УДК 81’27

КОЭФФИЦИЕНТ ЛЕКСИЧЕСКОГО РАЗНООБРАЗИЯ, ОБЪЕМ И ЧАСТЕРЕЧНОЕ НАПОЛНЕНИЕ ТЕКСТОВ КАК ПОКАЗАТЕЛИ СОЦИОЛИНГВИСТИЧЕСКОГО ВАРЬИРОВАНИЯ

Екатерина Сергеевна Худякова
к. филол. н., доцент кафедры теоретического и прикладного языкознания
Пермский государственный национальный исследовательский университет
614990, Пермь, Букирева, 15. khudiakova.es@gmail.com

Степан Денисович Киселев
студент филологического факультета
Пермский государственный национальный исследовательский университет
614990, Пермь, Букирева, 15. kiselevstde@gmail.com

В исследовании проверяется, насколько три показателя: коэффициент лексического разнообразия (КЛР), объем текста в токенах (словоформах) и частота реализации разных частей речи, демонстрируют социолингвистическое варьирование в устных монологических текстах. На основе сбалансированной выборки авторов текстов (N=48) по факторам пол, возраст, специальность и уровень образования, рассчитаны статистические показатели различий выборок. В исследовании применялись исключительно машинные методы анализа с помощью скриптов на языке программирования Python. В результате показано, что на исследованном материале параметр «КЛР» демонстрирует различия только в зависимости от фактора «возраст»; объем также различает тексты информантов молодого и старшего возраста. Частоты реализаций частей речи значимо варьируют в текстах мужчин и женщин, а их качественные различия, вероятно, показывают стратегии текстопорождения.

Ключевые слова: социолингвистическое варьирование, пол, возраст, специальность, уровень образования, коэффициент лексического разнообразия, объем текста в словоформах, частота частей речи.

Введение

Исследователи устного дискурса С. Дюбуа и А. Санкофф отмечают многообразие единиц, которые могут демонстрировать варьирование: «Дискурсивные формы являются структурно различными и могут обнаруживаться на различных уровнях анализа: это могут быть сложные процессы (повествование, описание, рассуждение), крупные единицы (повторы, риторические вопросы, непрямая речь) или более ограниченные единицы (маркеры и частицы). Вероятно, отсутствие общепринятых единиц анализа устного текста также создает сложности в его исследовании»[1] [Dubois, Sankoff 2001: 283].

В связи с этим утверждением в статье поставлена проблема проверки того, насколько три из давно применяемых в социолингвистике показателей, а именно коэффициент лексического разнообразия (КЛР), объем текста в токенах (словоформах) и частота реализации разных частей речи, демонстрируют социолингвистическое варьирование в устных спонтанных текстах в зависимости от факторов пол (мужской, женский), возраст (25–34, 35–44, 45–54), специальность (гуманитарная, негуманитарная) и уровень образования (средний, высший). Выборка текстов сбалансирована по указанным факторам, т. к. в работах представителей Пермской школы социолингвистики показано, что именно они могут обуславливать варьирование различных лингвистических параметров (см., напр., [Ерофеева 2009: 34–35]).

В исследовании применялись машинные методы анализа с помощью скриптов на языке программирования Python: от парсинга коллекции текстов для создания сбалансированной выборки, частеречного теггирования и до проведения статистических тестов, указывающих на варьирование признаков по рассматриваемым факторам.

Материалом исследования выступили 48 текстов на тему «О себе», общее количество обработанных токенов составило 14477, лемм – 2677. Выборка текстов сбалансирована по 4 факторам: пол говорящего, специальность, уровень образования и возраст.

Определение рабочих понятий исследования

Коэффициент лексического разнообразия (КЛР, lexical diversity) – это количественная характери-

стика, отражающая вариативность лексики в тексте. Наиболее известным способом его измерения является Type-Token Ratio (TTR) – отношение количества уникальных слов (англ. *types*) к общему числу слов (англ. tokens) [Zenker, Kyle 2021: 1]. Лексическое разнообразие «можно описать как диапазон и вариативность словарного запаса, используемого в тексте говорящим или пишущим» [McCarthy, Jarvis 2007: 459].

Этот показатель нашел широкое применение в различных областях – от литературоведения до психиатрии и исследования детской речи [Covington, McFall 2010: 94]. Оценка лексического разнообразия может служить индикатором успешности речевого акта, а также применятся «как мера прогресса овладения иностранным языком» [Захарова, Савина 2020: 21]. Также его описывают как меру «гибкости или вариативности словарного запаса», призванную указывать на аспекты «языковой адекватности» (англ. language adequacy) [Hess et al.1984: 51].

Объем текста (в западной традиции длина текста) – общее количество токенов (словоформ) в тексте. Как указывают Ф. Зенкер и К. Кайл, этот параметр связан с коэффициентом лексического разнообразия [Zenker, Kyle 2021: 1], однако в устном тексте возможна ситуация, когда достаточно длинный текст характеризуется низким КЛР из-за общей клишированности устной речи.

Частеречное наполнение текстов – относительные частоты реализации частей речи, размеченных в терминах OpenCorpora (https://pymorphy2.readthedocs.io/en/stable/user/grammemes.html) в текстах определенной социальной группы информантов.

Применение показателей КЛР, частеречного наполнения и длины текстов в социолингвистических исследованиях

Многие исследователи отмечали, что лексическое разнообразие может быть связано с социальными характеристиками говорящего. Так, например, С. Джарвис отмечает, что «лексическое разнообразие влияет на восприятие слушателями достоверности, компетентности, привлекательности, социально-экономического статуса и коммуникативной эффективности говорящего» [Jarvis 2013: 96]. Автор представляет лексическое разнообразие текста как показатель, связывающий словарный запас (vocabulary knowledge), использование слов (word use) и владение языком (language proficiency) [там же: 103], что подчеркивает важность данного параметра в социолингвистическом контексте.

Отмечаются противоречивые результаты исследований гендерных различий в показателях лексического разнообразия текстов. Так, С. Сингх указывает, что мужская речь отличается большим лексическим разнообразием и более длинными фразами, тогда как речь женщин характеризуется более короткими синтаксическими конструкциями и большей повторяемостью слов. Однако автор подчеркивает, что эти результаты следует рассматривать с осторожностью из-за ограниченного числа участников исследования (17 женщин и 13 мужчин) [Singh 2001: 261].

В противоположность этому, в исследовании Г. Ю не было выявлено значимых различий между мужчинами и женщинами по значениям коэффициента лексического разнообразия [Yu 2010: 248].

Различия в результатах исследования КЛР в текстах мужчин и женщин можно объяснить типами исследуемого материала. С. Сингх исследовал записи устной спонтанной речи, тогда как Г. Ю исследовал письменные тексты, созданные в ходе экзамена.

В своей работе Г. Ю обращается к связи коэффициента лексического разнообразия и с рядом других социальных характеристик. Автор анализировал письменные тексты учащихся, изучающих английский язык как второй. Было установлено, что между двумя представленными этническими группами – филиппинцами и китайцами – не оказалось существенных различий в значениях коэффициента лексического разнообразия [там же: 248]. Однако для другого показателя – цели создания письменного текста – различия в коэффициенте лексического разнообразия оказались существенными. У участников, проходивших тестирование с целью профессиональной аттестации, КЛР оказался выше, чем в текстах тех, кто сдавал тест для поступления в учебное заведение [там же: 249]. Это можно рассматривать как проявление влияния коммуникативной ситуации на лексическое наполнение текстов.

Для письменных русскоязычных коротких текстов КЛР, по мнению Т.А. Литвиновой и ее коллег, выступает наиболее важным «предиктором пола» автора [Litvinova 2017: 70]. В статье КЛР интерпретируется как социально обусловленная характеристика речи: в письменных текстах мужчин КЛР имеет значительно большие значения, в сравнении с текстами женщин. Авторы объясняют это тем, что в письменных текстах мужчин используется меньше наиболее частотных слов, большинство из которых являются служебными словами (англ. *function words*) [Litvinova 2017: 71].

Многие исследователи отмечают взаимосвязь частотности реализации различных частей речи в тексте с внеязыковыми параметрами, в частности с социальными характеристиками автора. Так, в исследовании А.С. Беляевой и Е.В. Ерофеевой показано, что частоты употребления частей речи варьируют как в зависимости от темы текста, так и от пола говорящего. В монологах на тему «Пельмени» женщины использовали значительно больше глаголов, тогда как при теме «Работа» существенных различий между текстами мужчин и женщин не наблюдалось. Кроме того, мужчины чаще женщин в устной речи употребляли существительные (независимо от темы), тогда как наречия, напротив, чаще реализовывались женщинами [Беляева, Ерофеева 2020: 80–81].

Относительно фактора «возраст» выяснилось, что младшая группа (25–34 года) чаще других использует глаголы и наречия в монологах на тему «Пельмени», однако в монологах на тему «Работа» у этой группы наблюдается наименьшая частотность реализации этих частей речи [Беляева, Ерофеева 2020: 81–82]. Таким образом, эти данные демонстрируют сложную соотнесенность социальных характеристик информантов, темы текста и частоты употребления различных частей речи.

В письменных же текстах, как отмечают Т.А. Литвинова и коллеги, мужчины чаще реализуют существительные, предлоги и «модификаторы» (*pronoun-like adjectives*) [Litvinova 2017: 71]. В другой статье указано, что в письменных текстах женщин строевые слова (служебные слова и местоимения), личные местоимения и союзы используются чаще, чем в текстах, написанных мужчинами [Литвинова 2015: 104–105].

Наконец, в работе Н.В. Богдановой-Бегларян указано, что частоты существительных и глаголов мало варьируют в зависимости от пола и возраста говорящего [Богданова-Бегларян и др. 2017: электр. ресурс].

Параметр объема текста не часто становится предметом исследования, т. к. требует материала одного типа (жанра) и даже, желательно, созданного на одну тему. Тем не менее в статье А. Лиматта показано, что объем текста выступает различительным для регистра языка, понимаемого скорее как стиль, параметром [Liimatta 2023].

Обратим внимание, что пока наблюдаются противоречивые данные о варьировании показателей КЛР и частоты реализации частей речи, чувствительных и к форме реализации языка (устной или письменной), и к теме текста, и к цели его создания, поэтому проверка их варьирования на материале однотемных устных текстов, сбалансированных по четырем социальным параметрам говорящих, представляется актуальной.

Методика анализа данных

Поскольку в исследовании применялся исключительно компьютерный анализ данных с применением различных библиотек и модулей для ЯП Python, покажем методику как последовательность препроцессинга данных.

Все файлы текстов «О себе», собранные в Лаборатории социокогнитивной и компьютерной лингвистики ПГНИУ (около 150), получили единообразные названия. С помощью цикла токены метаданных помещены в словарь, где ключ – градации исследуемых социолингвистических признаков, а значение – список файлов, соответствующих этим характеристикам. По наименьшему количеству текстов в группе (2) была организована выборка, в которую вошли 48 текстов. Создан словарь частотных паралингвистических характеристик, используемых в транскриптах (напр., «смех», «вздох», «кашель»), с помощью регулярного выражения все эти единицы были удалены из текстов.

Токенизация текстов произведена с помощью модуля word_tokenize библиотеки NLTK (см. https://www.nltk.org); лемматизация и получение частеречной разметки осуществлены с помощью библиотеки Pymorphy3 (https://pypi.org/project/pymorphy3). Пример токенизации, лемматизации и расчета на их основе параметра КЛР см. на Рисунке 1.

```
# Функция для расчета КЛР. Принимает на вход текст, очищает его,
# лемматизирует токены и возвращает значение КЛР и список токенов
def calculate(defaulttext): 1 usage
    defaulttext = re.sub( pattern: r'[^\w\s]', repl: '', defaulttext)
    tokens = nltk.word_tokenize(defaulttext)
    if not tokens: return 0.0, []
    morph = pymorphy3.MorphAnalyzer()
    normalforms = [morph.parse(i)[0].normal_form for i in tokens]
    uniquewords_count = len(set(normalforms))
    klr = uniquewords_count / len(tokens)
    return round(klr, 2), tokens
```

Рисунок 1. Пример автоматической токенизации, лемматизации и расчета параметра КЛР

Метаданные и значение рассчитанных параметров КЛР и объема текстов внесены в pandas dataframe (https://pandas.pydata.org). Частоты частей речи в текстах конкретной группы рассчитаны с помощью стандартного метода Counter и сохранены как словарь, в котором ключ – тег части речи, значение – абсолютная частота. Данные по КЛР и объемам текстов сохранены как pandas dataframe и с помощью метода .values преобразованы в массив numpy array (https://numpy.org).

Теги частей речи, сохраненные как словарь в формате .json, загружены как массив numpy array, далее преобразованы в pandas dataframe. Мето-

дом .merged библиотеки Pandas выборки по градациям объединены в единую выборку, рассчитаны относительные частоты каждой части речи, далее эти данные подавались на вход в библиотеку Scipy для осуществления статистического анализа (https://scipy.org).

Определение статистических мер значимости параметров КЛР, объема и частеречного наполнения текстов

Поскольку выбор статистического метода оценки сходства выборок зависит от нормальности распределения, была проведена оценка нормальности распределения в каждой выборке данных с помощью критерия Шапиро-Уилка (метод .shapiro модуля stats в библиотеке Scipy). Тест был выбран как наиболее достоверный при разных объемах выборки в сравнении с критериями Колмогорова-Смирнова, Лиллиефорса и Андерсона-Дарлинга [Mohd Razali, Yap 2011].

Результаты расчета p по тестам Шапиро-Уилка представлены в Таблице 1.

Таблица 1

Результаты теста Шапиро-Уилка на нормальность распределения, p

Показатель	Жен.	Муж.	Сред.	Высш.	Негум.	Гум.	25–34	35–44	45–54
КЛР	**0,947**	**0,229**	0,0191	**0,167**	0,099	**0,159**	**0,589**	0,025	**0,807**
Объем	**0,312**	0,0036	0,919	0,0015	0,059	0,029	0,856	**0,35**	**0,54**
Частеречное распределение	**0,165**	**0,21**	**0,17**	**0,15**	**0,52**	**0,67**	**0,32**	**0,35**	**0,34**

Как видно из Таблицы 1, тест на нормальность распределения прошли данные по частоте частей речи, к ним далее применен параметрический метод сравнения, по показателям КЛР и длины хотя бы одна из сравниваемых выборок не отвечает нормальному распределению, к ним применен непараметрический критерий сравнения.

Для выборок с ненормальным распределением был проведен тест Манна-Уитни с помощью метода .mannwhitneyu модуля stats указанной библиотеки, коэффициент корреляции Стьюдента применялся к выборкам с данными по частотам реализации частей речи с нормальным распределением. Расчет произведен методом ttest_rel. Описательные статистики распределения значений параметров КЛР и длины текстов получены с помощью графического метода .boxplot библиотеки Seaborn.

Результаты тестов Манна-Уитни и Т-критерия Стьюдента представлены в Таблице 2.

Таблица 2

Меры сходства выборок, уровень значимости 0,05

Критерий	Показатель	Пол	Образование	Специальность	25–34/ 35–44	35–44/ 45–54	25–34/ 45–54
U-test, p	КЛР	0,502	0,239	0,355	0,067	0,850	**0,047**
	Объем	0,297	0,549	0,714	0,113	0,924	**0,048**
t-test, p	Части речи	**0,018**	0,91	0,863	0,247	0,742	0,407

Для показателя «КЛР», не имеющего нормального распределения, все результаты теста Манна-Уитни, кроме данных для младшей и старшей возрастных групп (значение p=0,047 при уровне значимости 0,05), показали отсутствие статистически значимых различий во всех группах.

По показателю «объем текста», также не прошедшего проверку на нормальность распределения, результаты теста Манна-Уитни тоже продемонстрировали статистически значимые различия по фактору «Возраст». Так, среднее значение длины текстов в группе информантов 25–34 лет – 222 словоупотребления, в возрастной группе 35–44 лет – 364, в группе 45–54 лет – 318,6. Статистически значимые различия обнаружены между объемами текстов младшей и старшей возрастных групп (значение p=0,048 при уровне значимости 0,05).

Распределение частот частей речи в текстах во всех выборках соответствует нормальному, поэтому для сравнения выборок применялся параметрический t-критерий Стьюдента. Статистически значимыми оказались различия в частотах частей речи только в текстах мужчин и женщин (p=0,0018 при 5%-ном уровне значимости).

Далее в Таблице 3 рассмотрим различия в реализации параметров КЛР и объема в зависимости от градаций социолингвистических факторов.

Таблица 3

Средние значения показателей по градациям факторов

Показатель	Жен.	Муж.	Сред.	Высш.	Негум.	Гум.	25–34	35–44	45–54
КЛР	0,552	0,583	0,561	0,584	0,566	0,579	0,61	0,553	0,551
Объем	325,7083	277,5	294,5	308,7083	310,333	292,87	222,0	364,06	318,6875

Как видно из Таблицы 3, по значениям КЛР мужчины создавали чуть более богатые тексты, чем женщины, информанты с высшим образованием в сравнении с информантами со средним, гуманитарии в сравнении с негуманитариями, молодые информанты в сравнении с говорящими средней и старшей возрастной групп также создавали тексты с несколько большим значением КЛР. Длина текста довольно значительно больше у женщин, менее заметно – у информантов с высшим образованием, негуманитариев, представителей средней возрастной группы. Самый низкий показатель длины текста – в группе говорящих 25–34 лет.

Как видно по графику «ящик с усами» (см. Рисунок 2), КЛР в текстах мужчин и женщин имеет близкую медиану, но у женщин межквартильный размах несколько больше, но имеются выбросы в выборке текстов мужчин.

Медианные значения в выборке информантов со средним и высшим образованием (см. Рисунок 3) разнятся, размах больше в группе текстов, полученных от информантов с высшим образованием, выбросы имеются в группе со средним образованием.

В выборке гуманитариев и негуманитариев (см. Рисунок 4) медианные значения достаточно близки, как и размах, разнятся лишь границы статистически значимой выборки (границы «усов»).

Среди возрастных групп (см. Рисунок 5) медианное значение средней и старшей групп близки, в средней возрастной группе наблюдается больший размах варьирования данных.

По параметру длины текста мужские и женские тексты отличаются (см. Рисунок 6): мужские тексты по медиане и по межквартильному размаху короче женских, в выборке текстов мужчин также присутствует выброс.

По фактору «уровень образования» (см. Рисунок 7) выборки различаются незначительно, выбросы наблюдаются в выборке информантов с высшим образованием, а самая низкая граница значений – в выборке текстов информантов со средним образованием.

В текстах негуманитариев и гуманитариев (см. Рисунок 8) медианное значение длины текста оказалось выше у негуманитариев, в этой же выборке есть выбросы.

По возрастному фактору (см. Рисунок 9) средняя и старшая группы показывают более схожие результаты расчетов длин текстов, значительно отличается от них группа текстов, полученная от информантов 25–34 лет, в ней присутствуют выбросы, но по статистически важному разбросу значений (размеры «усов») она самая компактная.

Наиболее существенные различия в частотах реализации частей речи в зависимости от фактора «пол» имеют существительные, числительные и предлоги, а также глаголы (их больше в текстах мужчин) и частицы, союзы, наречия, полные прилагательные и местоимения (их больше в текстах женщин) (см. Рисунок 10).

По фактору «образование» (см. Рисунок 11) некоторые различия в частотах частей речи показали предлоги и частицы (их больше в текстах информантов со средним образованием) и полные прилагательные и компаративы (их больше в текстах информантов с высшим образованием), но частоты остальных частей речи практически не варьируют в рассматриваемых группах.

Как видно из Рисунка 12, в текстах гуманитариев чаще используются существительные, компаративы и полные причастия, негуманитариев – наречия и полные прилагательные, частоты остальных частей речи довольно схожи.

На Рисунке 13 видно, что по использованию существительных близки младшая и старшая возрастные группы информантов, частоты глаголов последовательно возрастают с возрастом; союзов, наречий и местоимений-существительных больше в средней возрастной группе.

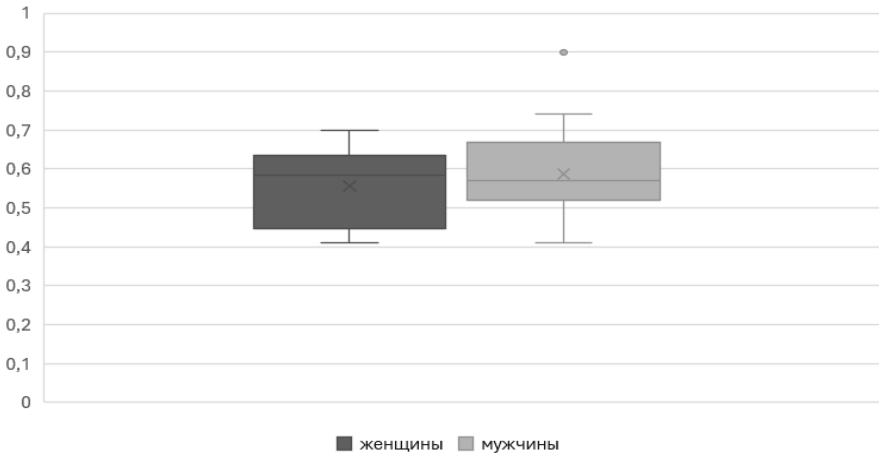

Рисунок 2. Распределение значений КЛР в зависимости от фактора «пол»

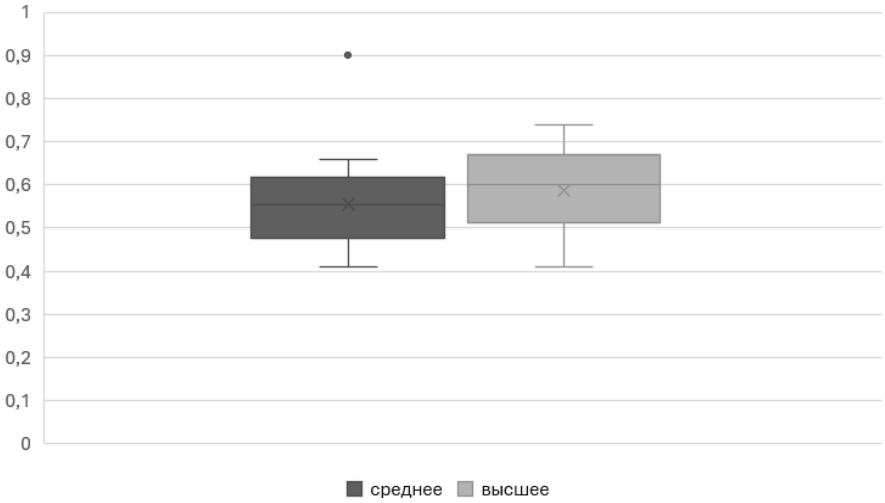

Рисунок 3. Распределение значений КЛР в зависимости от фактора «образование»

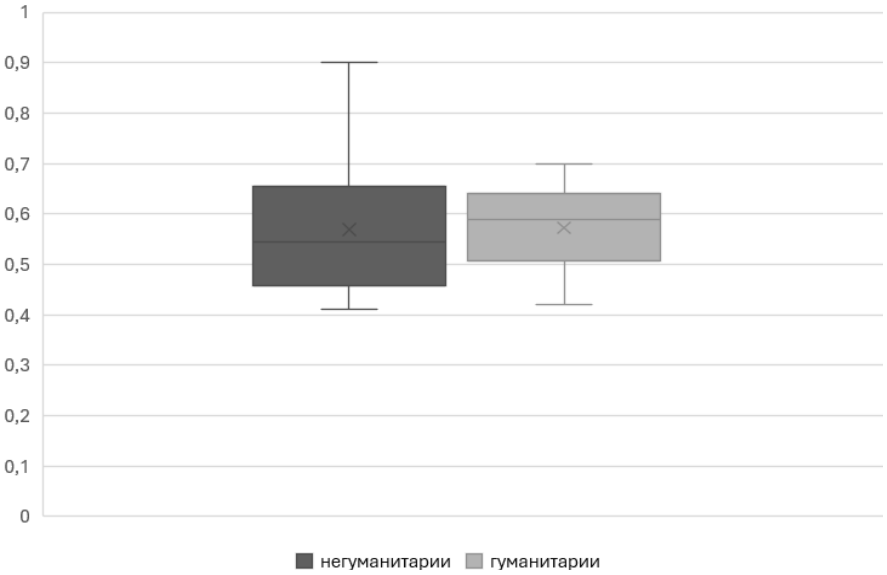

Рисунок 4. Распределение значений КЛР в зависимости от фактора «специальность»

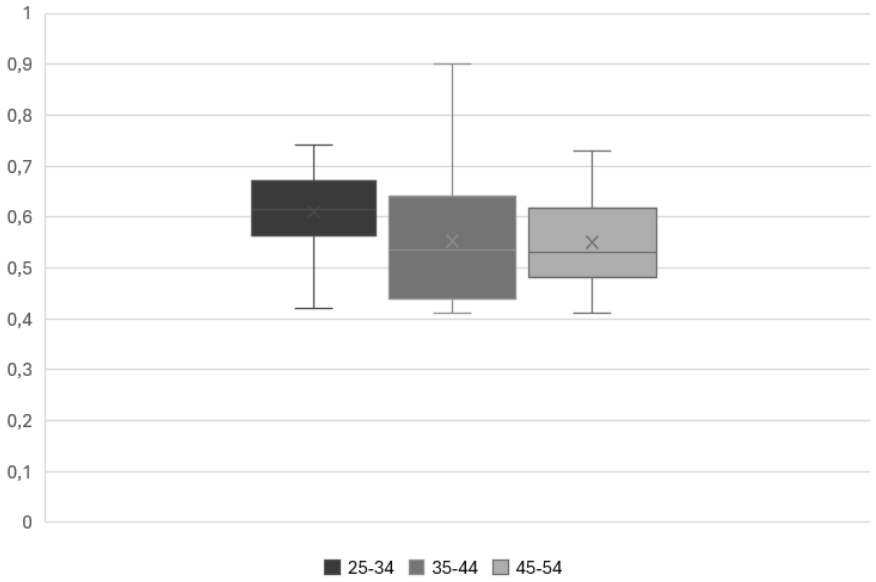

Рисунок 5. Распределение значений КЛР в зависимости от фактора «возраст»

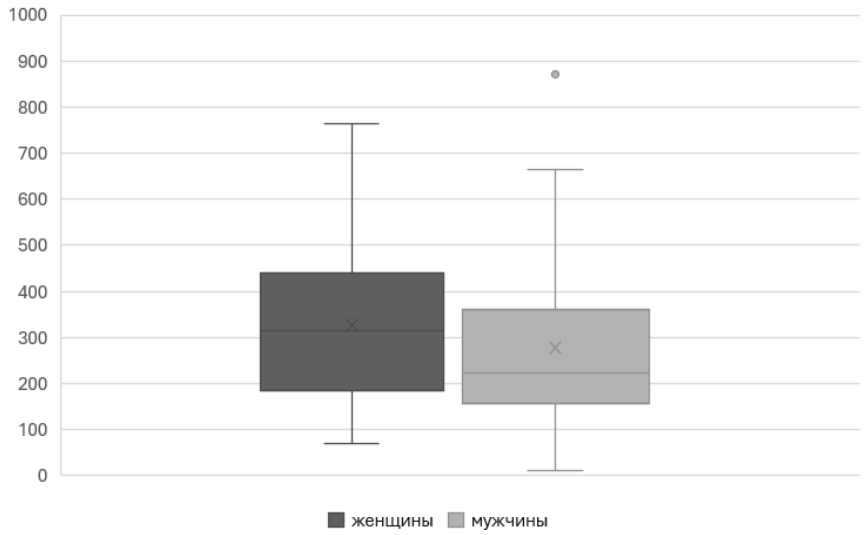

Рисунок 6. Распределение значений объема текста в зависимости от фактора «пол»

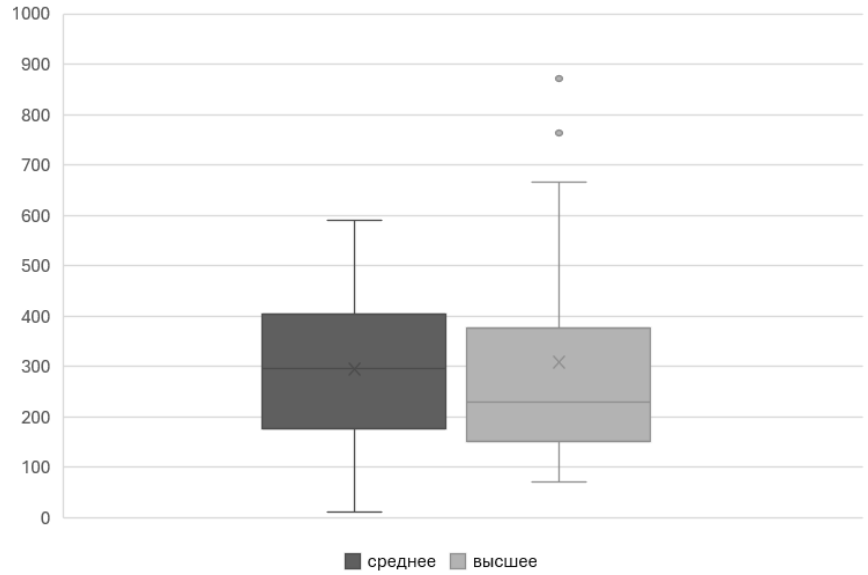

Рисунок 7. Распределение значений объема текста в зависимости от фактора «образование»

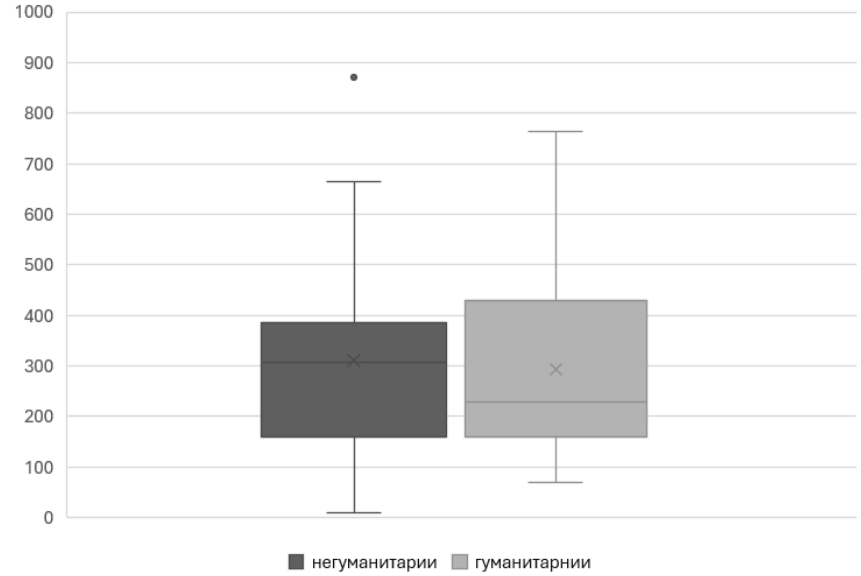

Рисунок 8. Распределение значений объема текста в зависимости от фактора «специальность»

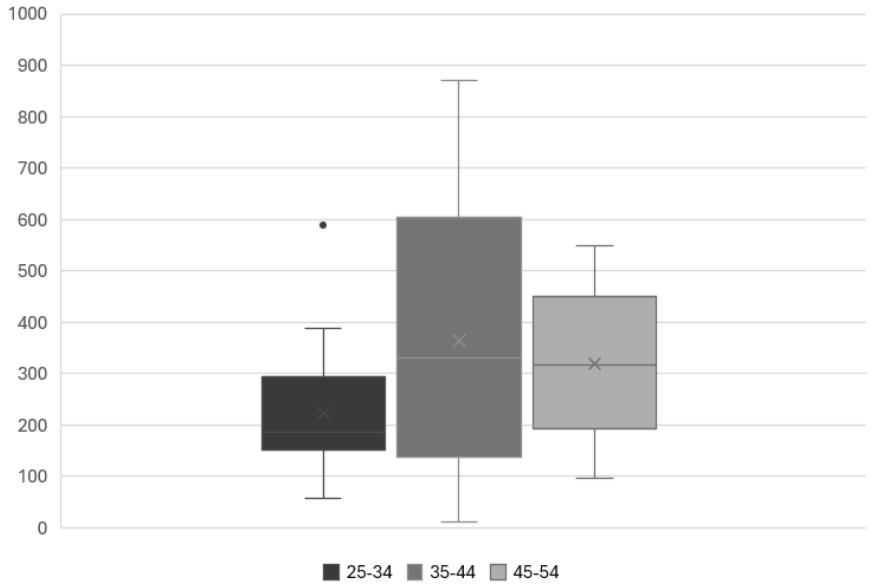

Рисунок 9. Распределение значений объема текста в зависимости от фактора «возраст»

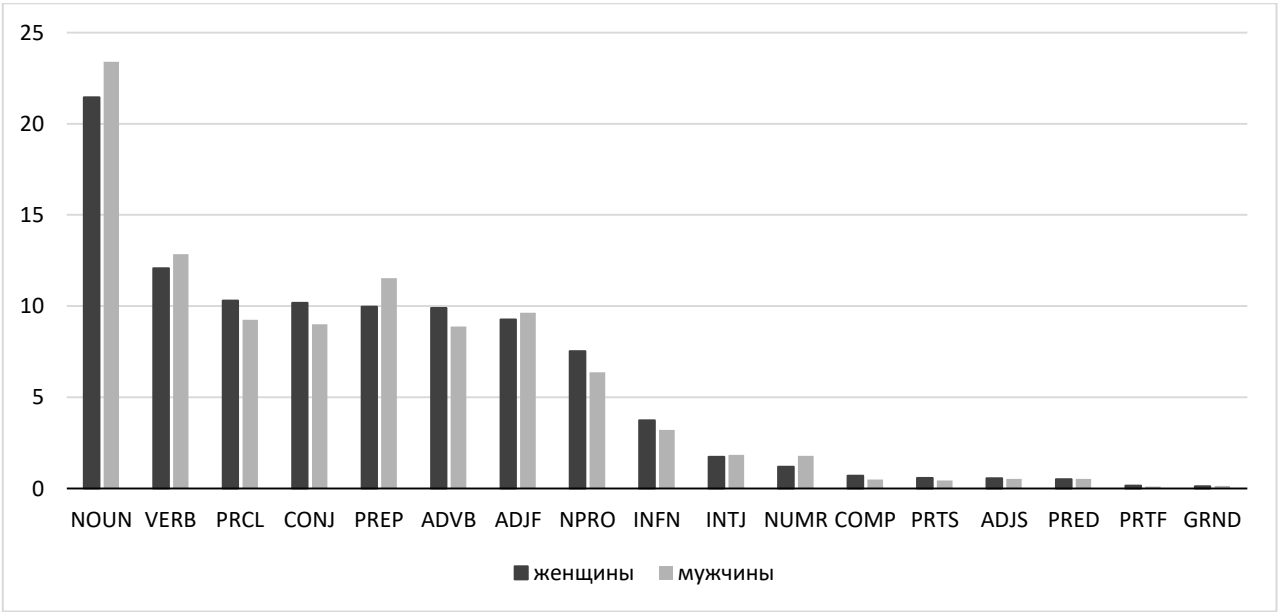

Рисунок 10. Частота реализации частей речи в зависимости от фактора «пол»

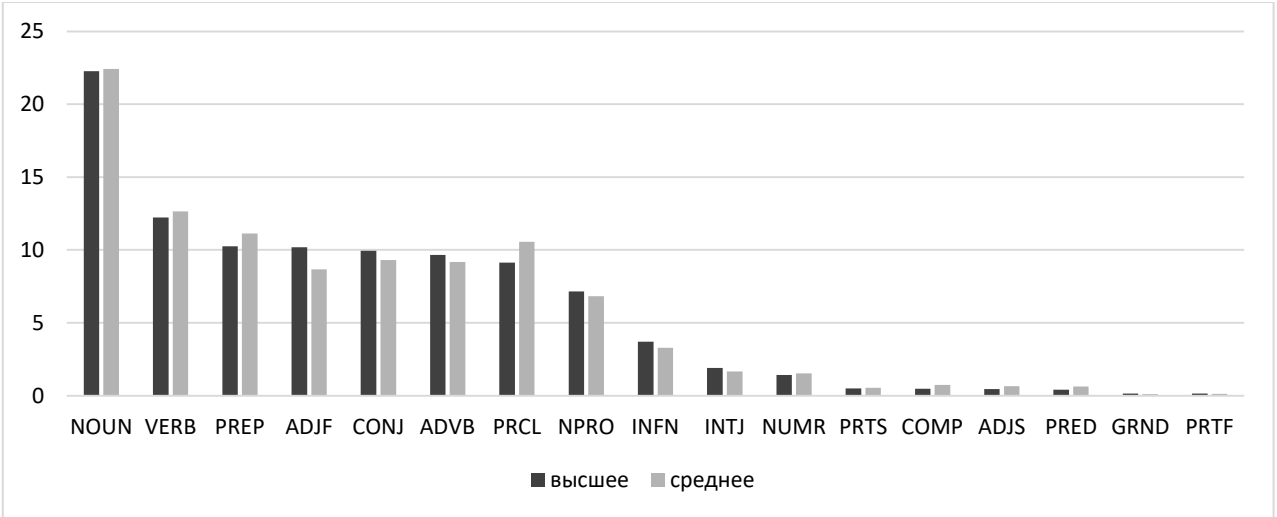

Рисунок 11. Частота реализации частей речи в зависимости от фактора «образование»

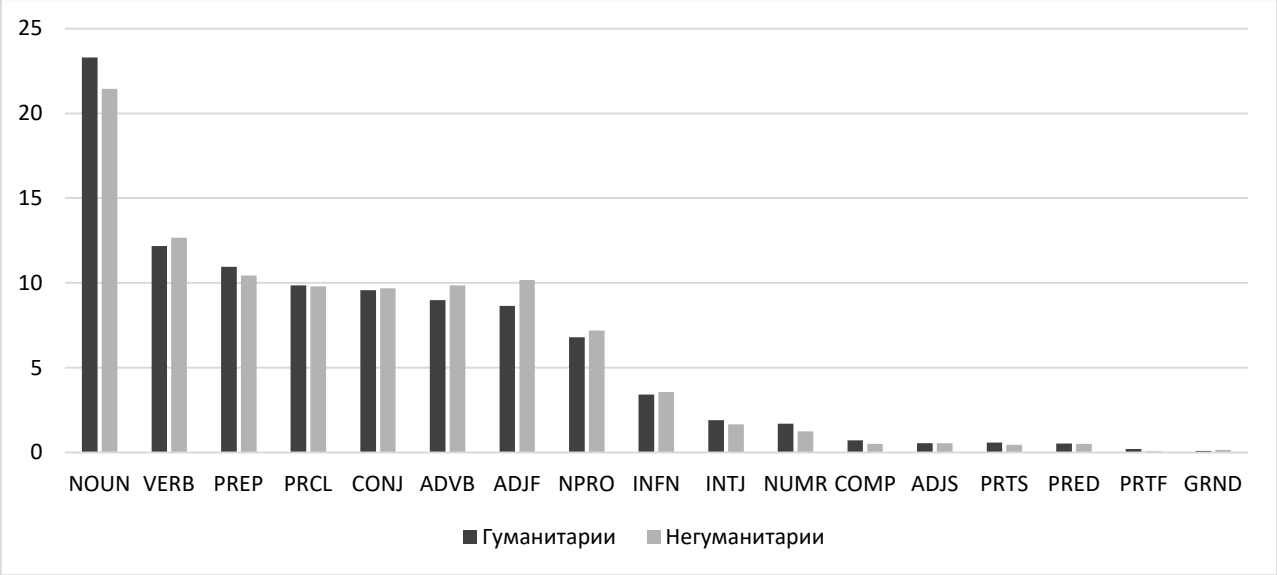

Рисунок 12. Частота реализации частей речи в зависимости от фактора «специальность»

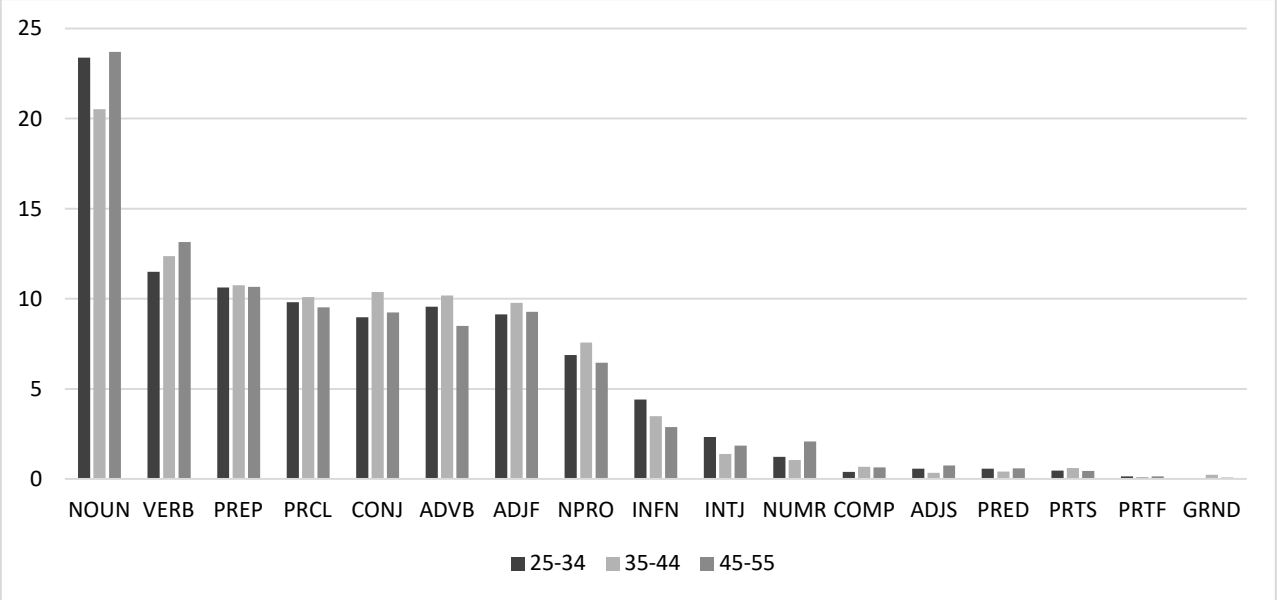

Рисунок 13. Частота реализации частей речи в зависимости от фактора «возраст»

Выводы и ограничения

В результате проведенного исследования показано, что на материале устных спонтанных текстов «О себе» параметр «КЛР» статистически значимо варьирует только в зависимости от фактора «возраст», объем различает также тексты информантов молодого и старшего возраста. Однако при применении поправки Бонферрони статистическая значимость варьирования не подтверждается. Вероятно, требуется дальнейшая проверка значимости указанных параметров на расширенной выборке текстов и, возможно, монологах на иную тему.

Тем не менее, наличие различий в значениях малоконтролируемых параметров длины и КЛР в двух возрастных группах подтверждает мнение Ф. Зенкера и К. Кейла о их прямой связи и в устных текстах.

Наиболее семантически нагруженный показатель – частоты частей речи – варьирует в текстах мужчин и женщин, что подтверждено параметрическим тестом: в текстах мужчин чаще используются существительные, числительные, предлоги и глаголы (предикативные компоненты высказывания), в текстах женщин чаще реализованы частицы, союзы, наречия, полные прилагательные и местоимения, что частично подтверждает выводы Т.А. Литвиновой и соавторов и опровергает результаты Н.В. Богдановой-Бегларян и соавторов, хотя так же, как в последнем

исследовании, в настоящей работе исследовались устные монологи. Качественные различия в реализациях частей речи, вероятно, показывают стратегии текстопорождения: называть и обозначить даты и количества у мужчин, характеризовать, описывать что-либо – в текстах женщин, однако, это заключение также требует проверки на расширенной выборке.

Примечание

1. Здесь и далее перевод иноязычных источников выполнен авторами.

Список литературы

Беляева А.С., Ерофеева Е.В. Зависимость частеречного варьирования устных спонтанных монологов от темы текста и социальных параметров говорящих // Филология в XXI веке. 2020. № 2(6). С. 77–88.

Богданова-Бегларян Н.В. и др. Некоторые инвариантные характеристики русской разговорной речи: фонетика, морфология, синтаксис / Богданова-Бегларян Н.В., Блинова О.В., Мартыненко Г.Я., Шерстинова Т.Ю. // Компьютерная лингвистика и интеллектуальные технологии: по материалам междунар. конф. 87 «Диалог 2017». 2017. [Электронный ресурс]. URL: https://publications.hse.ru/mirror/pubs/share/direct/213482429 (дата обращения: 27.10.2025).

Ерофеева Т.И. Социолект: стратификационное исследование: монография / Т.И. Ерофеева. Пермь: Перм. гос. ун-т, 2009. 240 с.

Захарова Е.Ю., Савина О.Ю. Лексическое разнообразие текста и способы его измерения // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanities. 2020. Т.6. №1 (21). С. 20–34.

Литвинова Т.А. и др. Исследование влияния пола и психологических характеристик автора на количественные параметры его текста с использованием программы Linguistic Inquiry and Word Count / Литвинова Т.А., Литвинова О.А., Рыжкова Е.С., Бирюкова Е.Д., Середин П.В., Загоровская О.В. // Научный диалог. 2015. № 12 (48). С. 101–109.

Covington M.A., McFall J.D. Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR) // Journal of Quantitative Linguistics. 2010. Vol. 17 (2). Pp. 94–100.

Dubois S., Sankoff D. The Variationist Approach toward Discourse Structural Effects and Sociointeractional Dynamics // The Handbook of discourse analysis / D. Schiffrin, D. Tannen and H. Hamilton (eds.). Malden, Massachusetts / Oxford: Blackwell Publishers Inc., 2001. Pp. 282–303.

Hess C.W. et al. The Type-Token Ratio and vocabulary performance / Hess C.W., Ritchie K.P., Landry R.G. // Psychological Report. 1984. Vol. 55 (1). Pp. 51–57.

Jarvis S. Capturing the Diversity in Lexical Diversity // Language Learning. 2013. Vol. 63. Pp. 87–106.

Liimatta A. Register variation across text lengths: Evidence from social media // *International Journal of Corpus Linguistics*. 2023. Vol. 28. №. 2. Pp. 202–231.

Litvinova T. et al. Differences in type-token ratio and part-of-speech frequencies in male and female Russian written texts / Litvinova T., Seredin P., Litvinova O., Zagorovskaya O. // Proceedings of the Workshop on Stylistic Variation / Copenhagen: Association for Computational Linguistics. 2017. Pp. 69–73.

McCarthy P.M., Jarvis S. Vocd: A theoretical and empirical evaluation // Language Testing. 2007. Vol. 24 (4). Pp. 459–488.

Mohd Razali N., Yap B. Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests // Journal of Statistical Modeling and Analytics (JOSMA). 2011. **№** 2. С. 21–33.

Singh S. A Pilot Study on Gender Differences in Conversational Speech on Lexical Richness Measures // Literary and Linguistic Computing. 2001. Vol. 16 (3). Pp. 251–264.

Yu G. Lexical Diversity in Writing and Speaking Task Performances // Applied Linguistics. 2010. Vol. 31 (2). Pp. 236–259.

Zenker F., Kile K. Investigating minimum text lengths for lexical diversity indices // Assessing Writing. 2021. Vol. 47. Pp. 1–15.

LEXICAL DIVERSITY, LENGTH OF TEXT, FREQUENCIES OF PARTS OF SPEECH AS INDICATORS OF SOCIOLINGUISTIC VARIATION

Ekaterina S. Khudiakova
Assistant Professor, Theoretical and Applied Linguistics Department
Perm State University

Stepan D. Kiselev
Student, Faculty of Philology
Perm State University

The study examines the extent to which three indicators (the lexical diversity (LD), the length of text in tokens (word forms), and the frequency of different parts of speech) demonstrate sociolinguistic variation in oral monologues. Based on a balanced sample of authors (N=48), adjusted for factors gender, age, specialty, and education level, statistical indicators of sample differences were calculated. The study utilized exclusively machine analysis methods using Python scripts. The results showed that, in the studied material, the lexical diversity parameter differs only depending on the "age" factor; the volume also differentiates the texts of younger and older informants. The frequencies of parts of speech vary significantly between the texts of men and women, and their qualitative differences likely reflect text-generation strategies.

Key words: sociolinguistic variation, gender, age, specialty, education level, lexical diversity, length of text in word forms, frequency of parts of speech.