OPEN ACCESS

[Научный диалог = Nauchnyi dialog = Nauchnyy dialog, 12(7), 2023]
[ISSN 2225-756X, eISSN 2227-1295]

Информация для цитирования:

Тельпов Р. Е. Типовые различия естественных и сгенерированных нейронной сетью текстов в квантитативном аспекте / Р. Е. Тельпов, С. В. Ларцина // Научный диалог. — 2023. — Т. 12. — № 7. — С. 47—65. — DOI: 10.24224/2227-1295-2023-12-7-47-65.

Telpov, R. E., Lartsina, S. V. (2023). Typological Differences of Natural and Neural Network-Generated Texts in a Quantitative Aspect. *Nauchnyi dialog, 12* (7): 47-65. DOI: 10.24224/2227-1295-2023-12-7-47-65. (In Russ.).

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

Журнал включен в Перечень ВАК

DOI: 10.24224/2227-1295-2023-12-7-47-65

Типовые различия естественных и сгенерированных нейронной сетью текстов в квантитативном аспекте

Тельпов Роман Евгеньевич *
orcid.org/0000-0001-9130-2941
кандидат филологических наук, доцент кафедры общего и русского языкознания,
**** корреспондирующий автор***
roman-tel-pov@yandex.ru

Ларцина Станислава Витальевна
orcid.org/0009-0008-8347-9710
магистрант, кафедра общего и русского языкознания
stasya-200@list.ru

Государственный институт русского языка им. А. С. Пушкина
(Москва, Россия)

Typological Differences of Natural and Neural Network-Generated Texts in a Quantitative Aspect

Roman E. Telpov *
orcid.org/0000-0001-9130-2941
PhD in Philology, Associate Professor, Department of General and Russian Linguistics,
**** Corresponding author***
roman-tel-pov@yandex.ru

Stanislava V. Lartsina
orcid.org/0009-0008-8347-9710
Master's degree student, Department of General and Russian Linguistics
stasya-200@list.ru

Pushkin State Russian Language Institute
(Moscow, Russia)

ОРИГИНАЛЬНЫЕ СТАТЬИ

Аннотация:

Авторы статьи выявляют отличительные черты в текстах, написанных людьми, и в текстах, созданных нейросетью GPT-3. Тексты, сгенерированные GPT-3, еще не становились предметом систематического углубленного изучения. Рассмотрено 160 текстов, распределенных по четырем темам («Высшее образование в моих глазах», «Как оставаться человеком в нечеловеческих условиях?», «Как я провёл лето?», «Педагог года»), 80 из которых созданы нейросетью, а 80 — людьми. Тексты проанализированы с использованием методов квантитативной лингвистики. К каждому из текстов при помощи программы AntConc составлен конкорданс, из которого были получены количественные значения, используемые для дальнейшего анализа. Сделаны следующие выводы: (1) в сгенерированных текстах слова, включённые в заголовок, встречаются с наибольшей частотностью; (2) относительная частота употребления слов, включённых в заголовок, нецелесообразно завышена; (3) в список 20-ти самых частотных слов во всех сгенерированных текстах входит наибольшее количество полнозначных слов; (4) коэффициент лексического разнообразия в естественных текстах значительно выше, нежели у сгенерированных. Результаты исследования могут быть полезны как преподавателям, так и специалистам в области машинного обучения.

Ключевые слова:
искусственный интеллект; чат-бот; нейросеть; квантитативная лингвистика; сгенерированный текст; лемма; конкорданс; коэффициент лексического разнообразия.

ORIGINAL ARTICLES

Abstract:

The authors of this article identify distinctive features in texts written by humans and texts generated by the GPT-3 neural network. Texts generated by GPT-3 have not yet been subject to systematic in-depth study. In total, 160 texts were analyzed in the article, distributed across four topics (“Higher Education in My Eyes,” “How to Remain Human in Inhuman Conditions,” “How I Spent the Summer,” “Teacher of the Year”), with 80 texts generated by the neural network and 80 texts written by humans. The texts were analyzed using quantitative linguistic methods. A concordance was compiled for each text using the AntConc program, from which quantitative values were obtained for further analysis. The authors reached the following conclusions: (1) in the generated texts, words included in the title occur with the highest frequency; (2) the relative frequency of words included in the title is unreasonably inflated; (3) the list of the 20 most frequent words in all generated texts includes the highest number of full-fledged words; (4) the lexical diversity coefficient in the examined natural texts is significantly higher than that of the generated texts. The findings of this research can be useful for both educators and machine learning specialists.

Key words:
artificial intelligence; chatbot; neural network; quantitative linguistics; generated text; lemma; concordance; lexical diversity coefficient.

OPEN ACCESS
[Научный диалог = Nauchnyi dialog = Nauchnyy dialog, 12(7), 2023]
[ISSN 2225-756X, eISSN 2227-1295]

УДК 81’42:004.89

Научная специальность ВАК
5.9.8. Теоретическая, прикладная и
сравнительно-сопоставительная лингвистика

Типовые различия естественных и сгенерированных нейронной сетью текстов в квантитативном аспекте

1. Введение = Introduction

Сегодня стремительно развиваются компьютерные технологии, и связано это, прежде всего, с разработкой новейших нейронных сетей. Одно из определений этого понятия может звучать следующим образом: ***нейронные сети*** — это «искусственные, многослойные, высокопараллельные логические структуры, составленные из формальных нейронов» [Галушкин, 2023], где ***нейрон*** — это «вычислительный элемент, каждый из которых имеет несколько входов-синапсов и один выход-аксон» [Бурнашев и др., с. 259]. Нейронные сети активно применяются для решения широкого ряда лингвистических задач: анализ настроений (Sentiment Analysis), распознавание речи (Speech Recognition), информационный поиск (Data Searching), машинный перевод (Machine Translation) и иные, объединяемые междисциплинарным направлением, интегрирующим достижения машинного обучения и лингвистики, — ***Natural Language Processing*** (NLP).

Одна из областей приложения NLP — генерация текста, и с этой задачей на данный момент лучше всего справляется разработанная американской компанией OpenAI нейросеть ChatGPT, стабильная версия которой была опубликована в открытом доступе 23 марта 2023 года. Полное название — Generative Pre-trained Transformer — означает, что нейросеть состоит из трансформаторов нескольких уровней, которые обрабатывают естественные тексты и генерируют на их основе когерентные выходные данные. Для непрерывного обучения и развития ChatGPT использует «огромные массивы текстовых данных» [Dinesh et al, p. 828], находящихся в свободном доступе в сети Интернет, а также данные, получаемые в ходе взаимодействия с пользователем. Версия ChatGPT 3 имеет доступ только к информации до 2021 года, но, по словам разработчиков, следующая версия Chat GPT 4, которая на данный момент недоступна для бесплатного пользования, сможет обращаться к данным из сети Интернет напрямую, тем самым актуальность генерируемых нейронной сетью текстов в значительной мере возрастёт.

Благодаря постоянному обучению, ChatGPT может изображать чью-либо манеру письма (в ответ на запросы от пользователей «перепиши

текст на древнерусском», «напиши это же, но понятнее», «представь это в стихах» нейросеть обрабатывает входные данные и изменяет их с учётом запроса). Тем не менее истинность выдаваемых результатов всегда следует ставить под сомнение и в дальнейшим перепроверять. Сегодня ChatGPT позволяет генерировать уникальные тексты по запросу в кратчайшие сроки — от нескольких секунд до нескольких минут, — из-за чего возможно широкое распространение в сети Интернет речевых произведений, созданных с помощью нейросети. Вследствие этого перед обществом встала проблема разграничения сгенерированных текстов и естественных.

На момент проведения исследования поискам типовых различий в естественных и сгенерированных текстах посвящено незначительное количество зарубежных работ. Это труды А. Кохен, Р. Мантенья и Ш. Хавлина [Cohen et al, 2022], которые также производили статистический анализ, но при анализе частотности опирались на *закон Ципфа* и *энтропию Шеннона*. Они пришли к выводу, что для более успешного различения авторства ИИ и человека необходимо использовать обратный закон Ципфа. В русскоязычном научном пространстве также присутствуют исследования на эту тему, представленные в статье-обзоре А. Ю. Красноярова, М. А. Аргузовой, Ж. А. Хужамурадова и С. Р. Рахимова [Речевое творчество …, 2022], но в них не указаны какие-либо статистические данные. При этом не обнаружены работы, в которых были бы рассмотрены тексты, сгенерированные ChatGPT 3. Вследствие революционной производительности чат-бота возможно утверждать, что тексты, сгенерированные им по пользовательским запросам (чаще всего при работе с чат-ботами применяют лексему *промт*, от англ. *Promt* — запрашивать, запрос), на порядок качественнее текстов, которые могли бы создать все предшествующие нейронные сети. Продолжение начатых научных изысканий на обозначенную тему обусловливает ***актуальность исследования***.

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

Цель настоящего исследования — с помощью разнообразных методов квантитативной лингвистики установить типовые отличия между текстами, созданными нейронной сетью, и текстами, сотворёнными людьми. В качестве ***материала*** исследования выступили 80 текстов, написанных людьми и размещённых в свободном доступе, на ряд тем (*«Высшее образование в моих глазах»*, *«Как оставаться человеком в нечеловеческих условиях?»*, *«Как я провёл лето?»*, *«Эссе для профессионального конкурса «Педагог года»* — по 20 текстов на каждую тему), а также 80 текстов, сгенерированных с помощью чат-бота GPT 3 на аналогичные темы.

Так как дальнейший анализ предполагает работу с количественными данными разнообразных текстов, следует предварить основную часть статьи необходимым замечанием: все рассматриваемые в настоящем исследовании образцы состоят из 500 и более знаменательных слов. Советский лингвист Б. Н. Головин, являющийся первым теоретиком направления, рассматривающего язык с точки зрения статистики, анализировал в своих работах выборки (другое название — пробы) размером именно 500 и более знаменательных слов. Количество выборок каждого текста — 20, так как именно такой минимальный объём текстов позволяет получить надёжные результаты. При меньшем их количестве полученные данные могут быть нерелевантными («увеличение числа выборок… увеличит и надёжность результатов» [Головин, 1970, с. 65]).

Для автоматического сравнения текстов, созданных нейронной сетью, и текстов, написанных людьми, была использована программа для составления конкордансов — AntConc [Anthony, 2023]. Помимо возможности извлечь из текста всю употребляемую лексику и автоматически подсчитать частотность всех используемых слов, приложение также обладает рядом полезных для исследования инструментов. Например, во избежание случаев, когда словоформы одного слова — *стол, стола, столу* — понимаются автоматическим обработчиком как три разных лексемы, текст был предварительно лемматизирован с помощью морфологического анализатора MyStem (версия 3.1).

В целях осуществления квантитативного сравнительного анализа текстов, сгенерированных нейросетью и написанных людьми, были использованы следующие статистические меры:

1) TF-IDF (Term Frequency — Inverse Document Frequency)

Мера TF позволяет рассчитать периодичность использования какого-либо слова в тексте. Она рассчитывается по следующей формуле:

1) $\mathbf{TF(t,d) = \frac{n_t}{\Sigma_k n_k}}$, где:

n_t – сумма всех ключевых слов, частоту употребления которых мы вычисляем, а $\sum_k n_k$ — сумма слов во всех выборках. Вторую часть — IDF — применять для сравнения двух выборок в нашем случае нецелесообразно ввиду того, что значение IDF всегда будет равно нулю, так как формула сведётся к 0 в любом случае (это связано с тем, что все рассматриваемые ключевые слова наличествуют в каждом тексте):

2) $idf(t,D) = \log \frac{|D|}{|\{d_i \in D \mid t \in d_i\}|}$; $idf(t,D) = \log \frac{20}{20}$;

$idf(t,D) = \log 1 = 0$

Тогда при дальнейшем умножении TF и IDF, чего требует статистическая мера, произведение всегда будет равняться нулю. Это означает, что значимость рассматриваемого ряда ключевых слов, которые встречаются в каждом тексте выборки, для всей совокупности текстов в целом невелика.

2) TTR (Type-Token Ratio)

Стандартное значение TTR вычисляется согласно следующей формуле:

3) $TTR = \frac{Nlex}{N}$,

где *Nlex* — это количество уникальных лексем, а *N* — это общее количество всех словоформ в тексте.

С применением этой формулы связано множество дискуссий. Дело в том, что с увеличением количества употребляемых в тексте словоформ общая лексическая уникальность текста уменьшается. Многие исследователи предпринимали попытки расширить формулу, вводили дополнительные переменные и / или алгебраические операции (логарифмы, корни, возведение в квадрат), тем самым утяжеляя процесс вычисления. Одна из усовершенствованных вариаций расчёта TTR, при этом не требующая глубоких математических знаний, — это *индекс Мааса* (от англ. *Maas` index*), вычисляемый следующим образом:

4) $a^2 = \frac{\log N - \log V}{\log N}$,

где log N — это логарифм к общему количеству словоформ во всём тексте, log V — это логарифм к количеству уникальных лексем. Важная в рамках этого исследования особенность индекса Мааса заключается в том, что данный показатель «независим от длины текста для… письменных жанров» [Захарова, 2020, с. 25].

Для достоверности исследования мы применили ещё две наиболее известные и вызывающие одобрение у исследователей [McCarthy et al, 2007, 2010] формулы расчёта TTR, но теперь такие, которые вычисляются автоматически ввиду необходимости учёта большого количества переменных. Первая из них, к которой мы обратимся, — MTLD (от англ. *The Measure of textual lexical diversity*). Ещё одна метрика, которая была разработана в трудах современных исследователей, — vocd-D (от англ. *Vocabulary diversity*). Vocd-D особо чувствительна к размеру текста. В основе её работы лежит понятие гипергеометрического распределения (hypergeometric distribution) из области теории вероятности. В настоящем исследовании была использована работающая схожим образом метрика, которая компенсирует недостатки (чувстви-

тельность к размеру текста) Vocd-D, — HD-D (от англ. *Hypergeometric distribution D*).

Для выявления коэффициента лексического разнообразия с помощью мер MTLD и HD-D существует ряд свободно распространяемых библиотек для Python. Нами была выбрана библиотека lexical diversity 0.1.1, разработанная и опубликованная в 2020 году Кристофером Кайлом. Им также был рассмотрен параметр лексического разнообразия [Zenker, 2021].

3. Результаты и обсуждение = Results and Discussion

3.1. Распределение частотных слов в выборках естественных и сгенерированных текстов

Прежде чем сравнивать значения вышеупомянутых статистических мер (TF-IDF, TTR), необходимо обратить внимание на то, как именно распределены самые частотные слова в выборках естественных и сгенерированных текстов на каждую из тем.

С помощью программы AntConc были получены первые двадцать самых часто употребляемых слов для совокупности каждой группы текстов. Результаты представлены ниже (рис. 1, рис. 2, рис. 3, рис. 4).

	Type	Rank	Freq
1	и	1	689
2	в	2	473
3	образование	3	422
4	высокий	4	376
5	для	5	179
6	что	6	170
7	свой	7	162
8	это	8	153
9	на	9	148
10	мочь	10	136
11	который	11	133
12	не	12	132
13	я	13	130
14	возможность	14	129
15	жизнь	15	126
16	человек	16	117
17	знание	17	103
18	то	18	102
19	быть	19	94
20	получать	20	88

	Type	Rank	Freq
1	в	1	610
2	и	2	607
3	не	3	266
4	образование	4	256
5	на	5	239
6	что	6	226
7	я	7	197
8	высокий	8	186
9	быть	9	165
10	это	10	148
11	который	11	142
12	то	12	138
13	человек	13	128
14	с	14	116
15	свой	15	112
16	а	16	108
17	этот	17	106
18	как	18	103
19	для	19	101
20	но	20	99

Рис. 1. Самые частотные слова в сгенерированных (слева) и естественных (справа) текстах на тему «Высшее образование в моих глазах»

OPEN ACCESS [Научный диалог = Nauchnyi dialog = Nauchnyy dialog, 12(7), 2023]
[ISSN 2225-756X, eISSN 2227-1295]

	Type	Rank	Freq
1	и	1	671
2	в	2	565
3	мы	3	320
4	свой	4	312
5	человек	5	279
6	условие	6	240
7	сохранять	7	233
8	мочь	8	206
9	это	9	187
10	не	10	185
11	что	11	181
12	наш	12	138
13	быть	13	137
14	на	13	137
15	ситуация	15	134
16	который	16	131
17	то	17	129
18	человеческий	18	122
19	оставаться	19	118
20	человечность	20	111

	Type	Rank	Freq
1	в	1	644
2	и	2	630
3	человек	3	397
4	не	4	254
5	что	5	252
6	быть	6	249
7	жизнь	7	207
8	он	8	205
9	на	9	181
10	который	10	171
11	свой	11	166
12	с	12	136
13	к	13	135
14	это	14	128
15	как	15	127
16	мочь	16	122
17	они	17	121
18	то	18	106
19	этот	19	104
20	весь	20	98

Рис. 2. Самые частотные слова в сгенерированных (слева) и естественных (справа) текстах на тему «Как оставаться человеком в нечеловеческих условиях»

	Type	Rank	Freq
1	и	1	803
2	я	2	748
3	в	3	380
4	на	4	251
5	лето	5	221
6	быть	6	214
7	новый	7	185
8	с	8	156
9	свой	9	153
10	что	10	138
11	мы	11	133
12	проводить	12	131
13	время	13	122
14	мой	14	115
15	этот	15	105
16	это	16	99
17	не	17	97
18	для	18	94
19	много	19	93
20	который	20	83

	Type	Rank	Freq
1	и	1	752
2	в	2	709
3	мы	3	455
4	я	4	448
5	на	5	431
6	быть	6	300
7	с	7	264
8	не	8	263
9	что	9	219
10	этот	10	208
11	это	11	172
12	а	12	161
13	по	13	142
14	то	14	134
15	но	15	130
16	который	16	120
17	как	17	119
18	очень	17	119
19	лето	19	116
20	у	20	107

Рис. 3. Самые частотные слова в сгенерированных (слева) и естественных (справа) текстах на тему «Как я провёл лето»

OPEN ACCESS [Научный диалог = Nauchnyi dialog = Nauchnyy dialog, 12(7), 2023]
[ISSN 2225-756X, eISSN 2227-1295]

	Type	Rank	Freq
1	и	1	715
2	я	2	362
3	в	3	321
4	свой	4	262
5	ученик	5	231
6	что	6	228
7	педагог	7	207
8	быть	8	206
9	учитель	9	196
10	год	10	183
11	это	11	176
12	не	12	153
13	работа	13	136
14	который	14	128
15	наш	15	127
16	для	16	124
17	каждый	17	115
18	к	18	112
19	на	18	112
20	но	20	111

	Type	Rank	Freq
1	и	1	949
2	в	2	672
3	я	3	659
4	учитель	4	338
5	не	5	334
6	свой	6	325
7	что	7	323
8	быть	8	294
9	на	9	284
10	ребенок	10	263
11	с	11	238
12	мой	12	235
13	это	13	232
14	они	14	211
15	к	15	184
16	как	16	180
17	а	17	175
18	то	18	172
19	он	19	165
20	ученик	20	157

Рис. 4. Самые частотные слова в сгенерированных (слева) и естественных (справа) текстах на тему «Педагог года»

При анализе обращать внимание необходимо именно на знаменательные слова, так как во всех текстах служебные слова обладают наибольшими показателями частотности, но при этом они не несут никакого самостоятельного значения и не влияют на результаты анализа. Как было упомянуто ранее, Б. Н. Головин при статистических исследованиях использовал выборки именно знаменательных слов [Головин, 1970].

В совокупности текстов на тему «Высшее образование в моих глазах» (рис. 1) в сгенерированных текстах слова *образование* и *высокий* занимают позиции 3 и 4 соответственно, в естественных — 4 и 8. Слово *образование* в естественных текстах встречается 256 раз, а в сгенерированных — 422, что чуть более чем в полтора раза больше (~ 1,65). *Высокий* встречается в текстах нейронной сети 376 раз, а в написанных людьми — всего лишь 186, что уже в два раза меньше.

В сгенерированных текстах на тему «Как оставаться человеком в нечеловеческих условиях» в двадцать наиболее частотных слов входят следующие единицы, включённые в заголовок: *человек* (позиция 5), *условие* (позиция 6), *оставаться* (позиция 19). В естественных текстах в число двадцати наиболее употребительных слов вошло единственное заголовочное слово — *человек*, которое занимает позицию под номером 3, что даже выше, чем в созданных нейросетью речевых произведениях.

В текстах на тему «Как я провёл лето» включённая в заголовок лексическая единица *лето* занимает позицию 5 в сгенерированных текстах и 19 — в естественных. При этом слово *проводить* в написанных людьми текстах встречается гораздо реже и не входит в список двадцати самых употребительных слов, а в текстах ChatGPT занимает 12-ю позицию по количеству словоупотреблений.

Примечательна последняя выборка текстов на тему «Педагог года». Оба слова, входящих в заголовок, входят в список двадцати наиболее употребительных — *педагог* (позиция 7) и *год* (10), при этом в естественных текстах они встречаются очень редко. В отличие от рассмотренного ранее, здесь в список самых частотных слов в текстах нейронной сети включён *ученик* (позиция 5) и синоним *педагога* — *учитель* (позиция 9). В естественных текстах указанные слова занимают иные позиции (*учитель* — 4, *ученик* — 20). На 10-й позиции находится слово *ребёнок*, которое является контекстуальным синонимом к слову *ученик*.

Из всех рассмотренных случаев только в двух группах текстов наблюдается высокая частотность употребления слов, включённых в заголовок, в естественных текстах в сравнении со сгенерированными. В текстах на тему «Как оставаться человеком в нечеловеческих условиях» слово *человек* (397 в естественных текстах против 279 в сгенерированных), а также слова *учитель* (употребляется чаще, чем аналогичное слово в текстах ChatGPT, а также синоним *педагог*) и *ребёнок* (употребляется чаще, чем контекстуальный синоним *ученик*, но само слово *ученик* в естественных текстах занимает последнюю позицию). Подобные исключения объясняются тем, что авторы используют различные средства художественной выразительности. В частности, используются намеренные повторы, а также риторические вопросы, восклицания, намеренное использование коротких предложений, ярко выделяющихся на фоне длинных. К примеру, отрывок из эссе на тему «Педагог года» выглядит следующим образом:

«*Шагнув вперёд, я размышляю: Счастливый ли я человек? Я* ***учитель****!* ***Учитель*** *— необыкновенный человек, в котором должны присутствовать сразу два противоположных качества: доброта и строгость. В их разумном сочетании и заключается мудрость* ***учителя****.* ***Учитель*** *учит детей, а дети, в свою очередь, —* ***учителя***» (эссе «Я — Учитель — логопед!», Цыбенова Туяна Валерьевна).

В данном контексте слово *учитель* употребляется 5 раз в одном абзаце, чтобы подчеркнуть индивидуальную значимость профессии для педагога. Рассмотрим пример того, как в тексте ChatGPT употребляется наиболее частотное полнозначное слово *ученик*.

«*Мы должны быть гордыми учителями, любящими свою работу и помнящими о своей ответственности перед* ***учениками*** *и обществом. Стать "Педагогом года" — это значит, быть лучшим примером для коллег, для* ***учеников*** *и для общества в целом. Но это также означает, что мы должны постоянно развиваться, улучшаться и творить что-то новое, что было бы полезно и интересно для наших* ***учеников***».

В абзаце слово *ученик* употреблено три раза, но при этом в приведённом отрывке отсутствуют какие-либо заметные средства художественной выразительности. Подобные повторы означают, что нейронная сесть на текущем этапе развития испытывает затруднения, приводящие к тавтологии.

При анализе четырёх групп текстов из двадцати наиболее частотных слов, мы обращали внимание на полнозначные слова, включённые в заголовок текстов. Если перечислить все такие слова, то можно получить следующий список:

Сгенерированные тексты:

1. Высшее образование в моих глазах.

Образование, высокий, мочь, возможность, жизнь, человек, знание, быть, получать.

2. Как оставаться человеком в нечеловеческих условиях?

Человек, условие, сохранять, мочь, быть, ситуация, человеческий, оставаться, человечность.

3. Как я провёл лето?

Лето, быть, новый, проводить, время, много.

4. Педагог года.

Ученик, педагог, быть, учитель, год, работа, каждый.

Естественные тексты:

1. Высшее образование в моих глазах.

Образование, высокий, быть, человек.

2. Как оставаться человеком в нечеловеческих условиях?

Человек, быть, жизнь, мочь.

3. Как я провёл лето?

Быть, очень, лето.

4. Педагог года.

Учитель, быть, ребёнок, ученик.

На основе рассмотренного выше можно прийти к следующему выводу: и в естественных, и в сгенерированных текстах включённые в заголовок полнозначные слова встречаются с преобладающей частотностью. При этом в большинстве случаев в частотном списке присутствует разное количество слов с самостоятельным значением. В сгенерированных тек-

стах перечень подобных лексем приблизительно в два раза выше, нежели в естественных.

Для более точной оценки количественного варьирования в употреблении включённых в заголовок слов в настоящем исследовании была использована статистическая мера TF-IDF.

3.2. Статистическая оценка количественного варьирования в употреблении включённых в заголовок слов

В предыдущем параграфе внимание было уделено преимущественно тому, какие слова содержатся в списке двадцати наиболее частотных слов. Было упомянуто, что в сгенерированных текстах ряд лексем употреблён чаще (например, слово *образование* встречается 422 раза в текстах нейросети и 256 — в авторских), но подобные наблюдения позволяют судить лишь о том, какие слова встречаются чаще, а какие — реже.

Для того чтобы проанализировать количественное расхождение в употреблении слов, включённых в заголовок, мы обратились к статистической мере TF-IDF, которая позволила математически оценить наблюдаемую разницу. Были рассмотрены слова, включённые в заголовок, которые можно назвать ключевыми, так как они полностью отражают тему представленных текстов.

Статистическую меру TF обычно употребляют вместе с IDF (полное название формулы — TF-IDF (Term Frequency — Inverse Document Frequency), подробнее данную величину рассматривает исследователь Шахзад Кайзер [Qaiser, 2018]. TF-IDF позволяет оценить релевантность слов, включённых в заголовок, как в одном конкретном тексте, так и в корпусе текстов. TF-IDF также используется для решения проблемы автоматической классификации текстов по ключевым словам и для оценки значимости ряда слов для выборки текстов. В соответствии с замечаниями, сделанными во втором параграфе статьи, нами была использована только первая часть этой статистической меры — TF.

Воспользовавшись только мерой TF, мы получили ***относительную частоту*** употребления рассматриваемого слова во всей совокупности выборок. С помощью этой величины мы можем понять, какую долю в текстах занимает включённое в заголовок слово от количества всех лексем вообще.

Посредством меры TF была рассчитана относительная частота употребления следующих слов:

1. «Высшее образование в моих глазах» — *образование, высокий, глаз*.

2. «Как оставаться человеком в нечеловеческих условиях?» — *человек, оставаться, условие.*

3. «Как я провёл лето?» — *лето, проводить*.

4. «Педагог года» — *педагог, год*.

Таким образом, была получена следующая таблица, содержащая относительную частоту употребления ключевых слов (табл. 1).

Таблица 1

Относительная частота употребления включённых в заголовок полнозначных слов в выборке текстов на каждую из тем

Ключевое слово	**Частотность в текстах, сгенерированных Chat GPT**	**Частотность в текстах, написанных человеком**
Образование	0.036	0.016
Высокий	0.032	0.012
Глаз	0.002	0.003
Человек	0.022	0.023
Условие	0.018	0.005
Оставаться	0.009	0.004
Лето	0.019	0.006
Проводить	0.011	0.001
Педагог	0.018	0.002
Год	0.016	0.003

Как видно из таблицы, в 9 из 10 случаев частотность употребления ключевых слов в текстах ИИ более чем в два раза превышает употребление этих же слов в естественных текстах. Тенденция не соблюдается в употреблении лексемы *человек* (в текстах Chat GPT слово употребляется реже).

Подобное наблюдение позволило выдвинуть предположение о том, что в большинстве случаев слова, включённые в заголовок, повторяются в искусственных текстах неоправданно чаще. При этом повышенная относительная частота употребления включённых в заголовок слов в естественных текстах может быть связана с применением людьми разнообразных стилистических средств. Например, с целью эмоционального воздействия (многократные повторы, намеренное заострение внимание читателя на каком-либо ключевом слове).

Заметно повышенная частотность ключевых слов в текстах, сгенерированных нейронной сетью, позволяет предположить, что Chat GPT на текущем этапе разработки не может производить тексты, которые содержат отступления в сторону от первоначального запроса пользователя. К примеру, нейронная сеть не понимает, что при запросе «Педагог года» результат генерации не должен непременно содержать слова, содержащиеся в запросе.

3.3. Квантитативный анализ лексического разнообразия сгенерированных и естественных текстов

В параграфе 3.1 было отмечено, что в списке двадцати наиболее употребительных слов в сгенерированных текстах приблизительно в два раза чаще встречаются полнозначные слова, нежели в текстах, написанными людьми. В связи с этим можно предположить, что сгенерированные тексты, обладают меньшей лексической самобытностью. С целью проведения таких вычислений существует ряд методов. В настоящем исследовании мы обратились к коэффициенту лексического разнообразия TTR.

С помощью базовой формулы расчёта TTR (3) и её усовершенствованной версии (4) были получены следующие результаты вычислений, произведённых вручную (табл. 2).

Таблица 2

Результаты вычисления лексического разнообразия текстов с помощью простейших вариаций TTR

Темы проанализированных текстов	TTR		Индекс Мааса	
	Chat GPT	Человек	Chat GPT	Человек
«Как я провёл лето?»	0.246	0.333	0.150	0.111
«Высшее образование в моих глазах»	0.219	0.337	0.162	0.113
«Как оставаться человеком в нечеловеческих условиях»	0.221	0.303	0.160	0.123
«Педагог года»	0.220	0.288	0.162	0.124

Для автоматического анализа уникальности лексического разнообразия естественных и сгенерированных текстов мы воспользовались ещё двумя формулами: MTLD и HD-D. Использование разных способов вычисления TTR должно помочь избавиться от возможных помех, связанных с отличающимися размерами рассматриваемых выборок. Были получены следующие результаты (табл. 3).

Из приведённых таблиц видно, что в большинстве случаев лексическое разнообразие естественных текстов превышает таковое у сгенерированных. Серьёзные отличия наблюдаются у TTR, рассчитанного с помощью Индекса Мааса, но это явление объясняется тем, что с ростом LD (лексическое разнообразие) Индекс Мааса уменьшается. Это означает, например, что в выборке текстов «Высшее образование в моих глазах» (0,162 > 0,133)

Таблица 3

Результаты вычисления лексического разнообразия текстов
с помощью методов, требующих автоматизации подсчёта TTR

Темы проанализированных текстов	MTLD		HD-D	
	Chat GPT	Человек	Chat GPT	Человек
«Как я провёл лето?»	87.963	89.712	0.826	0.856
«Высшее образование в моих глазах»	104.275	173.617	0.842	0.909
«Как оставаться человеком в нечеловеческих условиях»	91.718	135.187	0.848	0.894
«Педагог года»	97.273	139.880	0.838	0.880

коэффициент TTR естественных текстов гораздо выше, чем у созданных нейронной сетью.

4. Заключение = Conclusions

Обобщим наблюдения, полученные в ходе анализа выборок (общей суммой в 160 текстов), произведённого с применением компьютерных средств и статистических методов:

I. Современные нейронные сети способны сгенерировать тексты, которые действительно похожи на естественные речевые произведения, например, по частотности используемой лексики (*образование* и *высокий* в текстах «Высшее образование…»). Тем не менее наблюдаются некоторые расхождения. Например, в сгенерированных текстах слова, включённые в заголовок, встречаются значительно чаще, нежели в естественных текстах.

II. Относительная частота употребления слов, включённых в заголовок, в обоих текстах отличается: в выборке естественных текстов 9 из 10 посчитанных ключевых слов употребляются как минимум в два раза реже, чем в сгенерированных. Только в одной выборке («Как оставаться человеком в нечеловеческих условиях») употребительность ключевого слова в текстах, созданных нейронной сетью, ниже, чем в естественных (0,022 и 0,023). Отсюда следует, что в искусственных текстах относительная частота употребления ключевых слов нецелесообразно завышена. Редкое завышение количества их употребления в естественных текстах может быть обусловлено применением различных средств художественной выразительности, например лексических повторов.

III. Отмечено, что в представленных выборках текстов на 4 разные темы в список 20-ти наиболее частотных слов входит разное количество полнозначных слов. В текстах, созданных ChatGPT, в частотных списках приблизительно в два раза чаще встречаются слова с самостоятельным значением. Это подвело к необходимости исследовать коэффициент лексического разнообразия сгенерированных и естественных текстов, так как многочисленные повторы одних и тех же слов свидетельствуют о том, что при генерации текстов использовано меньше синонимов частотных слов.

IV. Коэффициент лексического разнообразия в рассмотренной выборке естественных текстов на четыре темы во всех случаях значительно выше, нежели у сгенерированных.

Результаты исследования могут быть полезны как людям, работающим в сфере образования и стремящимся избежать фальсификации выполняемых учащимися письменных работ, так и специалистам в области обработки естественного языка, которые используют выборки «человеческих» текстов для машинного обучения нейронных сетей. Использование сгенерированных текстов для указанной цели может ухудшить финальное речевое произведение, так как нейронная сеть на основе несовершенных искусственных текстов столкнётся с обработкой зачастую неестественных речевых оборотов. Исследования, направленные на поиск различий между сгенерированными и естественными текстами, позволят вычленить критерии отбора обучающего материала для нейросетей, чтобы повысить качество выходного продукта.

Сегодня нейронные сети значительно продвинулись в имитации естественной письменной речи, однако они всё ещё сохраняют ряд отличительных черт, которые позволяют предположить, кем был написан текст — человеком или искусственным интеллектом.

Заявленный вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации.	**Contribution of the authors:** the authors contributed equally to this article.
Авторы заявляют об отсутствии конфликта интересов.	**The authors declare no conflicts of interests.**

Источники и принятые сокращения

1. *Anthony L.* AntConc (Version 4.2.0) [Electronic resource] / L. Anthony. — Tokyo, Japan : Waseda University. — 2014. — Access mode : https://www.laurenceanthony.net/software.html (accessed 26.08.2023).

Литература

1. *Борунов, А. Б.* Разнообразие речи и методы его измерения в тексте (лингвостатистический подход) / А. Б. Борунов // Litera. — 2017. — № 4. — С. 81—86.

2. *Бурнашев Р. Ф.* Роль нейронных сетей в лингвистических исследованиях / Р. Ф Бурнашев, А. С. Аламова // Science and Education. — 2023. — № 3. — С. 258—269.

3. *Бурнашев Р. Ф.* Квантитативная лингвистика и искусственный интеллект / Р. Ф. Бурнашев, А. С. Аламова // Science and Education. — 2022. — Т. 3, № 2. — С. 1390—1402.

4. *Галушкин А. И.* Нейронные сети [Электронный ресурс] / А. И. Галушкин // Большая российская энциклопедия. — 2022. — 16 ноября. — Режим доступа : https://old.bigenc.ru/technology_and_technique/text/4114009 (дата обращения: 20.06.2023).

5. *Головин Б. Н.* Язык и статистика / Б. Н. Головин. — Москва : Просвещение, 1971. — 190 с.

6. *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.

7 *Насырова Г. Н.* Обзор современных сервисов и программного обеспечения квантитативной лингвистики / Г. Н. Насырова, Ш. Х. Амонова, Р. Ф. Бурнашев // Science and Education. — 2022. — Т. 3, №. 12. — С. 450—462.

8. «*Речевое* творчество» искусственного интеллекта : какие тексты пишет машина и чем они отличаются от людских / А. Ю. Краснояров, М. А. Аргузова, Ж. А. Хужамурадов, С. Р. Рахимов // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Серия 6: Языкознание. Реферативный журнал. — 2022. — № 2. — С. 41—49. — DOI: 10.31249/ling/2022.02.02.

9. *Юхан Т.* Проблемы и методы квантитативно-системного исследования лексики / Т. Юхан. — Таллин : Валгус, 1987. — 204 с.

10. *Cohen A.* Numerical Analysis of Word Frequencies in Artificial and Natural Language Texts / A. Cohen, R. Mantegna, S. Havlin // Fractals. — 2011. — Vol. 5, no. 01. — Pp. 1—19. — DOI: 10.1142/S0218348X97000103.

11. *Dale R.* GPT-3: What's it good for? / R. Dale // Natural Language Engineering. — 2021. — Vol. 27, no. 1. — Pp. 113—118. — DOI: 10.1017/S1351324920000601.

12. *Dinesh K.* Study and Analysis of Chat GPT and its Impact on Different Fields of Study / K. Dinesh, S. Nathan // International Journal of Innovative Science and Research Technology (IJISRT). — 2023. — Vol. 8, no. 3. — Pp. 827—833. — DOI: 10.5281/zenodo.7767675.

13. *Floridi L.* GPT-3: Its Nature, Scope, Limits, and Consequences / L. Floridi, M. Chiriatt // Minds and Machines. — 2020. — Vol. 30, no. 2. — Pp. 1—14. — DOI: 10.1007/s11023-020-09548-1.

14. *Kettunen K.* Can Type-Token Ratio be Used to Show Morphological Complexity of Languages? / K. Kettunen // Journal of Quantitative Linguistics. — 2014. — Vol. 21, no. 3. — Pp. 223—245. — DOI: 10.1080/09296174.2014.911506.

15. *Klee T.* Utterance length and lexical diversity in American and British–English speaking children: What is the evidence for a clinical marker of SLI? / T. Klee, W. J. Gavin, S. F. Stokes // Language Disorders From a Developmental Perspective. — New York, 2017. — Pp. 103—140. — DOI: 10.4324/9781315092041-4.

16. *McCarthy P. M.* Voc-D: a theoretical and empirical evaluation / P. M. McCarthy, S. Jarvis // Language Testing. — 2007. — Vol. 24, no. 4. — Pp. 459—488. — DOI: 10.1177/0265532207080767.

17. *McCarthy P. M.* MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment / P. M. McCarthy, S. Jarvis // Behavior Research Methods. — 2010. — Vol. 42, no. 2. — Pp. 381—392.

18. *Qaiser S.* Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents / S. Qaiser, R. Ali // International Journal of Computer Applications. — 2018. — 181 (1). — Pp. 25—29.

19. *Somers H. H.* Statistical methods in literary analysis / H. H. Somers // The Computer and Literary Style / J. Leeds (ed.). — Kent, OH : Kent State University. — 1966. — Pp. 128—140.

20. *Tweedie F. J.* How variable may a constant be? Measures of lexical richness in perspective / F. J. Tweedie, R. H. Baayen // Computers and the Humanities. — 1998. — Vol. 32. — Pp. 323—352.

21. *Zenker F.* Investigating minimum text lengths for lexical diversity indices / F. Zenker, K. Kyle // Assessing Writing. — 2021. — Vol. 47, no. 2. — DOI: 10.1016/j.asw.2020.100505.

Статья поступила в редакцию 21.06.2023,
одобрена после рецензирования 10.09.2023,
подготовлена к публикации 20.09.2023.

Material resources

Anthony, L. (2014.) *AntConc (Version 4.2.0)*. Tokyo, Japan: Waseda University. Available at: https://www.laurenceanthony.net/software.html (accessed 08.26.2023).

References

Borunov, A. B. (2017). Diversity of speech and methods of measuring it in text (linguostatistical approach). *Litera, 4:* 81—86. (In Russ.).

Burnashev, R. F., Alamova, A. S. (2022). Quantitative linguistics and artificial intelligence. *Science and Education, 3(2):* 1390—1402. (In Russ.).

Burnashev, R. F., Alamova, A. S. (2023). The role of neural networks in linguistic research. *Science and Education, 3:* 258—269. (In Russ.).

Cohen, A., Mantegna, R., Havlin, S. (2011). Numerical Analysis of Word Frequencies in Artificial and Natural Language Texts. *Fractals, 5 (1):* 1—19. DOI: 10.1142/S0218348X97000103.

Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering, 27 (1):* 113—118. DOI: 10.1017/S1351324920000601.

Dinesh, K., Nathan, S. (2023). Study and Analysis of Chat GPT and its Impact on Different Fields of Study. *International Journal of Innovative Science and Research Technology (IJISRT), 8 (3):* 827—833. DOI: 10.5281/zenodo.7767675.

Floridi, L., Chiriatt, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines, 30 (2):* 1—14. DOI: 10.1007/s11023-020-09548-1.

Galushkin, A. I. (2022). Neural networks. In: *Great Russian Encyclopedia, 16 november*. Available at: https://old.bigenc.ru/technology_and_technique/text/4114009 (accessed 06.20.2023). (In Russ.).

Golovin, B. N. (1971). *Language and statistics*. Moscow: Education. 190 p. (In Russ.).

Kettunen, K. (2014). Can Type-Token Ratio be Used to Show Morphological Complexity of Languages? *Journal of Quantitative Linguistics, 21(3):* 223—245. DOI: 10.1080/09296174.2014.911506.

Klee, T., Gavin, W. J., Stokes, S. F. (2017). Utterance length and lexical diversity in American and British–English speaking children: What is the evidence for a clinical marker

of SLI? In: *Language Disorders From a Developmental Perspective*. New York. 103—140. DOI: 10.4324/9781315092041-4.

Krasnoyarov, A. Yu., Arguzova, M. A., Khuzhamuradov, Zh. A., Rakhimov, S. R. (2022). "Speech creativity" of artificial intelligence: what texts a machine writes and how they differ from human ones. *Social and Humanitarian Sciences. Domestic and foreign literature. Episode 6: Linguistics. Abstract journal, 2:* 41—49. DOI: 10.31249/ling/2022.02.02. (In Russ.).

McCarthy, P. M., Jarvis, S. (2007). Voc-D: a theoretical and empirical evaluation. *Language Testing, 24 (4):* 459—488. DOI: 10.1177/0265532207080767.

McCarthy, P. M., Jarvis, S. (2010). MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42 (2):* 381—392.

Nasyrova, G. N., Amonova, Sh. Kh., Burnashev, R. F. (2022). Review of modern services and software of quantitative linguistics. *Science and Education, 3 (12):* 450—462. (In Russ.).

Qaiser, S., Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications, 181 (1):* 25—29.

Somers, H. H. (1966). Statistical methods in literary analysis. In: *The Computer and Literary Style*. Kent, OH: Kent State University. 128—140.

Tweedie, F. J., Baayen, R. H. (1998.).How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities, 32:* 323—352.

Yukhan, T. (1987). *Problems and methods of quantitative-systematic research of vocabulary*. Tallinn: Valgus. 204 p. (In Russ.).

Zakharova, E. Yu., Savina, O. Yu. (2020). Lexical diversity of text and ways of measuring it. *Bulletin of Tyumen State University. Humanities studies. Humanities, 6 (1):* 20—34. DOI: 10.21684/2411-197X-2020-6-1-20-34. (In Russ.).

Zenker, F., Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing, 47 (2)*. DOI: 10.1016/j.asw.2020.100505.

The article was submitted 21.06.2023;
approved after reviewing 10.09.2023;
accepted for publication 20.09.2023.