CC BY

OPEN ACCESS

Научный диалог = Nauchnyi dialog = Nauchnyy dialog
ISSN 2225-756X, eISSN 2227-1295

Шишков М. С. Лексический портрет устного рассказа : корпусное исследование речи двух поколений носителей русского языка / М. С. Шишков // Научный диалог. — 2022. — Т. 11. — № 10. — С. 121—138. — DOI: 10.24224/2227-1295-2022-11-10-121-138.

Shishkov, M. S. (2022). Lexical Portrait of an Oral Story: a Corpus Study of Speech of Two Generations of Native Speakers of Russian. *Nauchnyi dialog, 11(10):* 121-138. DOI: 10.24224/2227-1295-2022-11-10-121-138. (In Russ.).

WEB OF SCIENCE ERIH PLUS ULRICH'S PERIODICALS DIRECTORY НАУЧНАЯ ЭЛЕКТРОННАЯ БИБЛИОТЕКА eLIBRARY.RU EBSCO host

Журнал включен в Перечень ВАК

DOI: 10.24224/2227-1295-2022-11-10-121-138

Лексический портрет устного рассказа : корпусное исследование речи двух поколений носителей русского языка

Шишков Максим Сергеевич [1,2]
orcid.org/0000-0002-3629-2602
ResearcherID: I-4253-2013
кандидат филологических наук, доцент
max-shishkov@yandex.ru

[1] Шанхайский университет иностранных языков (Шанхай, Китай)

[2] Санкт-Петербургский государственный университет (Санкт-Петербург, Россия)

Благодарности:
Исследование выполнено за счет гранта Российского научного фонда (проект № 21-48-04401)

Lexical Portrait of an Oral Story: a Corpus Study of Speech of Two Generations of Native Speakers of Russian

Maxim S. Shishkov [1,2]
orcid.org/0000-0002-3629-2602
ResearcherID: I-4253-2013
PhD in Philology, Associate Professor
max-shishkov@yandex.ru

[1] Shanghai International Studies University (Shanghai, China)

[2] St. Petersburg State University (St. Petersburg, Russia)

Acknowledgments:
The study was supported by a grant from the Russian Science Foundation (project № 21-48-04401)

© Шишков М. С., 2022

CC BY

OPEN ACCESS

Научный диалог = Nauchnyi dialog = Nauchnyy dialog
ISSN 2225-756X, eISSN 2227-1295

ОРИГИНАЛЬНЫЕ СТАТЬИ

Аннотация:

Статья посвящена количественно-статистическому анализу лексики корпуса устных рассказов по картинкам. Испытуемые — 42 человека, монолингвальные носители русского языка (представители 15 семей). Цель исследования — выявить и описать лексические параметры, различающие устные рассказы представителей разных поколений. Делается попытка рассмотреть связь этих параметров с формированием у носителей языка представлений о лексической структуре нарратива. Получены следующие результаты. Характер стимульного материала обеспечил примерно равную степень сложности текстов обеих групп, хотя объем активного словаря у родителей значительно превышает объем словаря детей. На уровне отдельных текстов обе группы используют одинаково разнообразную лексику на единицу объема текста. В рассказах детей основная «сюжетная» лексика довольно разнородна, в то время как у родителей наблюдается тенденция к лексической унификации нарратива. В текстах обеих групп локальная отнесенность и направление движения выражены одинаково, что говорит об их «базовой» нарративной функции, не зависящей от возраста испытуемых. Рассказы родителей унифицированы и с точки зрения лексики, показывающей логическое строение нарратива и разные виды семантических отношений в тексте.

Ключевые слова:
корпус устной речи; количественно-статистический анализ; лексика; два поколения испытуемых; межпоколенческие различия.

ORIGINAL ARTICLES

Abstract:

The article is devoted to the quantitative and statistical analysis of the vocabulary of the oral stories corpus based on pictures. The subjects were 42 people, monolingual native speakers of the Russian (representatives of 15 families). The purpose of the study is to identify and describe the lexical parameters that distinguish the oral stories of representatives of different generations. An attempt is made to consider the connection of these parameters with the formation of native speakers' ideas about the lexical structure of the narrative. The following results are obtained. The nature of the stimulus material ensured an approximately equal degree of complexity in the texts of both groups, although the volume of the parents' active vocabulary significantly exceeded the volume of the children's vocabulary. At the level of individual texts, both groups use equally diverse vocabulary per unit of text volume. In the children's stories, the main "plot" vocabulary is quite heterogeneous, while the parents tend to lexical unification of the narrative. In the texts of both groups, local reference and direction of movement are expressed in the same way, which indicates their "basic" narrative function, which does not depend on the age of the subjects. The stories of the parents are also unified in terms of vocabulary, showing the logical structure of the narrative and different types of semantic relations in the text.

Key words:
corpus of oral speech; quantitative and statistical analysis; vocabulary; two generations of test subjects; intergenerational differences.

CC BY

OPEN ACCESS

Научный диалог = Nauchnyi dialog = Nauchnyy dialog
ISSN 2225-756X, eISSN 2227-1295

УДК 811.161.1’33: 81’373

Лексический портрет устного рассказа: корпусное исследование речи двух поколений носителей русского языка

© Шишков М. С., 2022

1. Введение = Introduction

В последние годы усилилось внимание лингвистов к устной речи, исследование которой получило новый импульс с развитием корпусных методов (см., например: [Корпус ..., 2019; Рассказы ..., 2009]). Устная речь позволяет увидеть ряд закономерностей в развитии лексической и грамматической семантики (см. [Богданова-Бегларян, 2017]), выявить социолингвистические основания происходящих изменений (см. [Богданова-Бегларян, 2016]) и мн. др. В то же время применение методов количественного анализа и интерпретации полученных данных имеет большой потенциал и на уровне анализа смысловой организации устных текстов.

В настоящем исследовании проведена попытка с помощью корпусных методов описать лексический состав текстов, записанных от представителей двух поколений носителей русского языка, и выявить базовые различия этих текстов, которые могут быть объяснены психолингвистическими особенностями освоения лексической системы и постепенным формированием у носителей языка связей лексики с представлением о правилах построения нарратива.

Важным представляется тот факт, что все тексты были получены на основе одного и того же стимульного материала: испытуемым предъявлялась книга М. Майера «Лягушка, где ты?» [Mayer, 1969]. Книга состоит из серии картинок, показывающих процесс поиска мальчиком и его собакой лягушки, сбежавшей ночью из банки.

К материалу книги М. Майера исследователи обращаются довольно часто прежде всего в аспекте изучения особенностей построения детьми нарратива. Как отмечает Т. Хоэл, «впервые “Лягушка” была использована Бамбергом в 1985 году для получения нарративов в научных целях» [Хоэл, 2016, с. 68]. О. В. Федорова, описывая разные методы психолингвистических исследований, отмечает, что рассказ по картинкам — «классический и в то же время самый популярный на сегодняшний день метод сбора корпусов элицитированных текстов» [Федорова, 2016, с. 10].

Использование доступного по содержанию стимульного материала позволяет на примере текстов, полученных от разных групп носителей

CC BY

языка, анализировать целый ряд лингвистических параметров, в том числе определяя языковые характеристики развития ребенка и диагностируя возможные отклонения. Так, например, имеются исследования, в которых книга М. Майера служит основой для изучения процессов языкового развития детей с различными нарушениями работы головного мозга и специфическими нарушениями речи: рассматриваются такие показатели, как количество морфологических ошибок и сложность повествования [“Frog ...”, 2004].

Другое, более частное направление работы с такого рода материалом избрано А. И. Севастьяновой: рассказ на основе книги М. Майера используется для исследования отклонений в употреблении детьми дошкольного возраста глаголов движения [Севастьянова, 2021].

Новизна нашего проекта заключается в том, что в нем исследование лексических параметров устных текстов проводится на основе оценки количественно-статистических показателей лексического разнообразия и сопоставления лексических параметров устных рассказов на одну тему у разных возрастных групп испытуемых.

При проектировании исследования была поставлена задача выявить параметры, которые помогут оценить различие между лексическим составом текстов, продуцируемых детьми и их родителями. В широком смысле данная проблема затрагивается в работах практической направленности (машинное обучение), посвященных кластеризации текстов на разных языках по возрасту автора (одна из частных проблем кластеризации). Как отмечает С. С. Верхозин, «широкое применение количественные методы находят <...> при описании и классификации текстов, например, для авторизации или атрибуции текстов (определение авторства)» [Верхозин, 2013, с. 149]. Исследование статистических характеристик появления новых слов в текстах часто связывается и с проблемой атрибуции текстов (см. методику в: [Baker, 1988]).

Применительно к материалу данного исследования, как представляется, за счет обращения к такому подходу можно увидеть тенденции в освоении и использовании лексики разными поколениями носителей языка (родители и дети), а также проследить влияние этого фактора на уровень развития навыков построения нарратива.

2. Материал, методы, обзор = Material, Methods, Review

В качестве материала исследования выступает корпус текстов 42 устных рассказов монолингвальных носителей русского языка, представителей двух поколений (всего 15 семей: 19 родителей в возрасте от 34 до 50 лет, средний возраст — 40 лет; 23 ребенка в возрасте от 7 до 13 лет,

CC BY

средний возраст — 10 лет). Полученные данные составили два подкорпуса в соответствии с поколением испытуемого. Тексты корпуса были предварительно обработаны для проведения анализа лексической составляющей: устранены обозначения явлений, свойственных устной речи (оборванные слова в случае самокоррекции, хезитативы и т. п.), проведена токенизация (без исключения стоп-слов) и лемматизация текстов корпуса.

Цель исследования — выявить и описать текстовые лексические метрики, которые различают подкорпусы носителей русского языка разных поколений, а также установить связь частотной представленности униграмм и биграмм в текстах двух подкорпусов с формированием у носителей языка представлений о лексической структуре нарратива.

Гипотеза исследования заключалась в предположении, что полученные нарративы будут однородны по содержанию и лексике (наименования реалий, действий героев и т. п.), что позволит более ярко выявить различия между текстами разных поколений испытуемых как в плане использования лексических единиц и их сочетаний, так и в тактиках построения нарратива.

В соответствии с целью исследования анализ корпусного материала состоял из двух этапов.

На первом этапе проведен количественно-статистический анализ текстов подкорпусов с применением методов вычисления ряда текстовых метрик и индексов удобочитаемости.

В качестве методологической базы было принято исследование А. А. Соболева и др., в котором авторы продемонстрировали принципы создания модели машинного обучения с целью определения возраста автора сообщений (дети и взрослые) на основе учета трех групп параметров: базовых статистик текста (количество слов, количество уникальных слов и др.), индексов удобочитаемости, а также коэффициентов лексического разнообразия [Методика ..., 2022, с. 48].

Данный подход к количественно-статистической оценке представляется убедительным и может быть использован не только в прикладной задаче машинного обучения, но и для лингвистической (и психолингвистической) интерпретации полученных текстовых метрик и таблиц, содержащих сведения о текстах подкорпусов носителей языка разного возраста.

Анализ текстовых параметров производился с помощью модулей библиотеки ruTS для языка Python (разработчики С. С. Шкарин, Е. Ю. Смирнова, https://github.com/SergeyShk/ruTS). При проведении анализа базовых статистик учитывались только те параметры, которые могут быть соотнесены в письменном и устном типах текстов.

Использование индексов удобочитаемости в последнее время широко представлено в практической деятельности (см., например: [Ляшев-

CC BY

ская, 2016; Рыбанов, 2011]). Был произведен расчет и анализ классических индексов Флеша-Кинкайда [Derivation ..., 1975] и Флеша [Flesch, 1948], которые опираются на оценку количества слов, предложений и слогов в текстах, а также индекса удобочитаемости SMOG [McLaughlin, 1969].

Отдельно рассматривались коэффициенты лексического разнообразия, вычисляемые по разным методикам (TTR, RTTR, CTTR, MATTR, MSTTR) (реферативное описание методик расчета коэффициентов см., например, в: [Захарова и др., 2020]).

Выбор данных параметров основан на степени значимости информативных признаков, выявленной в исследовании А. А. Соболева и др. в результате оценки дисперсии между группами текстов. Среди наиболее информативных авторами были выделены: индекс удобочитаемости Флеша, тест Флеша-Кинкайда, индекс SMOG, статистика использования длинных, сложных и многосложных слов, а также метрика TTR и несколько других параметров [Методика ..., 2022, с. 49]. На материале русских интернет-текстов, созданных носителями русского языка двух возрастных групп, данные параметры при использовании их для обучения методов классификации текстов показали точность определения до 83,2 % [Методика ..., 2022, с. 51], что говорит и о возможности применения метрик для характеризации группы текстов, в нашем случае — подкорпусов устных рассказов.

На втором этапе составления лексического портрета подкорпусов были сформированы списки униграмм и биграмм отдельно по двум подкорпусам (дети и родители). Проводился анализ единиц, встречающихся в большинстве текстов, в сопоставлении с сюжетом стимульного материала рассказа «Лягушка, где ты?».

3. Результаты и обсуждение = Results and Discussion

3.1. Базовые статистики текстов подкорпусов

Рассмотрим базовые статистики (табл. 1) в абсолютных (общее число единиц в подкорпусе) и относительных (от числа токенов в подкорпусе) значениях.

Анализ данных показывает, что подкорпус родителей характеризуется большими значениями по параметрам, связанным с объемом речевого материала. Большим количеством предложений и слов (в два раза больше, чем в подкорпусе детей) в рассказах обеспечивается большая развернутость высказываний. Средняя длина текста у родителей — 520 слов, у детей — 189 слов.

В абсолютных значениях заметны расхождения в количестве уникальных слов — в 2,66 раза больше в подкорпусе родителей, чем в подкорпусе детей. Количество длинных слов в подкорпусе родителей больше в 2,5 раза, а сложных — в 3,4 раза.

CC BY

Таблица 1

Базовые статистики текстов двух подкорпусов

Метрика	Подкорпус			
	родители		дети	
Предложения	1070		597	
Слова	9874		4352	
Уникальные слова	2428	25 %	933	21 %
Длинные слова	3730	38 %	1486	34 %
Сложные слова	1110	11 %	322	7 %
Простые слова	8280	84 %	3806	87 %
Односложные слова	2752	28 %	1174	27 %
Многосложные слова	6638	67 %	2954	68 %
		(от числа токенов)		(от числа токенов)

При этом частотность появления всех этих типов слов оказывается пропорциональной числу слов во всем подкорпусе: немного выше оказываются у родителей показатели появления уникальных (на 4 %), длинных (на 4 %) и сложных (на 6 %) слов. Дети немного чаще используют простые слова (на 3 %).

3.2. Удобочитаемость текстов подкорпусов

Опишем характеристики текстов подкорпусов с точки зрения индексов удобочитаемости (табл. 2).

Таблица 2

Метрики удобочитаемости текстов двух подкорпусов

Метрика	Подкорпус	
	родители	дети
Тест Флеша-Кинкайда	3,10	1,42
Индекс удобочитаемости Флеша	69,98	78,47
Индекс SMOG	9,05	6,54

Оценка текста по формуле Флеша-Кинкайда указывает на уровень образования: считается, что «значения в интервале 0—10 показывают число классов школы», которые необходимо окончить для понимания текста [Рыбанов, 2011].

При оценке степени трудности (индекс Флеша) детский текст оказывается на уровне легкого чтения (70—80), текст родителей — на границе стандартного и легкого понимания (65—70).

CC BY

Однако данные параметры учитывают только количество слов, предложений и слогов [Методика ..., 2022, с. 48], то есть, скорее, речь идет о сложности восприятия на уровне слова. С другой стороны, при расчете индекса удобочитаемости SMOG учитываются такие параметры, как количество сложных слов и количество предложений в тексте: сложность будет оцениваться на основе характеристик речевых единиц большего объема, требующего более развитых способностей восприятия коммуникативных единиц текста. По этому показателю значения получены выше, но при этом сохраняется такой же разрыв в сложности текстов между двумя подкорпусами.

Детский текст оказывается более понятным, чем взрослый. В то же время можно было ожидать более высокую степень легкости текстов обеих категорий испытуемых, так как стимульный материал — картинки с историей лягушки — подразумевает использование бытовой лексики, а сам сюжет выстраиваемого рассказа ориентирован на детей. Это проявляется и в форме изложения сюжета: многие взрослые испытуемые строят повествование в виде сказки или рассказа для ребенка. Такая нарративная стратегия обеспечивает и на лексическом уровне более высокий коэффициент удобочитаемости. Заметим, что по метрике Флеша-Кинкайда тексты взрослых испытуемых по оценке сложности примерно соответствуют среднему возрасту их детей (средний возраст — 10 лет, возраст обучения в 3-м классе школы).

3.3. Лексическое разнообразие текстов подкорпусов

Объем использованного авторами текстов словаря отражается в другой статистической характеристике — в коэффициентах лексического разнообразия (табл. 3).

Таблица 3

Коэффициенты лексического разнообразия
совокупности текстов двух подкорпусов

Метрика	Подкорпус	
	родители	дети
Type-Token Ratio (TTR)	0,25	0,21
Root Type-Token Ratio (RTTR)	24,43	14,14
Corrected Type-Token Ratio (CTTR)	17,28	10,00
Moving Average Type-Token Ratio (MATTR)	0,85	0,79
Mean Segmental Type-Token Ratio (MSTTR)	0,85	0,79

Метрика TTR (Type-Token Ratio) является самой простой и зависит от объема текста. Получение примерно одинаковых значений TTR для обоих подкорпусов связано именно с разницей в их объеме (количество то-

CC BY

кенов в подкорпусе родителей в два раза больше). Производные от этой метрики — Root Type-Token Ratio (RTTR) и Corrected Type-Token Ratio (CTTR) — показывают большую разницу за счет внесения количества слов в текстах под корень, однако по сути не решают проблему зависимости метрики от длины текста.

Этот недостаток исправляется в метриках Moving Average Type-Token Ratio (MATTR) и Segmental Type-Token Ratio (MSTTR), которые также являются модификациями метрики TTR, но используют скользящую среднюю (MATTR) и сегментирование (MSTTR), что позволяет вычислять метрику независимо от длины текста.

По всем параметрам лексического разнообразия в совокупности текстов видим, что в рассказах родителей наблюдается большее варьирование лексики на единицу объема текста: различие сохраняется и при использовании метрик MATTR и MSTTR, не зависящих от длины текста.

Говоря об интерпретации параметра лексического разнообразия, А. П. Варфоломеев пишет, что «стандартов для коэффициентов разнообразия речи <...> не существует, но ориентиром для сопоставления и, следовательно, оценки какого-либо текста в однородной группе текстов вполне может служить среднестатистическая норма величины коэффициента для равных по длине отрывков» [Психосемантика слова ..., 2000, с. 28].

Рассмотрим среднее значение и значение стандартного отклонения для всех текстов каждого из подкорпусов (табл. 4).

Таблица 4

Среднее значение и стандартное отклонение значений метрик лексического разнообразия (по отдельным текстам подкорпусов)

Метрика	**Подкорпус**			
	родители		**дети**	
	среднее	**станд. отклонение**	**среднее**	**станд. отклонение**
TTR	0,02	0,01	0,04	0,01
RTTR	1,01	0,21	1,41	0,16
CTTR	0,72	0,15	1,00	0,12
MATTR	0,44	0,01	0,44	0,01
MSTTR	0,44	0,01	0,44	0,01

Сопоставляя данные, полученные по совокупности текстов подкорпусов, можно увидеть, что в целом активный словарь, применяемый при описании одинаковых стимульных картинок, у детей меньше, чем словарь родителей. Однако на уровне отдельных текстов оба поколения в среднем используют одинаково разнообразные лексические средства (совпадаю-

CC BY

щие значения средних метрик MATTR и MSTTR). Некоторые рассказы детей кажутся более разнообразными (метрики RTTR и CTTR), что, скорее, будет связано с меньшим объемом текстов этого подкорпуса.

Значения TTR 21 % текстов родителей выходят за пределы стандартного отклонения, среди детских таких текстов оказывается 30 %. По параметрам RTTR, CTTR и MATTR у взрослых доля превышения стандартного отклонения составляет 37 %, у детей — 30 %. При этом количество отклонений в меньшую и большую сторону примерно равно и одинаково представлено у детей и взрослых. Вероятно, данный фактор связан с индивидуальными характеристиками речи испытуемых.

3.4. Частотный портрет лексики подкорпуса речи детей

В текстах детских рассказов только 15 слов (знаменательные части речи) встречаются более чем в трех четвертях текстов: *банка, дерево, домой, дупло, искать, лес, лягушка, лягушонок, мальчик, норка, послать, проснуться, спать, увидеть, улей*. Эти слова связаны с основными моментами сюжета: действующие персонажи, места поиска сбежавшей из банки лягушки, обозначение действий персонажей.

Самое частотное слово — *лягушка* (96 % текстов), непосредственно связанное с темой рассказа. Среди самых частотных слов оказывается это же слово с уменьшительно-ласкательным суффиксом (*лягушонок*, в 74 % текстов). В подкорпусе встречаются и иные обозначения этого героя рассказа: гипероним *животное*, обозначение сходного животного *жаба* (по 2 раза), уменьшительно-ласкательные *лягушечка* (2 раза) и *лягушоночек* (1 раз), а также «детские» варианты *квак, квакушка* (по 1 разу).

Лягушка *вылезла* из банки в 56 % рассказов и в 17 % — *выбралась* (само слово *банка* встретилось в 87 % текстов и не имеет иных обозначений). В незначительном числе рассказов употреблены слова, называющие процесс издавания лягушкой звуков — *кваканье*, *квакать*.

Главный герой в 83 % рассказов определяется как *мальчик*, а также по имени (имена варьируются). Его действия описываются с помощью глаголов *увидеть* (91 %), *искать* (83 %), *проснуться* (78 %), *спать* (70 %), *посмотреть* (65 %), *взять, кричать, звать* (по 48 %). Слова, связанные с процессом поиска: *посмотреть* (65 %), *проверить, поиск, поискать, отправиться* (по 9 %).

Интересно, что почти в четверти текстов для описания действий мальчика в начале рассказа использованы слова *наблюдать, любоваться, играть*.

Стоит отметить ряд глаголов, обозначающих преодоление препятствий для достижения какого-либо места, среди которых самым часто используемым оказывается глагол *залезть* (65 %). Помимо него, встречаются: *забраться, лазить, полезть, перелезть*.

CC BY

Менее чем в половине текстов использованы глаголы *найти, забрать, посадить* (по 39 %).

Все эти глаголы позволяют описать основной сюжет, представленный на картинках. Место жительства мальчика также обозначено в большинстве текстов с помощью слов *домой* (83 %), *комната* (26 %) и *дом* (22 %).

Нет единогласия в назывании собаки. Помимо самых частотных *собака* (52 %) и уменьшительно-ласкательного *собачка* (52 %), встречаются наименования *щенок* (22 %), *пёс* (9 %) и *пёсик, дружок, дог, барбос* (по 4 %). В текстах встречается большое количество глаголов, связанных с действиями собаки (чаще всего — единичные случаи употребления): *подпрыгнуть, лаять, затявкать, гавкнуть, выть, допрыгнуть, гавкать, обнюхать, обнюхивать, обыскивать* и др.

Обозначение места поиска — *лес* — встретилось в 91 % текстов. В группе лексики, связанной с лесом, стоит отметить следующие особенности:

— наибольшее число испытуемых использовало родовое понятие *дерево* (96 %), и лишь небольшая часть использовала видовую лексему *дуб* (17 %);

— поваленное дерево называется *бревном* (39 %), а отверстия в нем — *трещина, тоннель* (по 4 %);

— 78 % текстов включают обозначение *улей*, но имеются и иные обозначения — *кокон, гнездо* (по 1 случаю);

— в большинстве случаев испытуемые называют обитателей улья *пчёлами* (56 %), *пчёлками* (9 %) и используют прилагательное *пчелиный*, хотя встречается и прилагательное *осиный* (в одном тексте);

— в отношении отверстий (в дереве и в земле) испытуемые использовали большой набор слов: от конкретных — *дупло* (91 %) и *норка* (70 %), *нора* (26 %) до более общих — *дырка* (26 %), *дырочка* (9 %);

— обитателей норы и дупла также квалифицируют по-разному: *суслик* (17 %), *сурок* (13 %), *крот* (9 %), *зверёк, мышка, белка* (по 4 %);

— водоем называется испытуемыми *пруд, река, озеро* (по 13 %).

Время действия лексически маркируется с помощью глаголов *проснуться* (78 %) и заснуть (26 %), существительных *утро* (43 %) и *вечер* (22 %), наречия *вечером* (35 %).

В 30 % рассказов употребляются слова, связанные с темой семьи (в рассказах — преимущественно о семье лягушек на болоте): *семья, мама*.

Данные, полученные при анализе униграмм, дополняются информацией, которую можно почерпнуть из списка биграмм, сформированного по критерию их употребительности в рассказах. Обратимся к наиболее частотным (употреблены более чем в трети текстов подкорпуса).

CC BY

Самая частотная биграмма (слова приводятся в нормальной форме) — *из банка* — встречается в 78 % текстов. Среди частотных есть и другие подобные сочетания предлога и существительного: *из норка, на дерево* (по 48 %), *на улица, на утро* (по 35 %). Эти биграммы образуют общие для разных рассказов признаки направленности действий, связанных с ключевыми местами. К этой же группе примыкают частотные биграммы вида «глагол + предлог»: *залезть на* (61 %) и *вылезти из* (43 %).

В 43 % текстов встречается биграмма *маленький лягушонок*. Это словосочетание для данного рассказа можно принять за устойчивую коллокацию, связанную с темой рассказа. Встречается также биграмма *два лягушка* (39 %), используемая при описании найденной семьи лягушек.

Чуть больше половины (52 %) текстов содержит устойчивое сочетание *лечь спать* (тогда как сам глагол *спать* употреблен в 70 % текстов).

К числу частотных биграмм относится также сочетание *потом они* (52 %) и *лягушка потом* (39 %), служащие для организации повествования. Замыкают список частотных биграмм биграммы *увидеть что* (35 %) и *не быть* (35 %), необходимые для развития сюжета: первая вводит в рассказ информацию о результатах поиска, а вторая служит для констатации отрицательного результата.

3.5. Частотный портрет лексики подкорпуса речи родителей

Сопоставим данные подкорпуса детских рассказов с данными подкорпуса речи родителей.

Во всех рассказах родителей присутствуют слова *банка, дерево, дупло* и *лягушка*. В 95% рассказов употреблены слова *лягушонок, маленький, мальчик, очень* и *увидеть*. Более чем в половине текстов совпадают 92 слова, что в два раза больше совпадающих слов в корпусе речи детей (45 слов).

Обозначение персонажа-лягушки в текстах унифицировано: *лягушка* (100 %) и *лягушонок* (95 %), хотя встречаются и отдельные случаи синонимичных наименований, например, *квакуша* и *квакша* (по 5 %). Обозначение главного героя *мальчик* представлено в 95 % случаев, собаки — уменьшительно-ласкательное *собачка* в 58 % текстов. То есть в речи взрослых можно видеть стремление называть персонажа более нейтральным словом.

Количество слов, описывающих процесс поиска лягушки, также становится больше, и каждое из них характеризуется более высокой частотностью: *искать* (89 %), *звать* (84 %), *пойти* (68 %), *заглянуть, кричать, найти* (по 63 %), *идти, посмотреть, смотреть* (по 58 %), *заглядывать, залезть, вылезти, прыгать* (по 53 %).

В текстах родителей расширяется представленность глаголов *быть* (100 %), *стать* (63 %), *находиться* (58 %), а также модального глагола *мочь* (79 %).

CC BY

Увеличивается доля общих для большинства текстов прилагательных, связанных с характеризацией по признакам размера, расстояния и новизны: *маленький* (95 %), *большой* (68 %), *далёкий* (58 %), *новый* (53 %). В 95 % текстов используется наречие *очень*, что показывает желание родителей подчеркнуть степень проявления признаков.

Усиление конкретизации происходит за счет использования числительного *один* во всех текстах подкорпуса.

Становится больше слов, показывающих наличие в текстах рассказов временных (*когда* — 84 %), следственных (*поэтому* — 68 %), целевых (*чтобы* — 58 %), пространственных (*где* — 58 %), определительных (*который* — 68 %) отношений, выражаемых в сложных предложениях.

В число частотных средств выражения идеи отсутствия включаются слова *никто* и *пустой* (по 63 %).

Менее вариативными становятся и названия явлений, изображенных на картинках и связанных с лесом: *дерево* (100 %), *дупло* (100 %), *лес* (89 %), *норка* (74 %), *улей* (53 %), *пчела* (63 %).

Более часто в корпусе речи родителей представлена и лексика, связанная со временем действия: *утро* (74 %), *проснуться* (79 %), *спать* (74 %), *время* (63 %), *однажды* (53 %).

Совпадающих в текстах родителей биграмм значительно больше, чем в текстах детей, но есть и ряд общих частотных биграмм в двух подкорпусах: *маленький лягушонок* (в 74 % текстов родителей), *не быть* (63 %), *из банка* (58 %), *на дерево, на утро, увидеть что* (по 53 %), *из норка, на улица* (по 47 %), *залезть на* (42 %). Очевидно, что данные биграммы можно отнести к необходимым структурным элементам рассказа, не зависящим от возраста испытуемых: в них реализуются простые грамматические правила, используется простая лексика, они служат для выражения отсутствия предмета, пространственных отношений, а также описывают процесс восприятия действительности героями рассказа.

Однако в корпусе рассказов родителей каждая из подгрупп биграмм расширяется. Так, например, добавляется уточнение при отрицании *никто не* (58 %) и указание на степень проявления признака *быть очень* (74 %).

Среди частотных биграмм появляются устойчивые сочетания, в предложениях вводящие придаточные предложения: изъяснительные (*что лягушка, что это, что он, что банка*) и обстоятельственные (*потому что, когда они*). Это свидетельствует о более типичном построении сложных предложений родителями.

Расширяется и перечень биграмм, которые служат для описания процесса восприятия действительности: помимо констатации увиденного — *увидеть что* (по 53 %) и *они увидеть* (58 %), — добавляются частотные

CC BY

сочетания для обозначения процесса зрительного восприятия чего-либо (*смотреть на*, 47 %), а также сочетания, обозначающие процесс анализа увиденного: *понять что* (47 %), *они решить, решить что* (42 %).

В подкорпусе родителей более частотными оказываются сочетания с местоимением *свой*: *свой лягушонок, свой новый* (42 %) (в подкорпусе детей биграммы со словом *свой* встречались менее чем в четверти текстов).

Также можно отметить, что в речи родителей чаще и более устойчиво (37 %) встречается указание на определенность с помощью местоимения *это* (*это время, это дерево, эта лягушка*).

Устойчивым выражением, помимо *маленького лягушонка*, становится в рассказах родителей и *старое дерево* (42 %).

Таким образом, количество и качество лексики, используемой большинством взрослых информантов при описании картинок, свидетельствуют о том, что в этих текстах используется более стандартизованная форма нарратива, чем в текстах детей, к которой взрослые неосознанно стремятся в своих рассказах, опираясь на ключевую лексику. Помимо лексических единиц, связанных с наименованием действующих лиц, в число типичных попадают слова характеризации, глаголы с отвлеченной семантикой, а также большое число единиц, служащих грамматической и семантической организации логического построения текста.

Основной лексический каркас в текстах взрослых в большей степени унифицирован, чем у детей. При этом вариативность лексики в описании предметов и действий героев свидетельствует об объеме активного словарного запаса, возможно, ассоциативно связанного с лексикой бытового плана, известной детям.

4. Заключение = Conclusions

Использование для анализа текстов, полученных в результате предъявления испытуемым одинакового стимульного материала, как предполагалось, позволило выявить ряд особенностей как на уровне лексического состава полученных текстов, так и на связанном с ним уровнем построения нарратива.

Количественно-статистический анализ материалов двух подкорпусов текстов, полученных от монолингвальных русских испытуемых, позволяет сделать следующие выводы о лексических портретах двух подкорпусов устных рассказов.

1. Уровень сформированности словаря у родителей позволяет им при описании бытового сюжета использовать более вариативную лексику, непосредственно не связанную с сюжетом, и строить более длинный текст; при этом активный словарь, который может быть использован при описании одинакового стимульного материала, у детей меньше, чем у родителей.

CC BY

2. Характер предъявленного стимульного материала обеспечил примерно равную сложность текстов обеих групп (стандартное и легкое понимание), а также тот факт, что на уровне отдельных текстов рассказы обоих поколений демонстрируют использование одинаково разнообразной лексики на единицу объема текста.

3. Полученные тексты с точки зрения лексического разнообразия в большой степени зависят от индивидуальных особенностей испытуемого, а не от характера предъявляемого материала, что объясняет выход большого числа текстов за границы стандартного отклонения в обе стороны по этому параметру.

4. Дети наиболее часто используют в рассказах лексику, связанную с действующими лицами, при этом состав лексических групп довольно разнороден, в то время как у родителей наблюдается тенденция к лексической унификации при обозначении этих же понятий.

5. В детских рассказах основная роль отводится глаголам перемещения, а глаголы, связанные с процессом поиска, представлены незначительно; в подкорпусе родителей данный пласт лексики более разнообразен и частотен. Также в рассказах родителей расширяется использование лексики, обозначающей принадлежность и конкретность, и использование признаковых слов.

6. О процессе формирования представлений о категориях нарратива говорит тот факт, что в текстах детей локальная отнесенность и направление движения выражены преимущественно существительными с предлогами; такие же частотные биграммы встречаются и в текстах родителей, что позволяет сделать вывод о нарративной функции этих средств, не зависящей от возраста испытуемых. В то же время рассказы родителей в большинстве случаев включают в себя разнообразную лексику, показывающую логическое строение нарратива и разные виды семантических отношений в тексте, представленные в детских текстах лишь в единичных случаях.

Полученные данные представляют интерес для изучения языкового развития монолингвов в психолингвистическом аспекте, а также могут быть использованы в качестве базы для сопоставления по тем же параметрам с текстами аналогичных рассказов, записанных от информантов — билингвов или носителей унаследованного русского языка. Это позволит, с одной стороны, выявить разницу в объеме и качестве используемой лексики, а с другой — интерпретировать расхождения показателей между текстами разных поколений билингвов, которые объясняются возрастными особенностями испытуемых, а не влиянием фактора освоения второго или унаследованного языка.

CC BY

Источники и принятые сокращения

1. *Mayer M.* Frog, where are you? / M. Mayer. — New York : Dial Books for Young Readers, 1969. — 32 p. — ISBN 9780641931536.

Литература

1. *Богданова-Бегларян Н. В.* Устная спонтанная речь : судьба некоторых грамматических единиц / Н. В. Богданова-Бегларян // Корпусная лингвистика — 2017 : труды международной конференции, Санкт-Петербург, 27—30 июня 2017 года. — Санкт-Петербург : Издательство Санкт-Петербургского государственного университета, 2017. — С. 134-139.

2. *Богданова-Бегларян Н. В.* Функционирование некоторых прагматем русской устной речи в коммуникации представителей разных социальных групп / Н. В. Богданова-Бегларян // Вестник Пермского университета. Российская и зарубежная филология. — 2016. — № 2 (34). — С. 38—49. — DOI: 10.17072/2037-6681-2016-2-38-49.

3. *Верхозин С. С.* О статусе количественных методов в лингвистике / С. С. Верхозин // Вестник Иркутского государственного лингвистического университета. — 2013. — № 3 (24). — С. 145—150.

4. *Захарова Е. Ю.* Лексическое разнообразие текста и способы его измерения / Е. Ю. Захарова, О. Ю. Савина // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. — 2020. — Т. 6. — № 1 (21). — С. 20—-34. — DOI: 10.21684/2411-197X-2020-6-1-20-34.

5. *Корпус* русского языка повседневного общения «Один речевой день» (ОРД) : текущее состояние и перспективы / Н. В. Богданова-Бегларян, О. В. Блинова, Г. Я. Мартыненко, Т. Ю. Шерстинова // Труды института русского языка им. В. В. Виноградова. — 2019. — № 21. — С. 100—110.

6. *Ляшевская О. Н.* Индексы удобочитаемости как мера оценки сложности текста : [Доклад на семинаре НУГ] [Электронный ресурс] — Режим доступа : https://ling.hse.ru/data/2016/12/15/1111563794/Readability%20talk.pdf (дата обращения: 23.10.2022).

7. *Методика* определения возраста автора текста на основе метрик удобочитаемости и лексического разнообразия / А. А. Соболев, А. М. Федотова, А. В. Куртукова, А. С. Романов, А. А. Шелупанов // Доклады Томского государственного университета систем управления и радиоэлектроники. — 2022. — Т. 25. — № 2. — С. 45—52. — DOI: 10.21293/1818-0442-2022-25-2-45-52.

8. *Психосемантика* слова и лингвостатистика текста : Методические рекомендации к спецкурсу / сост. А. П. Варфоломеев ; Калининградский университет. — Калининград : Калининградский государственный университет, 2000. — 37 с.

9. *Рассказы* о сновидениях : корпусное исследование устного русского дискурса / под ред. А. А. Кибрика и В. И. Подлесской. — Москва : Языки славянских культур, 2009. — 736 с. — ISBN 978-5-9551-0303-7.

10. *Рыбанов А. А.* Оценка качества текстов электронных средств обучения // Школьные технологии. — 2011. — № 6. — С. 172—174.

11. *Севастьянова А. М.* Функционирование глаголов движения в речи русскоязычных детей дошкольного возраста (на материале рассказов по картинкам) / А. М. Севастьянова // Проблемы онтолингвистики — 2021 : языковая система ребенка в ситуации одно- и многоязычия : материалы ежегодной Международной научной конференции, Санкт-Петербург, 13—15 апреля 2021 года. — Санкт-Петербург : Издательство ВВМ, 2021. — С. 165—171.

12. *Федорова О. В.* Психолингвистические исследования дискурса в полевой лингвистике / О. В. Федорова // Социо- и психолингвистические исследования. — 2016. — Вып. 4. — С. 7-18.

13. *Хоэл Т.* Нарративы маленьких читателей: модель читателя и реальные читатели / Т. Хоэл // Современное дошкольное образование : теория и практика. — 2016. — № 10 (72). — С. 66—80.

14. "*Frog*, where are you?" Narratives in children with specific language impairment, early focal brain injury, and Williams syndrome / J. Reilly, M. Losh, U. Bellugi, B. Wulfeck // Brain and Language. — 2004. — Vol. 88. — Iss. 2. — Pp. 229—247. — DOI: 10.1016/S0093-934X(03)00101-9.

15. *Baker J. Ch*. Pace: A Test of Authorship Based on the Rate at which New Words Enter an Author's Text / J. Ch. Baker // Literary and Linguistic Computing. — 1988. — Vol. 3. — No. 1. — Pp. 36—39.

16. *Derivation* Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel [Electronic resource] / J. P. Kincaid, R. P. Jr. Fishburne, R. L. Rogers, B. S. Chissom ; Institute for Simulation and Training. — 1975. — Vol. 56. — ii, 39 p. — Access mode : https://stars.library.ucf.edu/istlibrary/56 (accessed 23.10.2022).

17. *Flesch R*. A new readability yardstick / R. Flesch // Journal of Applied Psychology. — 1948. — Vol. 32 (3). — Pp. 221—233. — DOI: 10.1037/h0057532

18. *McLaughlin H*. SMOG grading — a new readability formula / H. McLaughlin // Journal of Reading. — 1969. — No. 12 (8). — Pp. 639—646.

Material resources

Mayer, M. (1969). *Frog, where are you?* New York: Dial Books for Young Readers. 32 p. ISBN 9780641931536.

References

Baker, J. Ch. (1988). Pace: A Test of Authorship Based on the Rate at which New Words Enter an Author's Text. *Literary and Linguistic Computing, 3* (1): 36—39.

Bogdanova-Beglaryan, N V. (2017). Oral spontaneous speech: the fate of some grammatical units. *Corpus Linguistics*. St. Petersburg: St. Petersburg State University Publishing House. 134—139. (In Russ.).

Bogdanova-Beglaryan, N. V. (2016). Functioning of some pragmatems of Russian oral speech in the communication of representatives of different social groups. *Bulletin of the Perm University. Russian and foreign philology*, *2* (34): 38—49. DOI: 10.17072/2037-6681-2016-2-38-49. (In Russ.).

Bogdanova-Beglaryan, N. V., Blinova. O. V., Martynenko, G. Ya., Sherstinova, T. Yu. (2019). *Proceedings of the Russian Language Institute them. V.V. Vinogradov*, *21*: 100—110. (In Russ.).

Fedorova, O. V. (2016). Psycholinguistic studies of discourse in field linguistics. *Socio- and psycholinguistic studies*, *4*: 7—18. (In Russ.).

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology.* 32 (3): 221—233. DOI: 10.1037/h0057532

Hoel, T. (2016). Narratives of Little Readers: The Reader Model and Real Readers. *Modern Preschool Education: Theory and Practice, 10* (72): 66—80. (In Russ.).

Kibrik, A. A., Podlesskaya, V. I. (eds.). (2009) *Dream Stories: A Corpus Study of Oral Russian Discourse*. Moscow: Languages of Slavic cultures. 736 p. ISBN 978-5-9551-0303-7. (In Russ.).

CC BY

Kincaid, J. P., Fishburne, R. P. Jr., Rogers, R. L., Chissom, B. S. (eds.). (1975). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel*. (Institute for Simulation and Training. 56.) ii, 39 p. Available at: https://stars.library.ucf.edu/istlibrary/56 (accessed 23.10.2022).

Lyashevskaya, O. N. *Readability indices as a measure of text complexity assessment: [Report at the seminar of the NUG]*. Available at: https://ling.hse.ru/data/2016/12/15/1111563794/Readability%20talk.pdf (accessed 23.10.2022). (In Russ.).

McLaughlin, H. (1969). SMOG grading — a new readability formula. *Journal of Reading. 12* (8): 639—646.

Reilly, J., Losh, M., Bellugi, U., Wulfeck, B. (2004). "Frog, where are you?" Narratives in children with specific language impairment, early focal brain injury, and Williams syndrome. *Brain and Language, 88* (2): 229—247.DOI: 10.1016/S0093-934X(03)00101-9.

Rybanov, A. A. (2011). Evaluation of the quality of texts of electronic teaching aids. *School technologies*, *6*: 172—174. (In Russ.).

Sevastyanova, A. M. (2021). *The functioning of the verbs of motion in the speech of Russian-speaking children of preschool age (on the basis of stories from pictures)*. St. Petersburg: VVM Publishing House. 165—171. (In Russ.).

Sobolev, A. A., Fedotova, A. M., Kurtukova, A. V., Romanov, A. S., Shelupanov, A. A. (2022). Methodology for determining the age of a text author based on readability and lexical diversity metrics. *Reports of the Tomsk State University of Control Systems and radio electronics*, *25* (2): 45—52. DOI: 10.21293/1818-0442-2022-25-2-45-52. (In Russ.).

Varfolomeev, A. P. (ed.). (2000). *Psychosemantics of the word and linguostatistics of the text: Guidelines for the special course*. Kaliningrad: Kaliningrad State University. 37 p. (In Russ.).

Verkhozin, S. S. (2013). On the status of quantitative methods in linguistics. *Bulletin of the Irkutsk State Linguistic University*, *3* (24): 145—150. (In Russ.).

Zakharova, E. Yu., Savina, O. Yu. (2020). Lexical diversity of the text and methods of its measurement. *Bulletin of the Tyumen State University. Humanitarian research. humanitates*, *6* (1-21): 20—34. DOI: 10.21684/2411-197X-2020-6-1-20-34. (In Russ.).

Статья поступила в редакцию 15.11.2022,
одобрена после рецензирования 10.12.2022,
подготовлена к публикации 25.12.2022.