DOI: 10.31862/2500-2953-2024-3-87-111
УДК 81'23:81-114.4

А.Ю. Хоменко[1, 2], Д.П. Бальба[1], Д.А. Исаков[1]

[1] Национальный исследовательский университет
«Высшая школа экономики»,
603155 г. Нижний Новгород, Российская Федерация

[2] ООО «Центр экспертизы и оценки “ЕСИН”»,
603093 г. Нижний Новгород, Российская Федерация

Кластеризация лексики для решения задачи диагностики афазии

В данной работе реализуется алгоритм автоматической кластеризации лексики для решения задачи создания компьютерной модели диагностики афазии на основе результатов теста на семантическую вербальную беглость. Данный тест представляет собой называние максимального количества слов, соответствующих некоторой заданной категории, за определенный промежуток времени. При анализе результатов теста часто исследуют количественные и качественные значения: количество правильных ответов, количество переключений – переходов от лексики одной субкатегории, хранящейся в сознании в виде группы лексем, сгруппированных на основе некоторого общего признака, к лексике другой субкатегории. При этом объединения лексем на основе общего признака принято называть кластерами, размеры которых также являются одним из качественных показателей прохождения теста. Была выдвинута гипотеза, согласно которой респонденты, подверженные афазии, дают меньшее количество ответов с меньшим количеством переключений, но похожим размером кластеров, а также имеют менее выраженную семантическую близость в кластерах и между ними. Авторами использовался алгоритм автоматической кластеризации лексики на основе правил, предложенный в работе Н. Лундин и коллег (Psychiatry Research. 2022. Т. 309), а также материалы теста на вербальную беглость, собранные и описанные в работе О.В. Буйволовой (Российский журнал когнитивной науки. 2020. Т. 7). В результате была создана функциональная модель для определения афазии у носителей

русского языка. Было также показано, что респонденты с афазией склонны давать менее стандартизированные ответы, формирующие меньшее количество подкатегорий, однако с большей ассоциативной связью внутри кластеров. Также с помощью сравнительного анализа с данными Национального корпуса русского языка было выявлено, что для респондентов обеих исследуемых групп – контрольной группы и группы лиц с афазией – характерно давать стандартные для носителей русского языка ассоциации.
Ключевые слова: диагностика афатических расстройств, кластеризация лексики, вербальная беглость

Благодарности. Исследование осуществлено в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики».

ДЛЯ ЦИТИРОВАНИЯ: Хоменко А.Ю., Бальба Д.П., Исаков Д.А. Кластеризация лексики для решения задачи диагностики афазии // Рема. Rhema. 2024. № 3. С. 87–111. DOI: 10.31862/2500-2953-2024-3-87-111

DOI: 10.31862/2500-2953-2024-3-87-111

A. Khomenko[1, 2], D. Balba[1], D. Isakov[1]

[1] HSE University,
Nizhny Novgorod, 603155, Russian Federation

[2] LLC “Center for Expertise and Assessment ‘ESIN’ ”,
Nizhny Novgorod, 603093, Russian Federation

Vocabulary clustering for solving problems of aphasia diagnostics

In this article an algorithm of automatic vocabulary clustering is implemented to solve the problem of creating a computer model for aphasia diagnostics based on the results of the verbal fluency test. This test involves generating the maximum number of words that fit a specified category within a given time limit. When analyzing test results, researchers often examine both quantitative and qualitative metrics: the total number of correct responses

and the number of “switches” – transitions from one subcategory of lexicon to another. The subcategories are presented as groups of lexemes that are organized based on some common attributes. They are usually referred to as clusters and their sizes serve as additional qualitative indicators of test performance. The study hypothesizes that the respondents who were affected by aphasia produced a lower number of responses with fewer switches, but showed similar cluster size, as well as less distinct semantic similarity in and between clusters. The authors employed a rule-based algorithm of automatic vocabulary clustering, proposed in the work of N. Lundin et al. (Psychiatry Research. 2022. T. 309), along with the materials collected and described in the work of O. Buivolova et al. (Russian Journal of Cognitive Science. 2020. Vol. 7. No. 3). As a result, a functional model for determining aphasia among Russian speakers was created. The findings also indicate that respondents with aphasia tend to give less standardized answers, forming fewer subcategories, but with a greater associative score within clusters. Additionally, comparative analysis with data from The Russian National Corpus revealed that respondents of both groups tend to provide associations, that are typical and standardized for native Russian speakers.

Key words: aphasia diagnostic, vocabulary clustering, verbal fluency

Acknowledgments. This article is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

FOR CITATION: Khomenko A., Balba D., Isakov D. Vocabulary clustering for solving problems of aphasia diagnostics. *Rhema.* 2024. No. 3. Pp. 87–111. (In Rus.). DOI: 10.31862/2500-2953-2024-3-87-111

1. Введение

Ранняя диагностика когнитивных нарушений является крайне важной задачей в психиатрии, неврологии и педагогике, т.к. она позволяет запустить восстановительный процесс вовремя и предупредить осложнения. Клиническая лингвистика позволяет выявить эти нарушения при помощи анализа устной речи как на микролингвистическом (фонетическом, семантическом, лексическом), так и на макролингвистическом уровнях. Данная задача традиционно решается вручную, однако она нуждается в цифровых инструментах, которые помогут быстро и объективно оценить, требуется ли человеку помощь специалиста.

Для быстрой оценки расстройств мышления в различных клинических популяциях используется тест на вербальную беглость, входящий в состав батарей стандартизированных тестов и использующийся при проведении фундаментальных исследований психики. Данный тест характеризуется простотой процедуры и чувствительностью к когнитивным нарушениям и состоит в воспроизведении максимального количества слов, принадлежащих к одной семантической категории (например, «животные» или «фрукты»). Как правило, при оценке результатов теста учитывается только количество правильных ответов, однако современные исследования стремятся учитывать не только количественные, но и качественные показатели порожденного списка слов. Это обусловлено тем фактом, что семантические связи в последовательностях слов пациентов с нарушениями отличаются от семантических связей в группе нормы. Обработка результатов теста традиционно проводится исследователями вручную, однако в данной статье предлагается автоматический способ оценки семантических паттернов выполнения задания при помощи методов векторной семантики и языковых моделей. При этом недавние исследования показывают, что инструменты компьютерной лингвистики дают возможность извлекать детализированную семантическую информацию с большей эффективностью и надежностью, нежели при анализе, проведенном вручную [Corcoran, 2020, p. 163]. До настоящего момента используемый в работе алгоритм не применялся к русскоязычному материалу, полученному после проведения тестов на вербальную беглость для лиц, страдающих афазией, что обусловливает новизну работы.

В существующих аналогичных англоязычных исследованиях на материале тестирования лиц с психическими расстройствами проверяются гипотезы о том, что члены группы носителей с заболеваниями продуцируют меньшее количество ответов, а после кластеризации в этой группе обнаруживается меньшее количество переключений между кластерами, при этом средние размеры кластеров респондентов разных групп не различаются [Elvevåg et al., 2002; Lundin et al., 2020, 2022]. В данном исследовании мы проверяем эту гипотезу на русскоязычном речевом материале респондентов с афатическими расстройствами.

2. Теория категоризации как основа представления кластеров

При поиске семантически связанных групп слов в ответах респондентов опираются на теорию категоризации, согласно которой в сознании носителя для каждой единицы формируется категория и для каждой

категории – определенные единицы [Болдырев, 2014]. Существует три основных подхода к осмыслению данной теории.

Одним из них является прототипический подход, при котором в основе организации категории находится иерархия ее членов, где наиболее значимыми представителями этой категории называют прототипы [Rosch, 1978].

Прототипы отражают признаки категории в наибольшей степени и часто являются первой реакцией в качестве ответа на стимул-категорию. Внутри одной лексической категории вокруг прототипов в сознании человека группируются другие элементы, которые входят в данную категорию. Эти элементы формируют свои субкатегории, объединяясь между собой по какому-то общему характерному признаку, позволяющему опознавать субкатегорию [Elvevåg et al., 2002; Болдырев, 2014; Lundin et al., 2020]. Данные факты дают нам возможность говорить о том, что при продукции речи участники тестов создают связанные группы слов – кластеры [Lundin et al., 2022]. Во время переключения (смены одного кластера другим) говорящий переходит от одного слова к другому, менее связанному семантически.

Различные исследования указывают на то, что данные контрольных групп и групп с различными психоневрологическими нарушениями имеют расхождения, что позволяет опираться на результаты кластеризации при диагностике [Ibid].

3. Метод автоматической кластеризации лексики

Для изучения особенностей прохождения теста лицами из разных клинических групп можно использовать инструменты компьютерной лингвистики, в частности методы векторной семантики и языковые модели, которые позволяют извлекать детализированную семантическую информацию с большей скоростью и надежностью, нежели при анализе, проведенном вручную. Данная работа основана на алгоритме кластеризации, предложенном в исследовании, которое посвящено построению диагностической модели для определения наличия у англоговорящих респондентов психотических расстройств (psychosis) на ранней стадии при помощи теста на семантическую вербальную беглость [Ibid]. Данный алгоритм до настоящего момента не использовался в исследованиях афатических расстройств на русскоязычном материале.

Для реализации данного алгоритма в настоящем исследовании используется датасет, содержащий ответы респондентов на тест на вербальную беглость, который проводился Центром языка и мозга Национального исследовательского университета «Высшая школа экономики» [Буйволова и др., 2020].

При тестировании были собраны результаты по трем семантическим категориям: «животные», «профессии» и «города». После первичной предобработки с удалением пустых строк, возникших в результате ошибки сбора или расшифровки, были удалены элементы, заполняющие хезитационные паузы ([а-, э-, м-]-образные звуки, десемантизированная лексика). При помощи морфологического анализатора PyMorphy2[1] была проведена токенизация и лемматизация материала [Korobov, 2015]. Следующим этапом стал экспертный анализ и ручное тегирование некоторых лексем.

В результате данной предобработки был получен список токенов из ответов респондентов, содержащий существительные, соответствующие заданной категории. К нему применялся описанный в работе [Lundin et al., 2022] алгоритм кластеризации, на основе которого считались статистические метрики, описанные в разделе 4.

3.1. Описание материалов исследования

Используемый датасет представляет собой русскоязычный речевой материал трех групп: ответы контрольной группы неврологически здоровых лиц, ответы лиц с афазией различного генезиса и ответы группы дисфазии – 120, 60 и 20 человек соответственно [Буйволова и др., 2020]. Все участники теста являются носителями русского языка.

Под афазией понимается специфическое нарушение речи, возникающее при органических поражениях головного мозга различной этиологии [Арутюнян, 2013]. Отдельно стоит упомянуть понимание термина дисфазия (dysphasia). В этой работе термин дисфазия выступает как приобретенное нарушение устной и письменной речи в результате травмы. При этом трактовка данного понятия в различных источниках варьируется: в отечественной литературе традиционно под дисфазией понимается нарушение развития речи у детей, а в иностранной литературе понятия дисфазии и афазии часто воспринимаются как синонимичные или взаимозаменяемые. В данной работе термин дисфазия являет собой менее тяжелую форму афазии [Perry, 2013].

Во время первого этапа предобработки был удален речевой материал тех респондентов, ответы которых отсутствовали в результате ошибки сбора или расшифровки. Таким образом, после преобразования сохранились следующие значения: 94 ответа контрольной группы и 69 ответов в группе лиц с афатическими расстройствами и дисфазией. Так как данных группы дисфазии оказалось недостаточно для противопоставления ее двум другим группам, было решено объединить их с группой лиц

[1] URL: https://pymorphy2.readthedocs.io/en/stable/ (дата обращения: 12.01.2024).

с афазией различного генезиса (с сохранением метки «dys» в качестве типа афатического расстройства).

После первичной предобработки было выделено две группы: контрольная группа и группа лиц с афатическим расстройством. Данные классы использовались в качестве материалов исследования.

Возраст респондентов контрольной группы варьируется от 17 до 86 лет, группы лиц с афазией – от 26 до 79 лет.

Среди лиц с афазией встречаются носители следующих типов заболевания: акустико-мнестическая, моторная афферентная, амнестическая, проводниковая, тотальная сенсомоторная, динамическая, моторная эфферентная, сенсорная афазия и дисфазия. Преобладающим типом оказалась моторная афазия (табл. 1). Также следует учитывать, что многие респонденты подвержены нескольким типам афатического расстройства (например, частое сочетание – моторная афферентная и эфферентная одновременно).

Таблица 1

Количество респондентов с различными типами афатического расстройства [Number of respondents with different types of aphasic disorder]

Тип афатического расстройства [Type of aphasic disorder]	**Количество респондентов [Number of respondents]**
Акустико-мнестическая афазия [Acoustic-amnestic aphasia]	8
Моторная афферентная афазия [Motor afferent aphasia]	13
Амнестическая афазия [Amnestic aphasia]	1
Проводниковая афазия [Conduction aphasia]	16
Тотальная сенсомоторная афазия [Total sensorimotor aphasia]	1
Динамическая афазия [Dynamic aphasia]	6
Моторная эфферентная афазия [Motor efferent aphasia]	24
Сенсорная афазия [Sensory aphasia]	15
Дисфазия [Dysphasia]	16

В данной работе мы преследовали цель описания общих особенностей категоризации лексики, свойственных людям, страдающим афатическими расстройствами, и отличающих их от носителей языка группы нормы, что объясняет объединение лиц всех типов афатического расстройства в одну группу для противопоставления контрольной группе.

3.2. Предобработка материалов

Перед кластеризацией ответы респондентов были предобработаны для приведения их к общему машиночитаемому виду: в частности, были удалены вербально заполненные паузы хезитации разного рода: фонетико-фонологические (такие как *э-э*, *хм*, удлинения звуков и т.д.), лексико-семантические (*так*, *ну*, *в общем* и т.д.), синтаксические (непреднамеренные повторы слов и словосочетаний, незаконченные высказывания). Также убирались лексемы, не соответствующие задаче (например, упоминание *колобка в* категории «животные»; такие лексемы затрудняли исследование ассоциаций между словами конкретной категории, при этом, встретившись в малом количестве в данных, не позволяли сделать каких-то дополнительных выводов, мы воспринимали такие случаи как шум в данных), и повторы (в случае с повторами сохранялось только первое упоминание). Следующим этапом была произведена лемматизация ответов при помощи морфологического анализатора PyMorphy2 [Korobov, 2015].

Затем после экспертного анализа материала было произведено ручное тегирование некоторых групп лексики. В частности, для лексем с пояснениями (например *Свердловск*, *то есть Екатеринбург* или *козел горный*, *то есть архар*) был введен тег ЛСП, для лексем с уточнениями (*удав*, *нет*, *это пресмыкающееся* или *попугай*, *это ж*, *хоть и птица*, *но животное*) – ЛСУ, для семантически восстанавливаемых окказионализмов (*топибара* – семантическая группа «животные», предположительно, восстановимо до *капибара*) – СВО, для семантически невосстанавливаемых окказионализмов (*мовы*, *стантея*) – СНО. Чтобы избежать утраты полноты информации, были созданы два набора данных – с тегами (т.н. набор «чистых» данных), где лексемы были заменены на соответствующие теги, и набор данных со всеми лексемами, где мы сохранили лексемы СВО и СНО в исходном виде, а для ЛСП и ЛСУ сохранили обе лексемы (главное слово и пояснение/уточнение) в лемматизированном виде.

3.3. Алгоритм кластеризации

Для преобразования ответов в векторное представление использовалась предобученная статическая эмбеддинговая модель geowac_lemmas_

none_fasttextskipgram_300_5_2020 сервиса RusVectōrēs[2], с помощью которой были получены эмбеддинги типа fastText [Bojanowski, 2017]. Данная модель представляет собой полносвязную нейронную сеть, обученную на русскоязычном подкорпусе корпуса GeoWAC [Dunn, Adams, 2020] на задачу предсказания по данному (центральному) слову других слов из его контекста (контекст определяется как окно длиной $2k + 1$, где k – количество слов до/после центрального). При этом для создания эмбеддингов типа FastText каждое слово перед обучением модели делится на токены – символьные n-граммы, где n – количество символов в одном токене (для данной модели были взяты $n = 3, 4$ и 5). Вектор слова, таким образом, определяется как сумма векторов всех n-грамм, входящих в это слово, и вектора самого слова. Данный тип векторных представлений слов был выбран в связи с тем, что описанный выше алгоритм обучения эмбеддингов позволяет кодировать окказионализмы или слова с фонетической эрозией (т.е. слова, которые не могли встретиться в обучающем корпусе текстов, но легко получаются из обученных n-грамм), продуцируемые респондентами с афазией.

При кластеризации использовалось правило, предложенное в работе [Lundin et al., 2022]. Согласно этому правилу, для последовательности слов A, B, C, D переключение находится после слова B в случае, когда косинусная близость $S(X, Y)$ между векторами слов A и B больше косинусной близости между векторами слов B и C, а косинусная близость между векторами слов B и C меньше, чем косинусная близость между векторами слов C и D: $S(A, B) > S(B, C) < S(C, D)$.

Для проведения кластеризации лексики был написан код на языке Python с использованием библиотек Gensim[3] и Pandas[4].

4. Результаты

Были получены наборы кластеров для каждого респондента по каждой из трех семантических категорий: «животные», «профессии» и «города». Также с помощью программного кода на Python для каждого набора кластеров были посчитаны следующие метрики: количество переключений между кластерами, средний размер кластера, среднее расстояние между кластерами, средние значения метрик t-score и silhouette-score. Далее все эти значения были агрегированы по категориям

[2] RusVectōrēs: семантические модели для русского языка. URL: https://rusvectores.org/ (дата обращения: 12.01.2024).

[3] URL: https://radimrehurek.com/gensim/ (дата обращения: 12.01.2024).

[4] URL: https://pandas.pydata.org/ (дата обращения: 12.01.2024).

и респондентам и были получены средние, медианы и стандартные отклонения метрик для каждой группы респондентов и каждого типа предобработки (с тегированием и без).

4.1. Результаты кластеризации

Результаты применения алгоритма кластеризации лексики к ответам отдельных респондентов для семантической категории «животные» показан на графиках (рис. 1–3).

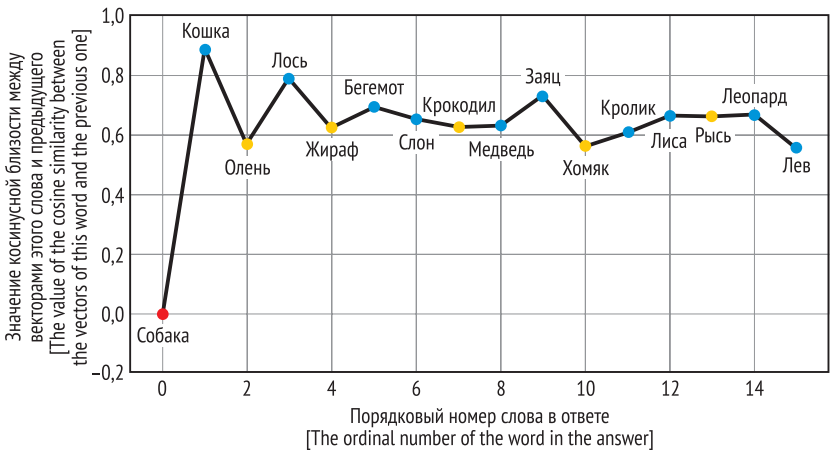

Рис. 1. Визуализация кластеризации ответа респондента (ID C6) контрольной группы для семантической категории «животные»

Для первого слова в ответе значение косинусной близости между векторами этого слова и предыдущего равно 0

Начало ответа отмечено красным, переключение (первое слово в новом кластере после переключения) – желтым цветом

Fig. 1. Visualization of the clustering of the respondent's answer (ID C6) of the control group for the semantic category "animals"

For the first word in the answer, the value of the cosine similarity between the vectors of this word and the previous one is 0

The beginning of the answer is marked in red, the switch (the first word in the new cluster after the switch) is in yellow

На приведенных примерах (см. рис. 1–3) можно заметить, что у респондентов контрольной группы графики располагаются ближе к единице по оси ординат, что говорит о большей семантической близости слов в пределах кластеров и ответа в целом (т.к. семантическая

близость между последним словом одного кластера и словом-переключением все равно сохраняет высокие значения, что говорит о том, что несмотря на переключение, слова являются семантически близкими).

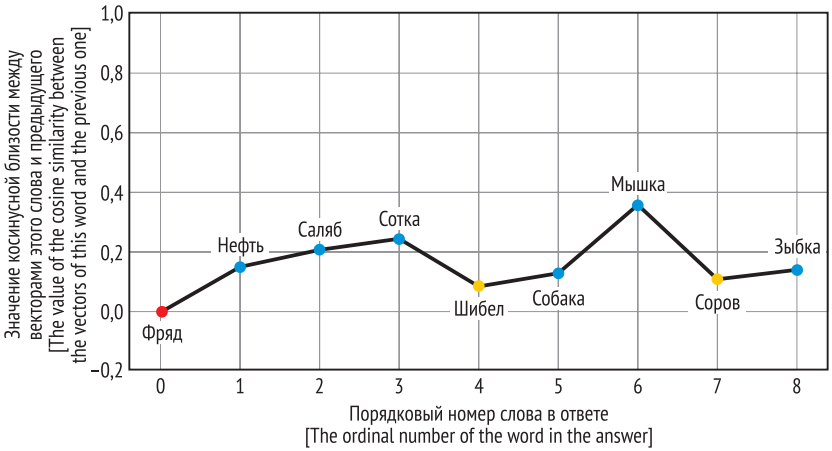

Рис. 2. Визуализация кластеризации ответа респондента (ID P012) группы с афатическим расстройством для семантической категории «животные»

Для первого слова в ответе значение косинусной близости между векторами этого слова и предыдущего равно 0

Начало ответа отмечено красным, переключение (первое слово в новом кластере после переключения) – желтым цветом

Fig. 2. Visualization of clustering of the respondent's answer (ID P012) from the group with aphasic disorder for the semantic category "animals"

For the first word in the answer, the value of the cosine similarity between the vectors of this word and the previous one is 0

The beginning of the answer is marked in red, the switch (the first word in the new cluster after the switch) is in yellow

Для респондента группы с афатическом расстройством можно наблюдать низкие значения *y*-координаты на всем графике, что говорит о низкой семантической близости слов (рис. 2). В данном случае это также определяется окказионализмами, приведенными респондентом в качестве ответов.

Для сравнения приведен график ответов респондента группы афазии, не включающий в себя окказионализмы (рис. 3). Можно наблюдать присутствие стандартизированных (прототипических) ответов, для

которых значение семантической близости близко к единице, однако график представляет собой более «ломаную» траекторию ответов, чем в группе нормы, что свидетельствует о меньшей семантической близости ответов.

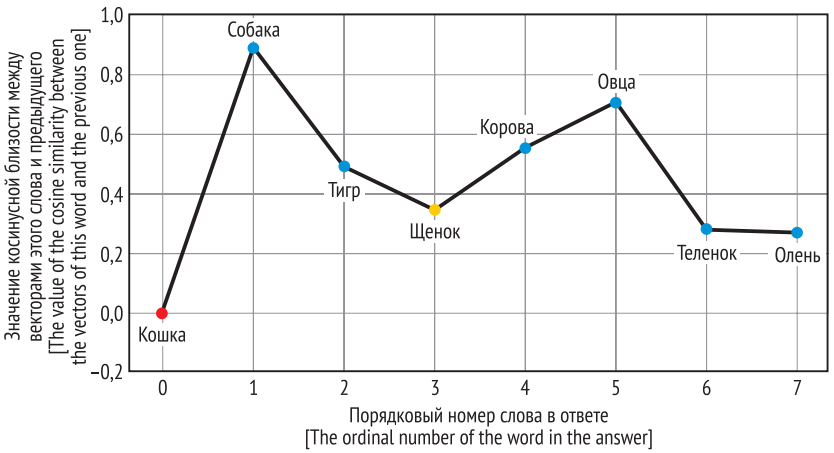

Рис. 3. Визуализация кластеризации ответа респондента (ID P001) группы с афатическим расстройством для семантической категории «животные»

Для первого слова в ответе значение косинусной близости между векторами этого слова и предыдущего равно 0

Начало ответа отмечено красным, переключение (первое слово в новом кластере после переключения) – желтым цветом

Fig. 3. Visualization of clustering of the respondent's answer (ID – P001) from the group with aphasic disorder for the semantic category "animals"

For the first word in the answer, the value of the cosine similarity between the vectors of this word and the previous one is 0

The beginning of the answer is marked in red, the switch (the first word in the new cluster after the switch) is in yellow

4.2. Статистический анализ результатов кластеризации

После работы алгоритма были подобраны такие метрики результатов кластеризации, которые оценивают особенности категоризации лексики в сознании носителей языка из двух исследуемых групп и показывают различия между ними. Так, были выделены следующие метрики:

– количество переключений для каждого человека по каждой категории;

– среднее расстояние между кластерами для каждого человека по каждой категории (считается как расстояние между центроидами кластеров; за центроид берется среднее значение векторов, находящихся в одном кластере);
– средний размер кластера для каждого человека.

Данные метрики уже использовались ранее в аналогичных исследованиях кластеризации лексики, продуцируемой респондентами теста на вербальную беглость, и показывали значимые различия между исследуемыми группами [Lundin et al., 2022]. Также нами были протестированы следующие метрики, способные показывать количественную информацию о свойствах ассоциативных связей ответов респондентов.

1. Среднее значение *t*-score в кластере для каждого человека по каждой категории. *T*-score – это адаптированная корпусная метрика, отображающая то, насколько неслучайной является сила ассоциации между коллокатами, т.е. насколько неслучайно два слова встретились рядом друг с другом в корпусе, насколько сильна их семантическая связь [Захаров, Хохлова, 2010]. В качестве коллокатов в нашем случае берутся все последовательности из двух слов внутри одного кластера. Метрика *t*-score для одной пары слов рассчитывалась следующим образом:

$$t\text{-score} = \frac{f(n,\ c) - \frac{f(n)\cdot f(c)}{N}}{\sqrt{f(n,\ c)}},$$

где $f(n)$, $f(c)$ – частоты слов n и c в данной семантической категории соответственно; $f(n,\ c)$ – частота совместной встречаемости слов n и c (под совместной встречаемостью мы понимаем нахождение двух слов в одном кластере при условии отсутствия других слов между ними); N – количество всех слов в данной семантической категории.

2. Метрика silhouette-score отображает, насколько близок объект к своему кластеру по сравнению с другими кластерами: чем больше значение данной метрики, тем больше объект «похож» на объекты своего собственного кластера [Rousseeuw, 1987]. Silhouette-score для каждого слова рассчитывается следующим образом:

$$\text{silhouette-score} = \frac{a - b}{\min(a,\ b)},$$

где a – средняя косинусная близость между данным словом и другими словами в том же кластере, в котором находится данное слово; b – средняя косинусная близость между данным словом и словами из ближайшего кластера к тому, в котором находится данное слово (в нашем

случае ближайший кластер – это следующий или предыдущий кластер в последовательности). Отметим, что silhouette-score принимает значения от –1 до 1. Также экспериментальным путем мы наложили ограничения на значения *a* и *b* так, что $|a| > 0{,}01$ и $|b| > 0{,}01$.

3. Коэффициент лексического разнообразия – type-token ratio (TTR). Данная метрика позволяет оценить лексическое разнообразие выборки, диапазон и вариативность словарного запаса, который говорящий реализует. Нами был применен самый известный способ измерения коэффициента лексического разнообразия, вычисляемый следующим образом:

$$TTR = \frac{V}{N},$$

где *V* – количество уникальных лексических единиц; *N* – общее количество словоформ. Данная метрика принимает значения от 0 (не включительно) до 1 [Захарова, 2020]. При этом, т.к. в исследовании мы использовали леммы, *N* сводится к общему количеству слов, которые названы некоторой выборкой, а *V* – к уникальным словам, которые эта выборка называет.

4. Средний размер словарного запаса (количество уникальных слов, названных респондентами одной группы). Анализ статистических метрик проводился с помощью языка программирования Python и библиотек Numpy[5] и Scipy[6]. Был использован статистический *t*-критерий Стьюдента для сравнения средних, исходя из предположения о независимости выборок и нормальном распределении данных. В качестве уровня значимости было взято стандартное значение $\alpha = 0{,}05$. После анализа статистических метрик было обнаружено, что значимых различий между значениями их средних для подвыборки с тегированием и подвыборки с исходными лексемами не обнаружено (табл. 2), поэтому далее анализ проводился только на втором наборе данных. Значения метрик можно увидеть в таблице 3.

Значение метрики *t*-score оказалось выше для контрольной группы. Это манифестирует, что в ней ассоциации между реакциями более стандартизированы, т.е. в ответах контрольной группы чаще встречаются одинаковые устойчивые пары слов-ассоциаций.

Еще одним подтверждением более стандартизированных ассоциаций в группе нормы является большее лексическое разнообразие TTR в группе афазии, которое, кроме менее стандартизированных связей,

[5] URL: https://numpy.org/ (дата обращения: 12.01.2024).

[6] URL: https://scipy.org/ (дата обращения: 12.01.2024).

мотивировано присутствием большого числа окказионализмов в ответах данной группы. Это объясняется тем, что значение TTR прямо пропорционально количеству уникальных слов в данных, которое оказывается выше для группы лиц с афазией из-за присутствия окказионализмов. Кроме того, меньшее значение лексического разнообразия свидетельствует о прототипических ответах в группе нормы. Несмотря на меньшее лексическое разнообразие, словарный запас (количество уникальных слов) в контрольной группе превышает словарный запас лиц с афазией в 1,7 раз. Можно заметить, что значение silhouette-score выше для лиц с афазией, что свидетельствует о большей связанности слов внутри одного кластера.

Таблица 2

Усредненные метрики кластеризации для подвыборок группы лиц с афазией с тегированием и без по всем категориям [Average clustering metrics for subsamples of aphasia group with and without tagging across all categories]

Статистические метрики [Statistical metrics]	**Контрольная группа с тегированием [Control group with tagging]**	**Контрольная группа с сохранением исходных лексем [Control group with preservation of original lexemes]**
Количество переключений [Number of switches]	2,38	2,39
Расстояние между кластерами [Distance between clusters]	0,65	0,66
t-score	22,01	21,68
Silhouette score	0,25	0,25
Размер кластера [Cluster size]	3,08	3,09

4.3. Анализ частотных коллокаций

Были также проанализированы наиболее частотные коллокации внутри кластеров, которые были отсортированы по относительной частоте внутри своей группы. Для подсчета наиболее частотных коллокаций использовалась относительная частота по корпусу всех коллокаций, которые образовывались в одном кластере. Наиболее частотные коллокации для категории «животные» приведены в таблице 4.

Таблица 3

Усредненные метрики кластеризации по всем категориям [Average clustering metrics across all categories]

Статистические метрики [Statistical metrics]	**Контрольная группа [Control group]**	**Лица с афазией [People with aphasia]**	**Статистическая значимость (по *t*-критерию Стьюдента, α = 0,05) [Statistical significance (Student's *t*-test, α = 0.05)]**
Количество переключений[*] [*Distance between clusters*[*]]	*5,23*; *1,88*[**]	*2,39*; *1,56*	$p < 0{,}001$
Расстояние между кластерами [Distance between clusters]	0,64; 0,04	0,66; 0,06	$p > 0{,}05$
t-score	*51,58*; *14,3*	*21,68*; *10,76*	$p < 0{,}001$
Silhouette score	*0,16*; *0,06*	*0,25*; *0,14*	$p < 0{,}001$
Размер кластера [Cluster size]	3,15; 0,31	3,09; 0,43	$p > 0{,}05$
TTR (мера лексического разнообразия) [Type-token ratio]	0,23	0,34	
Словарный запас (количество уникальных слов) [Vocabulary (number of unique words)]	412	241	

[*] Курсивом выделены статистически значимые метрики.

[**] Указаны среднее значение параметра (по каждому респонденту и категории) и его стандартное отклонение.

[[*] Statistically significant metrics are highlighted in italics.

[**] The average value of the parameter (for each respondent and category) and its standard deviation are indicated].

Эти коллокации имеют высокий сочетаемостный уровень в корпусе ответов респондентов. Заметим, что каждая из этих коллокаций входит в 10 наиболее частотных в Национальном корпусе русского языка (НКРЯ) (по состоянию на 13 февраля 2024 г.) (табл. 5). При поиске коллокатов в НКРЯ в качестве запроса использовалось первое слово

из коллокации, а коллокаты сортировались по значению *t*-score. Поиск производился по основному корпусу, для запроса первое слово задавалось как лемма, а для второго слова задавались параметры (существительное, нарицательное). Частотность данных пар говорит о репрезентации стандартных для русского языка коллокаций внутри кластеров. При этом относительная частота этих коллокаций в обеих группах равна, что может быть связано с тем, что разнообразие коллокаций в группе нормы мотивировано большим словарным запасом, а в группе афазии – большим количеством окказионализмов.

Таблица 4

Наиболее частотные коллокации со значением *t*-score для контрольной группы и группы лиц с афазией по семантической категории «животные» [The most frequent collocations with *t*-score value for the control group and the group of people with aphasia for the semantic category “animals”]

Контрольная группа [Control group]		**Лица с афазией [People with aphasia]**	
Коллокация [Collocation]	***t*-score**	**Коллокация [Collocation]**	***t*-score**
кошка–собака	23,93	*кошка–собака*	17,28
лев–тигр	13,53	*волк–медведь*	8,58
жираф–слон	9,53	*заяц–лиса*	7,91
корова–овца	7,53	*волк–лиса*	7,24
заяц–лиса	7,14	*лев–тигр*	5,90

Так как значения *t*-score выше для контрольной группы, можно сделать вывод о том, что ассоциации между словами внутри кластеров этой группы в среднем более стабильные, чем в ответах лиц с афазией. Похожую картину можно наблюдать для коллокаций категорий «профессии» (табл. 6, 7) и «города» (табл. 8, 9).

При этом можно заметить, что средние значения частоты коллокации и *t*-score максимальны для категории «животные» в обеих группах, т.е. там присутствуют наиболее стандартизированные ответы. Значение частоты и метрики *t*-score минимальны в категории «профессии», что говорит о наименее стандартизированных ответах в данной категории, а следовательно, и меньшей ее устойчивости. Это можно объяснить тем, что разнообразие ответов в категориях «профессии» и «города» зависит

ISSN 2500-2953

Таблица 5

Значения *t*-score коллокаций категории «животные» в Национальном корпусе русского языка (поиск по первому слову пары) для контрольной группы и лиц с афазией
[*T*-score values of collocations of the category “animals” in the Russian National Corpus (search by the first word of the pair) for the control group and individuals with aphasia]

	Коллокация [Collocation]	Относительная частота в корпусе ответов респондентов [Relative frequency in the corpus of respondents’ answers]	*t*-score в корпусе ответов респондентов [*t*-score in the corpus of respondents’ answers]	*t*-score в НКРЯ [*t*-score in the Russian National Corpus]	Место в списке коллокаций НКРЯ, отсортированном по частотности [Place in the list of collocations of the Russian National Corpus, sorted by frequency]
Контрольная группа [Control group]	*кошка–собака*	0,026	23,930	36,480	1
	лев–тигр	0,015	13,533	17,150	4
	жираф–слон	0,010	9,535	6,480	8
	корова–овца	0,008	7,535	26,460	1
	заяц–лиса	0,008	7,135	10,090	3
Лица с афазией [People with aphasia]	*кошка–собака*	0,028	17,280	36,480	1
	волк–медведь	0,014	8,580	22,370	3
	заяц–лиса	0,013	7,911	10,090	3
	волк–лиса	0,012	7,242	13,030	6
	лев–тигр	0,010	5,903	17,150	4

не только от культурного опыта респондента и наличия у него речевого нарушения, но во многом и от личностного опыта (например, респонденты склонны называть города своей родины, профессии врачей, которых они посещают, и т.п.).

Таблица 6

Наиболее частотные коллокации со значением *t*-score для контрольной группы и лиц с афазией групп по семантической категории «профессии» [The most frequent collocations with *t*-score value for the control group and individuals with aphasia groups by the semantic category "professions"]

Контрольная группа [Control group]		**Лица с афазией [People with aphasia]**	
Коллокация [Collocation]	***t*-score**	**Коллокация [Collocation]**	***t*-score**
врач–учитель	18,91	*врач–учитель*	13,04
преподаватель–учитель	10,53	*сантехник–электрик*	9,19
инженер–строитель	8,14	*сантехник–слесарь*	7,27
слесарь–токарь	8,14	*токарь–фрезеровщик*	7,27
врач–инженер	6,94	*профессор–учитель*	7,27

Однако в целом как респонденты без нарушений, так и респонденты с афазией склонны давать ответы, отражающие стандартные для носителей русского языка ассоциации (например, *кошка–собака*, *заяц–лиса*, *врач–учитель*, *Москва–Санкт-Петербург*), при этом ответы контрольной группы оказываются статистически более стабильными.

5. Выводы

В ходе данного исследования был создан драфт диагностического алгоритма для определения афатических нарушений речи у носителей русского языка. При этом он моделирует некоторые когнитивные особенности носителей языка с афазией и без нее. По результатам работы алгоритма кластеризации проявляются различия между двумя группами по нескольким метрикам: по количеству переключений между кластерами, по связанности на основе silhouette-score, по силе ассоциаций по *t*-score, а также по значениям лексического разнообразия TTR и объему словарного запаса. Также можно сказать, что обе группы склонны давать предсказуемые для носителей русского языка ассоциации,

ISSN 2500-2953

Таблица 7

Значения *t*-score коллокаций категории «профессии» в Национальном корпусе русского языка (поиск по первому слову пары) для контрольной группы и лиц с афазией
[*T*-score values of collocations of the category “professions” in the Russian National Corpus (search by the first word of the pair) for the control group and individuals with aphasia]

	Коллокации [Collocation]	Относительная частота в корпусе ответов респондентов [Relative frequency in the corpus of respondents’ answers]	*t*-score в корпусе ответов респондентов [*t*-score in the corpus of respondents’ answers]	*t*-score в НКРЯ [*t*-score in the Russian National Corpus]	Место в списке коллокаций НКРЯ, отсортированном по частотности [Place in the list of collocations of the Russian National Corpus, sorted by frequency]
Контрольная группа [Control group]	*врач–учитель*	0,008	18,914	25,020	2
	преподаватель–учитель	0,005	10,533	10,490	9
	инженер–строитель	0,004	8,138	17,800	5
	слесарь–токарь	0,004	8,138	10,680	1
	врач–инженер	0,003	6,941	20,170	4
Лица с афазией [People with aphasia]	*врач–учитель*	0,009	13,043	25,020	2
	сантехник–электрик	0,007	9,197	6,710	1
	сантехник–слесарь	0,005	7,273	7,680	3
	токарь–фрезеровщик	0,005	7,273	6,080	1
	профессор–учитель	0,005	7,273	10,490	9

однако контрольная группа в большей мере отображает общеязыковую норму согласно значениям *t*-score и результатам сравнения с данными НКРЯ. Наиболее стабильной оказалась категория «животные».

Таблица 8

Наиболее частотные коллокации со значением *t*-score для контрольной группы и лиц с афазией по семантической категории «города» [The most frequent collocations with *t*-score value for the control group and individuals with aphasia for the semantic category "cities"]

Контрольная группа [Control group]		Лица с афазией [People with aphasia]	
Коллокация [Коллокация]	*t*-score	Коллокация [Collocation]	*t*-score
Москва–Санкт-Петербург	27,77	*Москва–Санкт-Петербург*	13,99
Лондон–Париж	11,81	*Москва–Питер*	12,56
Ленинград–Москва	10,22	*Ленинград–Москва*	11,13
Севастополь–Симферополь	8,62	*Омск–Томск*	8,27
Вашингтон–Нью-Йорк	9,42	*Киев–Москва*	8,27

Можно также заметить, что гипотеза, выдвинутая на основе англоязычных материалов, подтверждается на материалах данного исследования не полностью. Действительно, респонденты, подверженные когнитивным дисфункциям (в нашем случае – афатическому расстройству), в среднем дают меньшее количество ответов, имеют меньшее количество переключений, а также среди ответов лиц с афазией наблюдается менее выраженная семантическая близость. При этом средний размер кластеров и среднее расстояние между ними не имеют значимых различий (табл. 10), тогда как зарубежные ученые на основе исследования группы с шизофреническим спектром подчеркивали менее выраженную близость между кластерами [Elvevåg et al., 2002; Lundin et al., 2020, 2022].

Отметим, что данное исследование имеет ряд ограничений. Во-первых, небольшой объем выборки: для эффективного использования модели с методами машинного обучения и более точной диагностики расстройств желательно иметь в распоряжении несколько сотен объектов для обучения. Во-вторых, неравномерная репрезентация типов афазии: большинство респондентов подвержены моторной афазии, что не позволяет нам на данном этапе исследования изучать особенности отдельных типов речевых нарушений.

ISSN 2500-2953

Таблица 9

Значения *t*-score коллокаций категории «города» в Национальном корпусе русского языка (поиск по первому слову пары) для контрольной группы и лиц с афазией
[*T*-score values of collocations of the category “cities” in the Russian National Corpus (search by the first word of the pair) for the control group and individuals with aphasia]

	Коллокации [Collocation]	Относительная частота в корпусе ответов респондентов [Relative frequency in the corpus of respondents’ answers]	*t*-score в корпусе ответов респондентов [*t*-score in the corpus of respondents’ answers]	*t*-score в НКРЯ [*t*-score in the Russian National Corpus]	Место в списке коллокаций НКРЯ, отсортированном по частотности [Place in the list of collocations of the Russian National Corpus, sorted by frequency]
Контрольная группа [Control groups]	*Москва–Санкт-Петербург*	0,013	27,766	28,050	14
	Лондон–Париж	0,006	11,814	32,820	2
	Ленинград–Москва	0,005	10,218	38,440	2
	Севастополь–Симферополь	0,004	8,623	7,280	4
	Вашингтон–Нью-Йорк	0,004	9,421	10,950	2
Лица с афазией [People with aphasia]	*Москва–Санкт-Петербург*	0,010	13,990	28,050	14
	Москва–Питер	0,009	12,560	23,950	25
	Ленинград–Москва	0,008	11,131	38,440	2
	Омск–Томск	0,006	8,271	6,630	2
	Киев–Москва	0,006	8,271	27,780	12

Таблица 10

Полученные после кластеризации данных значения метрик, предложенных для сравнения в гипотезе [The values of the metrics proposed for comparison in the hypothesis obtained after data clustering]

	Контрольная группа [Control group]	**Группа лиц с афазией [Group of people with aphasia]**
Среднее количество ответов [Average number of responses]	19,46	10,52
Среднее количество переключений [Average number of switches]	5,23	2,39
Средний размер кластеров [Average cluster size]	3,14	3,08
Среднее расстояние между кластерами [Average distance between clusters]	0,64	0,65

Полученные результаты доказывают эффективность применения методов компьютерной лингвистики для задачи диагностики афазии на русскоязычном материале и открывают потенциал для использования предложенного алгоритма кластеризации лексики для разных типов афазии. Перспективной также выглядит диагностика ментальных расстройств (депрессии, шизофрении и пр.) с применением компьютерных алгоритмов кластеризации.

Библиографический список / References

Арутюнян и др., 2013 – Арутюнян В.Г., Клачек П.М., Кошелева И.Л. Мозг и афазия: нейролингвистические подходы как основа методики для компьютеризации диагностики и реабилитации // Труды XI международной научной конференции «Инновации в науке, образовании и бизнесе – 2013». Калининград, 2013. С. 338. [Arutyunyan V.G., Klachek P.M., Kosheleva I.L. Brain and aphasia: Neurolinguistic approaches as a basis for the methodology for computerization of diagnostics and rehabilitation. *Trudy XI mezhdunarodnoy nauchnoy konferentsii «Innovatsii v nauke, obrazovanii i biznese – 2013»*. Kaliningrad, 2013. P. 338. (In Rus.)]

Болдырев, 2014 – Болдырев Н.Н. Когнитивная семантика. Введение в когнитивную лингвистику: учебное пособие. 4-е изд., испр. и доп. Тамбов, 2014. [Boldyrev N.N. Kognitivnaya semantika. Vvedenie v kognitivnuyu lingvistiku [Cognitive Semantics. Introduction to cognitive linguistics]. Tambov, 2014.]

Буйволова и др., 2020 – Adaptation of the Aphasia Bedside Check for Russian / О.В. Буйволова и др. // Российский журнал когнитивной науки. 2020. Т. 7.

№ 3. С. 45–67. [Buivolova O.V. et al. Adaptation of the Aphasia Bedside Check for Russian. *Russian Journal of Cognitive Science*. 2020. Vol. 7. No. 3. Pp. 45–67.]

Захаров, Хохлова, 2010 – Захаров В.П., Хохлова М.В. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.). Вып. 9 (16). М., 2010. С. 137–143. [Zakharov V.P., Khokhlova M.V. Analysis of the effectiveness of statistical methods for identifying collocations in Russian-language texts. *Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference* "*Dialogue*" (*2010*). Issue 9 (16). Moscow, 2010. Pp. 137–143. (In Rus.)]

Захарова, Савина, 2020 – Захарова Е.Ю., Савина О.Ю. Лексическое разнообразие текста и способы его измерения // Вестник Тюменского государственного университета. Серия: Гуманитарные исследования. Humanitates. 2020. Т. 6. № 1 (21). С. 20–34. DOI: 10.21684/2411-197X-2020-6-1-20-34 [Zakharova E.Yu., Savina O.Yu. Lexical diversity measures' review and classification. *Tyumen State University Herald. Humanities Research. Humanitates*. 2020. Vol. 6. No. 1 (21). Pp. 20–34. DOI: 10.21684/2411-197X-2020-6-1-20-34 (In Rus.)]

Савчук и др., 2024 – Савчук С.О., Архангельский Т.А., Бонч-Осмоловская А.А. Национальный корпус русского языка 2.0: новые возможности и перспективы развития // Вопросы языкознания. 2024. № 2. С. 7–34. [Savchuk S.O., Arkhangelsky T.A., Bonch-Osmolovskaya. A.A. Voprosy yazykoznaniya. 2024. No. 2. Pp. 7–34. (In Rus.)]

Bojanowski, 2017 – Bojanowski P. et al. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*. 2017. Vol. 5. Pp. 135–146.

Corcoran, 2020 – Corcoran C.M. et al. Language as a biomarker for psychosis: A natural language processing approach. *Schizophrenia Research*. 2020. Vol. 226. Pp. 158–166.

Dunn, Adams, 2020 – Dunn J., Adams B. Geographically-balanced gigaword corpora for 50 language varieties. *Proceedings of the Language Resources and Evaluation Conference* (*LREC 2020*). 2020. Pp. 2528–2536.

Elvevåg et al., 2002 – Elvevåg B., Fisher J.E., Gurd J.M., Goldberg T.E. Semantic clustering in verbal fluency: Schizophrenic patients versus control participants. *Psychol. Med*. 2002. No. 32. Pp. 909–917.

Korobov, 2015 – Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages. *Analysis of Images, Social Networks and Texts*. 2015. Pp. 320–332.

Lundin et al., 2020 – Lundin N.B. et al. Semantic search in Psychosis: Modeling local exploitation and global exploration. *Schizophr Bull Open*. 2020. No. 1. Pp. 1–11.

Lundin et al., 2022 – Lundin N.B. et al. Semantic and phonetic similarity of verbal fluency responses in early-stage psychosis. *Psychiatry Research*. 2022. Vol. 309. P. 114404.

Rosch, 1978 – Rosch E. Principles of categorization. *Cognition and Categorization*. Routledge, 1978. Pp. 27–48.

Rousseeuw, 1987 – Rousseeuw P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 1987. No. 20. Pp. 53–65.

Статья поступила в редакцию 19.08.2024
The article was received on 19.08.2024

Сведения об авторах / About the authors

Хоменко Анна Юрьевна – кандидат филологических наук; старший научный сотрудник Центра языка и мозга, научный сотрудник Лаборатории теории и практики систем поддержки принятия решений, приглашенный преподаватель департамента фундаментальной и прикладной лингвистики, Национальный исследовательский университет «Высшая школа экономики», г. Нижний Новгород; эксперт-лингвист, эксперт-авторовед, эксперт-фоноскопист, ООО «Центр экспертизы и оценки “ЕСИН”», г. Нижний Новгород

Anna Yu. Khomenko – PhD in Linguistics; senior researcher at the Center for Language and Brain, Researcher at the Laboratory of Theory and Practice of Decision Support Systems, Visiting Lecturer at the Department of Fundamental and Applied Linguistics, HSE University, Nizhny Novgorod; Expert Linguist, Author Expert, Expert Phonoscopist, LLC “Center for Expertise and Assessment ESIN”, Nizhny Novgorod, Russian Federation

ORCID: https://orcid.org/0000-0003-3564-6293
E-mail: akhomenko@hse.ru

Бальба Дарья Петровна – студент бакалавриата образовательной программы «Фундаментальная и прикладная лингвистика» факультета гуманитарных наук, Национальный исследовательский университет «Высшая школа экономики», г. Нижний Новгород; стажер-исследователь Центра языка и мозга, Национальный исследовательский университет «Высшая школа экономики», г. Нижний Новгород

Darya P. Balba – BA student at the educational program “Fundamental and Applied Linguistics” of the Faculty of Humanities, HSE University, Nizhny Novgorod; research intern at the Center for Language and Brain, HSE University, Nizhny Novgorod, Russian Federation

ORCID: https://orcid.org/0009-0004-1434-6966
E-mail: dpbalba@edu.hse.ru

Исаков Данила Андреевич – студент бакалавриата образовательной программы «Фундаментальная и прикладная лингвистика» факультета гуманитарных наук, Национальный исследовательский университет «Высшая школа экономики», г. Нижний Новгород; стажер-исследователь Центра языка и мозга, Национальный исследовательский университет «Высшая школа экономики», г. Нижний Новгород

Danila A. Isakov – BA student at the educational program “Fundamental and Applied Linguistics” of the Faculty of Humanities, HSE University, Nizhny Novgorod; research intern at the Center for Language and Brain, HSE University, Nizhny Novgorod, Russian Federation

ORCID: https://orcid.org/0009-0000-3354-5048
E-mail: issakov.daanila@mail.ru

Все авторы прочитали и одобрили окончательный вариант рукописи
All authors have read and approved the final manuscript