Евгения Сергеевна – аспирант кафедры общего языкознания и русского языка. Уральский государственный педагогический университет (Екатеринбург, Россия).

Адрес: 620017, Россия, г. Екатеринбург, пр. Космонавтов, 26, 281.

E-mail: evgenia113@gmail.com

Krasnoperova Evgeniya Sergeevna – Postgraduate at the Department of General Linguistics and Russian Language. Ural State Pedagogical University (Yekaterinburg, Russia).