Научная статья

УДК 81.1
DOI 10.25205/1818-7919-2024-23-2-39-55

Методы машинного обучения для анализа морфологических и лексических особенностей речи мальчиков с расстройствами аутистического спектра и синдромом Дауна

Олеся Владимировна Махныткина [1]
Ольга Владимировна Фролова [2]
Елена Евгеньевна Ляксо [3]

[1] Университет ИТМО
Санкт-Петербург, Россия

[2, 3] Санкт-Петербургский государственный университет
Санкт-Петербург, Россия

[1] makhnytkina@itmo.ru, https://orcid.org/0000-0002-8992-9654
[2] olchel@yandex.ru, https://orcid.org/0000-0002-6293-009X
[3] lyakso@gmail.com, https://orcid.org/0000-0002-6073-0393

Аннотация

Предлагается подход к выявлению различий в речи мальчиков с типичным развитием (ТР), расстройствами аутистического спектра (РАС) и синдромом Дауна (СД) на основе сравнения морфологических и лексических характеристик их речи. В рамках исследования были интервьюированы 69 детей. Для каждой реплики ребенка выделено по 45 лингвистических признаков. На основании критерия Манна – Уитни выявлены различия между детьми с ТР и РАС, детьми с ТР и СД по 31 лингвистическому признаку, между детьми с РАС и СД по 15 признакам. Эти признаки были использованы для построения классификационных моделей методами машинного обучения. Выявленные признаки показали хорошую разделяющую способность, была достигнута точность классификации диалогов равная 88 %.

Ключевые слова

детская речь, лингвистические признаки, машинное обучение, расстройства аутистического спектра, синдром Дауна

Благодарности

Исследование выполнено при финансовой поддержке Российского научного фонда (проект RSF-DST № 22-45-02007)

Для цитирования

Махныткина О. В., Фролова О. В., Ляксо Е. Е. Методы машинного обучения для анализа морфологических и лексических особенностей речи мальчиков с расстройствами аутистического спектра и синдромом Дауна // Вестник НГУ. Серия: История, филология. 2024. Т. 23, № 2: Филология. С. 39–55. DOI 10.25205/1818-7919-2024-23-2-39-55

Machine Learning Methods for Analyzing Morphological and Lexical Characteristics of Speech of Boys with Autism Spectrum Disorders and Down Syndrome

Olesia V. Makhnytkina [1], **Olga V. Frolova** [2], **Elena E. Lyakso** [3]

[1] ITMO University
St. Petersburg, Russian Federation

[2, 3] St. Petersburg State University
St. Petersburg, Russian Federation

[1] makhnytkina@itmo.ru, https://orcid.org/0000-0002-8992-9654
[2] olchel@yandex.ru, https://orcid.org/0000-0002-6293-009X
[3] lyakso@gmail.com, https://orcid.org/0000-0002-6073-0393

Abstract

Purpose. In this paper, we propose an approach to identifying significant differences in the speech of typically developing boys (TD), boys with Autism Spectrum Disorder (ASD) and Down syndrome (DS) based on a comparison of morphological and lexical characteristics of their speech. The linguistic characteristics were extracted automatically using the morphological analyzer pymorphy2. Sixty nine boys were interviewed. In total, 45 linguistic features were extracted from each dialogue.

Results. The Mann – Whitney U test was used for assessing the differences in linguistic features of speech, and differences were identified for 31 linguistic features of speech of boys with TD and with ASD, 31 linguistic features of speech of boys with TD and with DS, and 15 linguistic features of speech of boys with ASD and with DS. These features were used to build classification models using machine learning methods: gradient boosting, random forest, and AdaBoost algorithm. The identified features showed good separability, and the accuracy of the classification of the dialogues of boys with typical development, autism spectrum disorders and Down syndrome equal to 88 % was achieved.

Keywords

children's speech, linguistic features, machine learning, autism spectrum disorder, Down syndrome

Acknowledgements

The study was funded by Russian Science Foundation (project RSF-DST no. 22-45-02007)

For citation

Makhnytkina O. V., Frolova O. V., Lyakso E. E. Machine Learning Methods for Analyzing Morphological and Lexical Characteristics of Speech of Boys with Autism Spectrum Disorders and Down Syndrome. *Vestnik NSU. Series: History and Philology*, 2024, vol. 23, no. 2: Philology, pp. 39–55. (in Russ.) DOI 10.25205/1818-7919-2024-23-2-39-55

Введение

Изучение речи как сложно организованной системы требует использования разных подходов. Традиционно вопросами строения языка, порождением и пониманием языковых высказываний, социальной вариативностью занимается лингвистика. В каждом конкретном случае, в зависимости от изучаемых аспектов лингвистики, анализируются определенные признаки речевых высказываний, например, графематические, лексико-морфологические. Большее число исследований посвящено лингвистике речи типично развивающихся (ТР) детей [Елисеева, 2015]. Для детей с атипичным развитием (АР), воспитывающихся в русскоязычной среде, такие исследования немногочисленны и посвящены изучению грамматических и лексических особенностей речи детей с расстройствами аутистического спектра (РАС) и синдромом Дауна (СД) [Лякso, Фролова, 2017; Городный, Лякso, 2018; Николаев и др., 2019; Лякso и др., 2020].

Согласно Международной классификации болезней (http://mkb–10.com) расстройства аутистического спектра определены как «группа расстройств, характеризующихся качественными отклонениями в социальных взаимодействиях и показателях коммуникабельности,

а также ограниченным, стереотипным, повторяющимся комплексом интересов и действий». Множественная симптоматика нарушений информантов с РАС объединена в «аутистическую триаду» [Kanner, 1943]. Степень выраженности аутистических расстройств и возраст проявления симптоматики, характерной для РАС, индивидуальны [Wing, 1993].

Речевое развитие детей с РАС варьирует от хорошо сформированной речи у высоко функциональных аутистов [Grossman et al., 2013] до использования отдельных вокализаций. Речь детей с РАС представлена в основном вокализациями, отдельными словами и короткими простыми фразами [Tek et al., 2014; Lyakso, Frolova, 2016]. Они имеют более бедный, по сравнению со сверстниками с ТР, лексикон и испытывают затруднения при построении предложений. В их лексике преобладают существительные [Tek et al., 2014], отмечается своеобразное словоупотребление, в частности использование слов в неправильном значении. Дети с РАС отличаются от детей с ТР по использованию личных [Mazzaggio, Shield, 2020] и предметных местоимений [Terzi et al., 2019].

Синдром Дауна (СД, трисомия по хромосоме 21) – одна из форм геномной патологии. Проявления СД представлены обширным спектром нарушений: психических – снижением умственного развития, задержкой речевого развития; морфо-анатомических – замедленным ростом, патологией речевого аппарата, мышечной гипотонией [Лебединский, 2003]. Когнитивные и физические нарушения у детей варьируют от легких до тяжелых. Особенности речи детей с СД обусловлены анатомо-функциональными особенностями строения речевого аппарата ребенка и уровнем когнитивного развития.

Картина нарушений речевого развития детей с СД индивидуальна [Cleland et al., 2010]. Показано, что дети с СД имеют меньшую длину высказываний по сравнению с детьми с ТР, их речь обладает плохой разборчивостью, которая улучшается в подростковом периоде. Дети могут четко произносить отдельные слова и фразы, состоящие из одного-двух слов, но их беглая и развернутая речь – неразборчива [Kumin, 2003]. На материале английского языка показано, что дети с СД чаще, чем дети с ТР, используют устойчивые словосочетания и произносят больше глаголов, чем существительных [Hessling, Brimo, 2019]. На материале немецкого языка у детей с СД отмечены трудности с глагольным словоизменением: образованием форм прошедшего времени [Penke, 2019], согласованием сказуемого в лице и числе с подлежащим [Penke, 2018].

Анализ частей речи, используемых русскоязычными детьми 6–7 лет, показал, что дети с ТР используют в речи преимущественно существительные и глаголы, дети с РАС и с СД – частицы и вокализации. Дети с РАС реже употребляют глаголы, прилагательные, наречия, числительные, местоимения, предлоги и союзы по сравнению со сверстниками с ТР. Для детей с СД обнаружена достоверно меньшая частота употребления глаголов, местоимений, прилагательных, наречий, числительных, предлогов и союзов, чем для детей с ТР [Городный, Ляксо, 2018]. Фонетический анализ показал в речи детей с РАС нетипичные для русского языка фонемы, несформированность некоторых групп согласных и замены согласных [Николаев и др., 2019]. Во всех представленных исследованиях лингвистический анализ речи детей проводили вручную.

В настоящее время широко обсуждается возможность использования автоматического распознавания психоневрологического состояния ребенка по его речи в качестве одного из методов дополнительной диагностики [Fusaroli et al., 2017; Matveev et al., 2021]. Создается большое количество мобильных приложений для детей, в том числе с АР [Adamu et al., 2019]. Встраивание автоматической классификации речи детей с ТР и АР в такие приложения позволит адаптироваться под пользователя. Для автоматической классификации речи детей используются как звучащая речь, так и тексты.

Исследователи выделяют акустические признаки, которые позволяют классифицировать детей на две группы – РАС / ТР [Fusaroli et al., 2017], и ставят задачу определения специфических акустических признаков, позволяющих отличить речь детей с РАС от речи детей

с другими диагнозами – умственной отсталостью, смешанными специфическими расстройствами психологического развития, СД [Lyakso et al., 2018; Matveev et al., 2021].

В работе [Cho et al., 2019] впервые была показана возможность совместного применения лингвистических и акустических признаков для автоматического определения методами машинного обучения аутистических расстройств по речи ребенка, ранее в работах рассматривались только акустические признаки [Pokorny et al., 2017]. На материале русского языка работы по автоматической классификации речи детей с использованием лингвистических признаков отсутствуют.

Целью исследования явилась автоматическая классификация речи типично развивающихся детей, детей с расстройствами аутистического спектра и с синдромом Дауна по специфическим лингвистическим признакам с использованием методов машинного обучения.

Методика исследования

Набор данных

В исследовании использовался оригинальный набор данных, содержащий диалоги экспериментатора с детьми с ТР, СД и РАС. В выборку включен речевой материал 69 мальчиков в возрасте от 8 до 11 лет (табл. 1). Выбор речевого материала мальчиков обусловлен большей частотой проявления аутистических расстройств у лиц мужского пола, чем женского [Nicholas et al., 2008]. Выбор возраста информантов связан с тем, что дети обладают достаточными вербальными навыками для общения с другими людьми. Все дети посещали школу (дети с ТР – общеобразовательную, дети с РАС и СД – коррекционную школу VIII вида). Уровень речевого развития детей с РАС и СД предполагал возможность использования ими отдельных слов и фраз. Диагноз детям с РАС был подтвержден детским психиатром СПбГПМУ. Для оценки степени выраженности аутистических расстройств использовали шкалу CARS (Childhood Autism Rating Scale) [Shopler et al., 1980]. В исследовании приняли участие дети с РАС, имеющие баллы по шкале CARS 31–42 (31–37 – средняя степень; 38–60 – тяжелая форма РАС, дети с высокофункциональным аутизмом не включены в исследование). По баллам вербального развития по шкале CARS все дети с РАС не различались.

Таблица 1

Число детей каждого возраста, принявших участие в исследовании

Table 1

Number of children of each age who took part in the study

Диагноз	Возраст, лет				Всего
	8	9	10	11	
ТР	5	5	5	5	20
РАС	8	10	5	12	35
СД	4	2	4	4	14
Всего	17	17	14	21	69

Запись речи детей производили в условиях школы, лаборатории и детского центра в стандартизированных условиях в модельной ситуации «диалог с экспериментатором». Диалог включал стандартный набор вопросов о семье, друзьях, прогулках, любимых занятиях, школе. Экспериментатор чередовал общие и специальные вопросы, чтобы создать впечатление естественного взаимодействия. Стратегия экспериментатора основана на максимальном привлечении внимания ребенка, речь и поведение взрослого эмоциональны, но регламентированы дизайном эксперимента. Диалоги с детьми с РАС и СД осуществляли в присутствии родителей.

Для записи речевого материала использована профессиональная аппаратура: магнитофон «Marantz PMD660» с выносным микрофоном «SENNHEIZER e835S». Расстояние от лица ребенка до микрофона составляло 30–50 см. Параллельно с аудиозаписью осуществляли видеозапись поведения детей на камеру «SONY HDR-CX560». Речевые файлы сохраняли в формате Windows PCM WAV, 44 100 Гц, 16 бит; видеофайлы – в формате AVI.

Транскрибирование диалогов осуществлялось экспертами – специалистами в области изучения детской речи. В рамках исследования анализировали только реплики детей.

В наборе данных содержится 69 файлов с репликами детей из диалогов. Описательные характеристики набора данных представлены в табл. 2.

Таблица 2

Описание набора данных

Table 2

Description of the dataset

Признаки	ТР	СД	РАС
Количество реплик	834	1 173	2 550
Количество предложений	1 105	1 206	2 930
Количество токенов	5 319	1 737	5 348

Методы машинного обучения

С целью определения возможности использования методов машинного обучения для задачи классификации речи детей с ТР, СД и РАС проведено данное исследование. На первом этапе проводились исследование и отбор лингвистических признаков, отражающих особенности речи детей с ТР, РАС и СД. Для выявления особенностей речи использовался лексический, морфологический и частотный анализ. На втором этапе исследования осуществлялся выбор значимых признаков для решения задачи классификации речи детей. Выбор значимых признаков является важным этапом, так как универсального набора признаков не существует, и для различных задач классификации речи набор может существенно отличаться. На третьем этапе исследования осуществлялся выбор метода машинного обучения для решения задачи классификации речи детей.

Лексический анализ – первичный этап в процессе автоматической обработки текстов на естественном языке. Основной задачей лексического анализа является выделение структурных единиц из входного текста, а именно предложений, абзацев, слов (токенов), знаков препинания и т. д. Транскрибирование реплик детей производилось вручную, что позволило выделить паузы в речи, используя знаки пунктуации (…). Деление текста на предложения производили по стандартным правилам, если последовательность слов заканчивается такими знаками препинания, как точка, вопросительный знак, восклицательный знак. В соответствии с разметкой использовали правило: если последовательность слов заканчивается паузой (…) и далее текст начинается с прописной буквы, то производится деление на предложения; если последовательность слов заканчивается паузой (…) и далее текст начинается со строчной буквы, то деления на предложения нет.

Пример:

Ну, нравятся … Но математика лучше нравится. – 2 предложения

Количество пауз рассчитывали на основании расчета количества вхождений последовательности символов «…» в тексте, при этом учитывали паузы после неоконченных слов.

Пример:

Больш… Большие такие как бы, красивые. – 1 пауза

Количество неоконченных слов рассчитывали на основе последовательности букв, не являющихся конечным словом, и наличием паузы после него (…).

Пример:

Больш… Большие такие как бы, красивые. – 1 неоконченное слово

Таким образом, после проведения лексического анализа были извлечены следующие признаки:

1) количество предложений в репликах респондента в диалоге;
2) количество пауз во всех репликах респондента в диалоге;
3) среднее количество пауз в реплике респондента в диалоге;
4) количество неоконченных слов во всех репликах респондента в диалоге;
5) среднее количество неоконченных слов в реплике респондента в диалоге;
6) среднее количество токенов в предложении респондента в диалоге;
7) среднее количество токенов в реплике респондента в диалоге.

Морфологический анализ текстов с использованием библиотеки pymorphy2 можно проводить в двух режимах: 1) простым поиском по словарю OpenCorpora; 2) простым поиском с использованием словаря и предсказателем на основе правил для незнакомых слов. В библиотеке pymorphy2 реализована возможность не только разбирать словарные слова, но и для несловарных слов автоматически задействовать предсказатель. В документации разработчики приводят пример: «Например, попробуем разобрать слово “бутявковедами” – pymorphy2 поймет, что это форма творительного падежа множественного числа существительного “бутявковед” и что “бутявковед” – одушевленный и мужского рода».

Для формирования частотного словаря лексики детей осуществлялась лемматизация, (приведения словоформы к лемме – ее нормальной (словарной) форме). В русском языке нормальными формами считаются следующие морфологические формы: для существительных – именительный падеж, единственное число; для прилагательных – именительный падеж, единственное число, мужской род; для глаголов, причастий, деепричастий – глагол в инфинитиве несовершенного вида.

После проведения морфологического анализа были вычислены относительные частоты употребления следующих частей речи для каждого ребенка:

1) имя прилагательное (полное) – ADJF;
2) имя прилагательное (краткое) – ADJS;
3) наречие – ADVB;
4) компаратив – COMP;
5) союз – CONJ;
6) глагол (инфинитив) – INFN;
7) междометие – INTJ;
8) несуществующее слово – None;
9) имя существительное – NOUN;
10) местоимение-существительное – NPRO;
11) числительное – NUMR;
12) частица – PRCL;
13) предикатив – PRED;
14) предлог – PREP;
15) причастие (краткое) – PRTS;
16) глагол (личная форма) – VERB.

Оценка употребления позитивных и негативных слов проводилась с использованием тонального словаря LinisCrowd 2015 (http://linis-crowd.org/). Для каждого диалога составляли список тональных слов и вычисляли среднее, максимальное, минимальное, суммарное значения тональности слов, количество позитивных и негативных слов.

Для выбора значимых признаков проводилась проверка гипотезы о различии двух независимых выборок по уровню выраженности изучаемого признака. Оценка нормальности распределения признаков с использованием теста Колмогорова – Смирнова показала, что выборка не происходит из нормального распределения. Так как количество наблюдений каждого класса достаточно мало, то для проверки гипотезы о различии двух независимых выборок по уровню выраженности изучаемого признака использовали критерий U Манна – Уитни. На втором этапе формировали набор данных с признаками, имеющими различия на уровне значимости 0,05.

Для подтверждения гипотезы о значимости различий значений лингвистических признаков решалась задача классификации диалогов на основе использования классических методов машинного обучения, с учетом наличия или отсутствия диагноза у детей. В качестве базового метода рассматривался градиентный бустинг (Gradient Boosted Decision Trees), который показал хорошие результаты при решении задачи классификации текстов детей с ТР и РАС [Cho et al., 2019]. Градиентный бустинг представляет собой ансамбль деревьев решений, обучение которых происходит последовательно. С целью повышения качества классификации рассматриваются другие ансамбли моделей на основе деревьев решений: случайный лес (Random Forest), алгоритм AdaBoost. Для оценивания обобщающей способности рассмотренных алгоритмов классификации использовался метод кросс-валидации LeaveOneOut. Все рассматриваемые методы реализованы с использованием библиотеки scikit-learn в Python [1].

Результаты исследования

Морфологический анализ текста позволил выявить наиболее частотные слова в речи детей с ТР, РАС и СД. Значительная часть наиболее употребительных слов относится к группе функциональных (предлоги, местоимения, артикли и союзы), которые обозначают отношения между словами, но сами не несут почти никакой смысловой нагрузки. В результате исключения из анализа функциональных слов (ADVB – наречие, CONJ – союз, INTJ – междометие, None – несуществующее, NPRO – местоимение-существительное, PRCL – частица, PRED – предикатив, PREP – предлог) получен словарь частотной лексики (табл. 3).

Для транскрипций реплик детей с ТР предсказатель позволял провести полный морфологический анализ большего числа слов, так как часть слов, имеющихся в нашей повседневной речи, в словаре отсутствовала:

универ – существительное, неодушевлённое, единственное число.

Для транскрипций диалогов детей с атипичным развитием с использованием предсказателя были получены неоднозначные результаты:

люиай – глагол, несовершенный вид, говорящий не включен в действие, повелительное наклонение, единственное число.

Далее в исследовании морфологический анализ проводился только на основе словаря (первый режим морфологического анализа). Частично несуществующим словам создателями набора данных были поставлены в соответствие слова из словаря, и статистический анализ употребления частей речи проводился на основе морфологического разбора слова из словаря:

здрасьте = здравствуйте – глагол, множественное число, несовершенный вид, говорящий не включен в действие.

[1] https://scikit-learn.org/stable/.

Таблица 3

Частотность 10 слов,
наиболее часто встречающихся в лексиконе детей

Table 3

Frequency of the 10 words
most frequently found in children's vocabulary

Диагноз	ТОП-10			
	слов		после удаления функциональных слов	
ТР	ну	213	все	52
	да	188	люблю	37
	и	174	нравится	29
	в	152	больше	20
	там	128	три	18
	я	109	хожу	15
	не	102	такие	15
	на	91	знаю	15
	а	81	один	14
	это	79	русский	14
СД	да	149	мама	32
	нет	83	стой	14
	не	38	коза	9
	там	32	тигр	9
	мама	32	пантера	8
	и	23	волк	8
	ыыы	22	лев	8
	а	22	морковка	7
	это	21	мой	7
	кто	19	медведь	7
РАС	да	342	мама	40
	а	120	папа	37
	это	102	тигр	31
	не	97	один	24
	я	79	три	19
	и	79	два	18
	ну	67	знаю	17
	угу	64	кот	16
	нет	59	улыбается	15
	что	55	хочет	14
			мальчик	14

Среди признаков по результатам морфологического анализа однородными оказались только относительные частоты основных частей речи в репликах детей с ТР: прилагательное, наречие, существительное, частица, глагол.

Частотный анализ: для наиболее употребляемых частей речи (имя существительное, глагол) определены граммемы по следующим категориям, принятым в OpenCorpora (http://opencorpora.org/dict.php?act = gram):

1. категория одушевленности (animacy):

1.1. anim – одушевленное;

1.2. inan – неодушевленное;
2. число (number):
2.1. sing – единственное число;
2.2. plur – множественное число;
3. категория вида:
3.1. perf – совершенный вид;
3.2. impf – несовершенный вид;
4. категория совместности (involvement):
4.1. incl – говорящий включен (идем, идемте);
4.2. excl – говорящий не включен в действие (иди, идите);
5. категория наклонения:
5.1. Indc – изъявительное наклонение;
5.2. Impr – повелительное наклонение;
6. категория лица (у глаголов будущего и настоящего времени):
6.1. 1per – 1 лицо;
6.2. 2per – 2 лицо;
6.3. 3per – 3 лицо;
7. категория времени (кроме повелительно наклонения):
7.1. pres – настоящее время;
7.2. past – прошедшее время;
7.3. futr – будущее время.

В тональном словаре для каждого слова определена его тональная оценка, например, для слова «лгать» тональная оценка равна 1,67, а для слова «мягкий» – 0,5. Показано, что позитивные слова преобладают в речи детей с ТР, негативные – в речи детей с СД. Выявлены значимые различия в речи детей с ТР и детей с РАС и СД на уровне лексических и морфологических признаков (табл. 4).

Таблица 4

Особенности в речи детей с типичным и атипичным развитием

Table 4

Features in the speech of children with typical and atypical development

Категория		ТР	СД	РАС
Количество реплик	ТР		+*	–
	СД	+*		–
	РАС	–	–	
Количество предложений	ТР		+**	+**
	СД	+**		–
	РАС	+**	–	
Среднее количество предложений в 1 реплике	ТР		+***	+***
	СД	+***		+**
	РАС	+***	+**	
Среднее количество токенов в реплике	ТР		+***	+***
	СД	+***		+**
	РАС	+***	+**	
Среднее количество токенов в предложении	ТР		+***	+***
	СД	+***		+***
	РАС	+***	+***	

Продолжение табл. 4

Категория		ТР	СД	РАС
Количество пауз	ТР		+***	–
	СД	+***		+*
	РАС	–	+*	
Количество неоконченных слов	ТР		–	+**
	СД	–		–
	РАС	+**	–	
ADJF	ТР		+***	+***
	СД	+***		+*
	РАС	+***	+*	
ADJS	ТР		+*	–
	СД	+*		–
	РАС	–	–	
ADVB	ТР		+***	+***
	СД	+***		–
	РАС	+***	–	
COMP	ТР		+***	+***
	СД	+***		–
	РАС	+***	–	
CONJ	ТР		+***	+***
	СД	+***		+*
	РАС	+***	+*	
INFN	ТР		+***	+***
	СД	+***		+*
	РАС	+***	+*	
INTJ	ТР		+***	+*
	СД	+***		+*
	РАС	+*	+*	
None	ТР		+***	+***
	СД	+***		–
	РАС	+***	–	
NPRO	ТР		+***	+***
	СД	+***		–
	РАС	+***	–	
NUMR	ТР		+***	+*
	СД	+***		+*
	РАС	+*	+*	
PRCL	ТР		+**	+**
	СД	+**		–
	РАС	+**	–	
PREP	ТР		+***	+**
	СД	+***		+*
	РАС	+**	+*	
VERB	ТР		+***	+*
	СД	+***		–
	РАС	+*	–	

Продолжение табл. 4

Категория		ТР	СД	РАС
NOUN_plur	ТР		+***	+***
	СД	+***		–
	РАС	+***	–	
NOUN_sing	ТР		+**	+***
	СД	+**		–
	РАС	+***	–	
NOUN_anim	ТР		–	+**
	СД	–		–
	РАС	+**	–	
NOUN_anim	ТР		–	+**
	СД	–		–
	РАС	+**	–	
NOUN_inan	ТР		+*	+**
	СД	+*		–
	РАС	+**	–	
VERB_plur	ТР		+***	+**
	СД	+***		–
	РАС	+**	–	
VERB_impf	ТР		+***	+*
	СД	+***		–
	РАС	+*	–	
VERB_excl	ТР		–	–
	СД	–		+*
	РАС	–	+*	
VERB_None	ТР		+***	+*
	СД	+***		+**
	РАС	+*	+**	
VERB_impr	ТР		–	–
	СД	–		+*
	РАС	–	+*	
VERB_indc	ТР		+***	+*
	СД	+***		+**
	РАС	+*	+**	
VERB_past	ТР		+***	+**
	СД	+***		+*
	РАС	+**	+*	
VERB_pres	ТР		+*	+*
	СД	+*		–
	РАС	+*	–	
VERB_1per	ТР		+***	+***
	СД	+***		–
	РАС	+***	–	

Окончание табл. 4

Категория		ТР	СД	РАС
Максимальное значение тональности слов	ТР		+**	+*
	СД	+**	–	–
	РАС	+*	–	–
Количество позитивных слов	ТР	–	+*	+*
	СД	+*	–	–
	РАС	+*	–	–

Условные обозначения: минус (–) – нет различий; плюс (+) – есть различия; * – $p < 0{,}05$; ** – $p < 0{,}01$; *** – $p < 0{,}001$.

Значимые различия выявлены на уровне лексических и морфологических признаков. По большинству критериев наблюдаются различия между речью детей с ТР и детей с СД, детей с ТР и детей с РАС. Значимые различия между речью детей с РАС и детей с СД выявлены для признаков: среднее количество предложений в одной реплике; среднее количество токенов в реплике; среднее количество токенов в предложении; доля союзов, междометий, глаголов; доля глаголов повелительного и изъявительного наклонения; доля глаголов, обозначающих, что говорящий не включен в действие.

С использованием значимых признаков проведены эксперименты по построению классификационных моделей, позволяющих по репликам из диалогов определять принадлежность реплики ребенку с ТР, СД, РАС. Результаты экспериментов представлены в табл. 5.

Таблица 5

Результаты классификации диалогов детей

Table 5

Results of classification of children’s dialogues

Метод	Диагноз	Точность / precision	Полнота / recall	F1-мера	Точность / accuracy
Градиентный бустинг	ТР	0,89	0,80	0,84	0,75
	СД	0,60	0,64	0,62	
	РАС	0,75	0,77	0,76	
Случайный лес	ТР	0,90	0,95	0,93	**0,88**
	СД	0,91	0,71	0,80	
	РАС	0,86	0,91	0,89	
AdaBoost	ТР	0,90	0,95	0,93	0,83
	СД	0,67	0,71	0,69	
	РАС	0,85	0,80	0,82	
Базовая модель	ТР	0,71	0,86	0,78	0,76
	РАС	0,82	0,66	0,73	

Использование метода «случайный лес» для предсказания меток классов позволило достичь точности 88 %. В результате 95 % диалогов детей с ТР были правильно классифицированы (19 из 20, один диалог был отнесен к РАС). Для детей с атипичным развитием правильно классифицированы 71 % диалогов детей с СД (11 из 14, 3 диалога были отнесены к РАС), 91 % диалогов детей с РАС (31 из 35, 2 диалога были отнесены к диалогу детей с ТР

и 2 – к диалогу детей с СД). Результаты экспериментов значительно превышают базовый уровень [Cho et al., 2019]. Однако это может быть связано, в том числе, с разной степенью тяжести РАС у детей в нашем эксперименте и экспериментах коллег. Полученное решение также превышает результаты, опубликованные в работе [Makhnytkina et al., 2021], за счет рассмотрения новых морфологических признаков и тональностей слов.

Заключение

В статье представлены результаты пилотного исследования различий в речи мальчиков с ТР, СД и РАС с использованием методов машинного обучения. Выявлены лингвистические особенности диалогов детей с типичным и атипичным развитием. Из 45 рассматриваемых признаков 36 являются различительными для изучаемых групп. В среднем значения значимых лингвистических признаков диалогов детей с РАС находятся между значениями признаков для диалогов детей с ТР и детей с СД. Подход на основе использования классификатора «случайный лес» позволил добиться точности классификации диалогов в 88 %, при этом наилучшее качество было достигнуто для диалогов детей с типичным развитием (accuracy = 95 %) и детей с РАС (accuracy = 91 %), а наихудшее – для детей с СД (accuracy = 71 %). В ряде исследований, например [Cho et al., 2019], было показано, что для повышения точности классификации диалогов детей с типичным и атипичным развитием полезным является объединение аудио- и текстовой модальности. В дальнейшем мы планируем проанализировать речевой материал детей с другими типами нарушений развития для уточнения и подтверждения специфичности выявленных в данном исследовании различий между группами детей и провести эксперименты по мультимодальному (текстовые и аудиоданные) распознаванию состояний детей, что может улучшить качество классификационных моделей.

Список литературы

Городный В. А., Ляксо Е. Е. Характеристика речи детей 6–7 лет с расстройствами аутистического спектра и синдромом Дауна // Теоретическая и прикладная лингвистика. 2018. Т. 4, № 2. С. 22–37.

Елисеева М. Б. Становление индивидуальной языковой системы ребенка. Ранние этапы. М.: ЯСК, 2015. 344 с.

Лебединский В. В. Нарушения психического развития в детском возрасте. М.: Академия, 2003. 140 с.

Ляксо Е. Е., Фролова О. В., Гречаный С. В., Матвеев Ю. Н., Верхоляк О. В., Карпов А. А. Голосовой портрет ребенка с типичным и атипичным развитием: Монография. СПб.: Изд.-полигр. ассоциация высших учебных заведений, 2020. 204 с.

Ляксо Е. Е., Фролова О. В. Анализ текстов речи «взрослый – ребенок», «взрослый – взрослый» при нормальном и атипичном развитии информантов // Теоретическая и прикладная лингвистика. 2017. Т. 2. С. 20–47.

Николаев А. С., Фролова О. В., Городный В. А., Ляксо Е. Е. Характеристика ответных реплик детей 5–11 лет с расстройствами аутистического спектра в диалогах со взрослыми // Вопросы психолингвистики. 2019. Т. 4, № 42. С. 92–105.

Adamu A. S., Abdullahi S. E., Aminu R. K. A Survey on Software Applications use in Therapy for Autistic Children // 15th International Conference on Electronics, Computer and Computation (ICECCO). 2019. P. 1–4. DOI 10.1109/ICECCO48375.2019.9043237

Cho S., Liberman M., Ryant N., Cola M., Schultz R. T., Julia Parish-Morris J. Automatic detection of Autism Spectrum Disorder in children using acoustic and text features from brief natural conversations // Proc. Interspeech 2019: 20th Annual Conference of the International Speech Communication Association. 2019. P. 2513–2517. DOI 10.21437/Interspeech.2019-1452

Cleland J., Wood S., Hardcastle W., Wishart J., Timmins C. Relationship between speech, oromotor, language and cognitive abilities in children with Down's syndrome // International Journal of Language and Communication Disorders. 2010. Vol. 45 (1). P. 83–95.

Fusaroli R., Lambrechts A., Bang D., Bowler D. M., Gaigg S. B. Is Voice a Marker for Autism Spectrum Disorder? // A Systematic Review and Meta-Analysis. Autism Research. 2017. Vol. 10. P. 384–407.

Grossman R. B., Edelson L. R., Tager-Flusberg H. Emotional facial and vocal expressions during story retelling by children and adolescents with high-functioning autism // Journal of Speech, Language, and Hearing Research. 2013. Vol. 56 (3). P. 1035–1044.

Hessling A., Brimo D. M. Spoken fictional narrative and literacy skills of children with Down syndrome // Journal of Communication Disorders. 2019. Vol. 79. P. 76–89. DOI 10.1016/j.jcomdis.2019.03.005

Kanner L. Autistic disturbances of affective contact // Nervous Child. 1943. Vol. 2. P. 217–250.

Kumin L. Early communication skills for children with Down syndrome: A guide for parents and professionals. Bethesda, MD: Woodbine House, 2003. 391 p.

Lyakso E., Frolova O. Speech features of typically developing children and children with autism spectrum disorders. In: Abstract book. BIT's 7th Annual World Congress of Neurotalk – 2016. Innovation of Neuroscience. 2016.

Lyakso E., Frolova O., Karpov A. A New Method for Collection and Annotation of Speech Data of Atypically Developing Children // International Conference on Sensor Networks and Signal Processing (SNSP). 2018. P. 175–180. DOI 10.1109/SNSP.2018.00040

Makhnytkina O. V., Grigorev A., Nikolaev A. Analysis of dialogues of typically developing children, children with Down syndrome and ASD using machine learning methods // Lecture Notes in Computer Science. 2021. P. 397–406. DOI 10.1007/978-3-030-87802-3_36.

Matveev Y., Matveev A., Frolova O., Lyakso E. Automatic Recognition of the Psychoneurological State of Children: Autism Spectrum Disorders, Down Syndrome, Typical Development // Speech and Computer. SPECOM 2021. Lecture Notes in Computer Science. 2021. Vol. 12997. P. 417–425. DOI 10.1007/978-3-030-87802-3_38

Mazzaggio G., Shield A. The Production of Pronouns and Verb Inflections by Italian Children with ASD: A New Dataset in a Null Subject Language // Journal of Autism and Developmental Disorders. 2020. Vol. 50. P. 1425–1433. DOI 10.1007/s10803-019-04349-7

Nicholas J. S., Charles J. M., Carpenter L. A., King L. B., Jenner W., Spratt E. G. Prevalence and characteristics of children with autism-spectrum disorders // Annals of Epidemiology. 2008. Vol. 18 (2). P. 130–136.

Penke M. Verbal Agreement Inflection in German Children With Down Syndrome // Journal of Speech, Language, and Hearing Research. 2018. Vol. 61 (9). P. 2217–2234. DOI 10.1044/2018_JSLHR-L-17-0241

Penke M. Regular and irregular inflection in Down syndrome – New evidence from German // Cortex. 2019. Vol. 116. P. 192–208. DOI 10.1016/j.cortex.2018.08.010

Pokorny F. B., Schuller B., Marschik P. B., Brueckner R., Nyström P., Cummins N., Bölte S., Einspieler C., Falck-Ytter T. Earlier Identification of Children with Autism Spectrum Disorder: An Automatic Vocalisation-Based Approach // Interspeech. 2017. P. 309–313. DOI 10.21437/Interspeech.2017-1007

Schopler E., Reichler R. J., DeVellis R. F., Daly K. Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS) // Journal of Autism and Developmental Disorders. 1980. March. No. 10 (1). P. 91–103. DOI 10.1007/BF02408436

Tek S., Mesite L., Fein D., Naigles L. Longitudinal analyses of expressive language development reveal two distinct language profiles among young children with autism spectrum disorders // Journal of Autism and Developmental Disorders. 2014. Vol. 44 (1). P. 75–89.

Terzi A., Marinis T., Zafeiri A., Francis K. Subject and Object Pronouns in High-Functioning Children with ASD of a Null-Subject Language // Frontiers in Psychology. 2019. Vol. 10, no. 1301. P. 1–8. DOI 10.3389/fpsyg.2019.01301

Wing L. The definition and prevalence of autism: a Review // European Child and Adolescent Psychiatry. 1993. Vol. 2 (1). P. 61–74.

References

Adamu A. S., Abdullahi S. E., Aminu R. K. A Survey on Software Applications use in Therapy for Autistic Children. In: 15th International Conference on Electronics, Computer and Computation (ICECCO), 2019, pp. 1–4. DOI 10.1109/ICECCO48375.2019.9043237

Cho S., Liberman M., Ryant N., Cola M., Schultz R.T., Julia Parish-Morris J. Automatic detection of Autism Spectrum Disorder in children using acoustic and text features from brief natural conversations. In: Proc. Interspeech 2019: 20th Annual Conference of the International Speech Communication Association, 2019, pp. 2513–2517. DOI 10.21437/Interspeech.2019-1452

Cleland J., Wood S., Hardcastle W., Wishart J., Timmins C. Relationship between speech, oromotor, language and cognitive abilities in children with Down's syndrome. *International Journal of Language and Communication Disorders*, 2010, vol. 45 (1), pp. 83–95.

Eliseeva M. B. Stanovlenie individual'noi yazykovoi sistemy rebenka. Rannie etapy [Formation of a Child's Individual Language System: Early Stages]. Moscow, Yazyki slavyanskoi kul'tury Publ., 2015, 344 p. (in Russ.)

Fusaroli R., Lambrechts A., Bang D., Bowler D. M., Gaigg S. B. Is Voice a Marker for Autism Spectrum Disorder? *A Systematic Review and Meta-Analysis. Autism Research*, 2017, vol. 10, pp. 384–407.

Gorodnyi V. A., Lyakso E. E. Kharakteristika rechi detei 6–7 let s rasstroistvami autisticheskogo spektra i sindromom Dauna [Characteristic of Speech of Children aged 6–7 years with Autism Spectrum Disorders and Down Syndrome]. *Teoreticheskaya i prikladnaya lingvistika* [*Theoretical and Applied Linguistics*], 2018, vol. 4, no. 2, pp. 22–37. (in Russ.)

Grossman R. B., Edelson L. R., Tager-Flusberg H. Emotional facial and vocal expressions during story retelling by children and adolescents with high-functioning autism. *Journal of Speech, Language, and Hearing Research*, 2013, vol. 56 (3), pp. 1035–1044.

Hessling A, Brimo D. M. Spoken fictional narrative and literacy skills of children with Down syndrome. *Journal of Communication Disorders*, 2019, no. 79, pp. 76–89. DOI 10.1016/j.jcomdis.2019.03.005

Kanner L. Autistic disturbances of affective contact. *Nervous Child*, 1943, vol. 2, pp. 217–250.

Kumin L. Early communication skills for children with Down syndrome: A guide for parents and professionals. Bethesda, MD, Woodbine House, 2003, 391 p.

Lebedinsky V. V. Narusheniya psikhicheskogo razvitiya v detskom vozraste [Disorders of Mental Development in Childhood]. Moscow, Akademiya Publ., 2003, 140 p. (in Russ.)

Lyakso E., Frolova O. Speech features of typically developing children and children with autism spectrum disorders. In: Abstract book. BIT's 7th Annual World Congress of Neurotalk – 2016. Innovation of Neuroscience. 2016.

Lyakso E., Frolova O., Karpov A. A New Method for Collection and Annotation of Speech Data of Atypically Developing Children. In: International Conference on Sensor Networks and Signal Processing (SNSP), 2018, pp. 175–180. DOI 10.1109/SNSP.2018.00040

Lyakso E. E., Frolova O. V. Analiz tekstov rechi "vzroslyi – rebenok", "vzroslyi – vzroslyi" pri normal'nom i atipichnom razvitii informantov [Analysis of "Adult-Child" and "Adult-Adult" Speech Texts Produced by Informant with Typical and Atypical Development] *Teoreticheskaya i prikladnaya lingvistika* [*Theoretical and Applied Linguistics*], 2017, vol. 2, pp. 20–47. (in Russ.)

Lyakso E. E., Frolova O. V., Grechanyi S. V., Matveev Yu. N., Verkholyak O. V., Karpov A. A. Golosovoi portret rebenka s tipichnym i atipichnym razvitiem [Speech Profile of a Child with Typical and Atypical Development]. St. Petersburg, Izdatel'sko-poligraficheskaya assotsiatsiya vysshikh uchebnykh zavedenii, 2020, 204 p. (in Russ.)

Makhnytkina O. V., Grigorev A., Nikolaev A. Analysis of dialogues of typically developing children, children with Down syndrome and ASD using machine learning methods. *Lecture Notes in Computer Science*, 2021, vol. 12997, p. 397–406. DOI 10.1007/978-3-030-87802-3_36

Matveev Y., Matveev A., Frolova O., Lyakso E. Automatic Recognition of the Psychoneurological State of Children: Autism Spectrum Disorders, Down Syndrome, Typical Development. *Speech and Computer. SPECOM 2021. Lecture Notes in Computer Science*, 2021, vol. 12997, pp. 417–425. DOI 10.1007/978-3-030-87802-3_38

Mazzaggio G., Shield A. The Production of Pronouns and Verb Inflections by Italian Children with ASD: A New Dataset in a Null Subject Language. *Journal of Autism and Developmental Disorders*, 2020, vol. 50, pp. 1425–1433. DOI 10.1007/s10803-019-04349-7

Nicholas J. S., Charles J. M., Carpenter L. A., King L. B., Jenner W., Spratt E. G. Prevalence and characteristics of children with autism-spectrum disorders. *Annals of Epidemiology*, 2008, vol. 18 (2), pp. 130–136.

Nikolaev A. S., Frolova O. V., Gorodnyi V. A., Lyakso E. E. Kharakteristika otvetnykh replik detei 5–11 let s rasstroistvami autisticheskogo spektra v dialogakh so vzroslymi [Features of Responses of 5–11 years old Children with Autism Spectrum Disorders in Dialogues with Adults]. *Voprosy psikholingvistiki* [*Journal of Psycholinguistics*], 2019, vol. 4, no. 42, pp. 92–105. (in Russ.)

Penke M. Regular and irregular inflection in Down syndrome – New evidence from German. *Cortex*, 2019, vol. 116, pp. 192–208. DOI 10.1016/j.cortex.2018.08.010

Penke M. Verbal Agreement Inflection in German Children With Down Syndrome. *Journal of Speech, Language, and Hearing Research*, 2018, vol. 61 (9), pp. 2217–2234. DOI 10.1044/2018_JSLHR-L-17-0241

Pokorny F. B., Schuller B., Marschik P. B., Brueckner R., Nyström P., Cummins N., Bölte S., Einspieler C., Falck-Ytter T. Earlier Identification of Children with Autism Spectrum Disorder: An Automatic Vocalisation-Based Approach. *Interspeech*, 2017, pp. 309–313. DOI 10.21437/Interspeech.2017-1007

Schopler E., Reichler R. J., DeVellis R. F., Daly K. Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS). *Journal of Autism and Developmental Disorders*, 1980, March, no. 10 (1), pp. 91–103. DOI 10.1007/BF02408436

Tek S., Mesite L., Fein D., Naigles L. Longitudinal analyses of expressive language development reveal two distinct language profiles among young children with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 2014, vol. 44 (1), pp. 75–89.

Terzi A., Marinis T., Zafeiri A., Francis K. Subject and Object Pronouns in High-Functioning Children with ASD of a Null-Subject Language. *Frontiers in Psychology*, 2019, vol. 10, p. 1301. DOI 10.3389/fpsyg.2019.01301

Wing L. The definition and prevalence of autism: a Review. *European Child and Adolescent Psychiatry*, 1993, vol. 2 (1), pp. 61–74.

Информация об авторах

Олеся Владимировна Махныткина, кандидат технических наук
Scopus Author ID 57208002090
WoS Researcher ID F-2283-2017
SPIN 1723-5004

Ольга Владимировна Фролова, кандидат биологических наук
Scopus Author ID 8521676200
WoS Researcher ID G-2649-2015
SPIN 4811-1118
Елена Евгеньевна Ляксо, доктор биологических наук, профессор
Scopus Author ID 24468656100
WoS Researcher ID H-9904-2013
SPIN 8669-2483

Information about the Authors

Olesia V. Makhnytkina, Candidate of Sciences (Engineering)
Scopus Author ID 57208002090
WoS Researcher ID F-2283-2017
SPIN 1723-5004
Olga V. Frolova, Candidate of Sciences (Biology)
Scopus Author ID 8521676200
WoS Researcher ID G-2649-2015
SPIN 4811-1118
Elena E. Lyakso, Doctor of Sciences (Biology), Professor
Scopus Author ID 24468656100
WoS Researcher ID H-9904-2013
SPIN 8669-2483

Вклад авторов:

О. В. Махныткина – разработка алгоритма анализа морфологических и лексических особенностей речи с использованием методов машинного обучения, проведение экспериментов.
О. В. Фролова – сбор исходного материала, доработка текста.
Е. Е. Ляксо – разработка концепции исследования, сбор оригинального материала.

Contribution of the Authors:

Olesia V. Makhnytkina – design of an algorithm for analyzing morphological and lexical features of speech using machine learning methods, conducting computational experiments.
Olga V. Frolova – collection of source material, finalization of the text.
Elena E. Lyakso – development of the research concept, collection of original material.

Статья поступила в редколлегию 10.07.2022;
одобрена после рецензирования 05.04.2023; принята к публикации 07.04.2023
The article was submitted on 10.07.2022;
approved after reviewing on 05.04.2023; accepted for publication on 07.04.2023