Социологические исследования
Sociological researches

DOI: 10.24412/2070-1381-2022-91-148-161

Семантический анализ структуры ценностей группы с помощью Тезауруса Роже: автоматизированный алгоритм

***Андреюк Денис Сергеевич*[1]**
Кандидат биологических наук, доцент, экономический факультет, МГУ имени М.В. Ломоносова; старший научный сотрудник, Государственное бюджетное учреждение здравоохранения города Москвы «Психиатрическая клиническая больница № 1 им. Н.А. Алексеева Департамента здравоохранения г. Москвы»; исполнительный директор Общероссийской общественной организации «Российская ассоциация содействия науке», Москва, РФ.
E-mail: denis.s.andreyuk@yandex.ru
SPIN-код РИНЦ: 8083-4058
ORCID ID: 0000-0002-3349-5391

Ливитина Арина Сергеевна
Аналитик, ООО «РА «ИНДЕКС 20», Москва, РФ.
E-mail: arina.livitina@gmail.com
ORCID ID: 0000-0002-1060-7688

Сушко Никита Сергеевич
Программист-разработчик ПО, АО «Медицинские Технологии Лтд», Москва, РФ.
E-mail: gorakievskaya@gmail.com
ORCID ID: 0000-0003-2245-7354

Аннотация
В работе предложен подход для векторизации и количественного анализа ценностей группы. Для демонстрации возможностей метода проведен анализ структуры ценностей группы молодых людей на предмет различий между женщинами и мужчинами и различий при использовании разных частей речи. Ценности группы были вербализованы в виде свободных ассоциаций «с чем-то самым важным в жизни». Полученный массив слов преобразовали в массив семантических групп с помощью Тезауруса Роже. Попарное сравнение векторов с частотами отдельных семантических групп показало высокий уровень косинусной близости (0,9664) между подгруппами, разделенными по гендерному признаку. Расчет статистически значимых различий в частотах отдельных семантических групп методом Хи-квадрат позволил выделить отдельные семантические группы, по которым гендерные подгруппы достоверно различаются. Векторы частот, полученные от преобразования массивов разных частей речи, имели низкий уровень косинусной близости во всех попарных сравнениях. Существительные наиболее часто использовали для выражения жизненных ценностей, имеющих отношение к причинно-следственным связям (14% семантических групп); прилагательными чаще всего выражали ценности, имеющие смысл личных пристрастий (18% сематических групп); глаголы чаще всего использовали для выражения ценностей, связанных с симпатиями (14% семантических групп). Разработанный автоматический алгоритм может быть полезен для количественного сопоставления ценностей между разными группами, а также расчета степени соответствия ценностей целевой группы заявляемым ценностям коммерческих брендов.

Ключевые слова
Семантический анализ, тезаурус Роже, программа на языке Python, векторизация ценностей, ценности молодежи.

Semantic Analysis of Group Values Structure Using Roget Thesaurus: Automated Algorithm

***Denis S. Andreyuk*[2]**
PhD, Associate Professor, Faculty of Economics, Lomonosov Moscow State University; Senior Researcher, Mental Health Clinic № 1 named after N.A. Alexeev; Executive Director of the Russian Association for the Advancement of Science, Moscow, Russian Federation.
E-mail: denis.s.andreyuk@yandex.ru
ORCID ID: 0000-0002-3349-5391

Arina S. Livitina
Data Analyst, LLC "RA "INDEX 20", Moscow, Russian Federation.
E-mail: arina.livitina@gmail.com
ORCID ID: 0000-0002-1060-7688

Nikita S. Sushko
IT Specialist, Program Developer, JSC "Medical Technology Ltd", Moscow, Russian Federation.
E-mail: gorakievskaya@gmail.com
ORCID ID: 0000-0003-2245-7354

[1] Корреспондирующий автор.
[2] Corresponding author.

Abstract
The article proposes an approach for vectorization and quantitative analysis of group values. To demonstrate the possibilities of the method, the value structure of a group of young people was analyzed for differences between women and men, and for differences in the use of different parts of speech. The values of the group were verbalized in the form of free associations "with something most important in life". The resulting array of words was converted into an array of semantic groups using Roget Thesaurus. Pairwise comparison of vectors with frequencies of individual semantic groups showed a high level of cosine similarity (0,9664) between subgroups separated by gender. Calculation of statistically significant differences in frequencies of separate semantic groups by chi-square test allowed us to single out separate semantic groups, for which gender subgroups differed significantly. Frequency vectors obtained from the transformation of arrays of different parts of speech had a low level of cosine similarity in all pairwise comparisons. Nouns were most frequently used to express life values related to cause-and-effect relationships (14% of semantic groups). Adjectives were most often used to express values having a sense of personal predilections (18% of semantic groups). Verbs were most often used to express values related to liking (14% of semantic groups). The developed automatic algorithm will be useful for quantitative comparison of values between different groups, as well as calculating the degree of consistency of the target group values with the declared values of commercial brands.

Keywords
Semantic analysis, thesaurus Roget, Python script, vectorization of values, youth values.

Введение

Структура ценностей вызывает интерес исследователей на протяжении многих десятилетий и даже веков, поскольку представляет собой основу мотивации в целенаправленном поведении социальных групп и актуальна для любых культур и для любого времени [TenHouten 1997]. Если мы знаем соотношение важности целей для данного человека в данный момент времени, мы можем предсказывать приоритетность его поступков. Например, если для человека удовольствие от познания стоит выше в системе ценностей, чем удовольствие от еды, то для такого человека покупка на последние деньги книги вместо еды не кажется удивительной. При этом статистически такая структура ценностей относительно редка, а значит, в случае среднестатистического человека выбор «книга вместо еды» парадоксален и маловероятен.

Индивидуальная структура приоритетов/целей сложна для изучения в силу своей динамичности — соотношение ценностей у конкретного человека постоянно меняется (см., например, [Баева 2004; Человек в условиях глобальных рисков 2020]). Особенно сильно это обстоятельство ограничивает практическую применимость знаний о ценностях. Мы вынуждены затратить много сил, времени и денег для того, чтобы определить соотношение приоритетов в выборе для данного человека, но эти приоритеты изменились уже в процессе измерения, и наши выводы даже относительно поведения данного индивида оказываются под сомнением. Еще больше сомнений возникает при необходимости экстраполяции выводов на «других таких же».

Из описанных выше ограничений следует очевидная схема: измерения структуры ценностей должны быть дешевыми, быстрыми, массовыми. Таким критериям удовлетворяют современные подходы семантического анализа цифровых следов. В частности, для задач маркетинга, когда интересна не сама глубинная структура ценностей, а только ее локальное следствие — те предпочтения, которыми группа «похожих друг на друга» людей руководствуется, когда типичному представителю группы необходимо принять решение о покупке или в более широком смысле обмене одних ценностей на другие. Для этой задачи достаточно знать соотношение ценности денег у представителей данной группы с ценностью товаров или услуг, имеющихся в ассортименте и подлежащих обмену. И тогда можно предложить к обмену те позиции, которые для данной целевой группы в данный момент времени представляют максимальную ценность. Такой обмен повысит одновременно и маржинальность сделки для продавца и удовлетворенность от сделки для покупателя.

В настоящее время целый ряд подобных задач решается с помощью методов дистрибутивной семантики. Принципиальная логика данного подхода базируется на гипотезе, что если какие-то слова часто встречаются в похожем контексте, то они близки по смыслу

[Sahlgren 2008]. На основе собранных корпусов текстов в сотни миллионов слов строятся многомерные векторы, характеризующие контекст данного слова. После этого формальные операции по определению векторных расстояний позволяют судить о семантической близости любых двух данных слов [Митрофанова 2007]. В рамках данного подхода словесный цифровой след может быть представлен векторами, а поиск человека, похожего по интересам данному, сводится к задаче определения векторных расстояний по массивам слов в следах одного и другого и поиска людей-следов с наименьшими отличиями.

У данного подхода есть ряд существенных ограничений:

— все люди из целевой группы должны иметь максимально похожий цифровой след, а собственно ценности при этом вторичны (например, два человека в равной степени обеспокоены проблемой изменения климата на планете; один, движимый этим беспокойством, предпочитает веганские рестораны мясным, а другой посещает научные конгрессы по океанологии; с точки зрения цифрового следа это будут абсолютно разные люди, хотя глубинная мотивация у обоих очень похожая);

— частота употребления слов в текстах только косвенно связана со смыслом этих слов, то есть с тем, какие ассоциации слово рождает в голове у говорящего и у слушающего. Во многих случаях могут существовать значительные расхождения между статистически наиболее распространенными и реальными ассоциациями в ответ на слово — эти случаи некорректно исследовать с помощью дистрибутивной семантики;

— в некоторых нишевых традициях считается дурным тоном использовать одно и то же слово слишком часто; например, у литераторов (довольно массовая профессиональная группа) показателем высокого уровня культуры и искусства считается умение передать одну и ту же мысль с помощью разных слов; используемые синонимы имеют заведомо разные частоты в стандартных корпусах и при этом призваны передавать заведомо одинаковый смысл.

В более общей формулировке подходы дистрибутивной семантики не очень подходят в тех случаях, когда исследователю важно иметь возможность непосредственно обращаться к смыслу слов. Для таких задач перспективным представляется использование идеографических тезаурусов — словарей, построенных по принципу каталогизации смысла. Одним из наиболее известных словарей этой группы является Тезаурус Роже [Roget 1991]. Его первоначальная версия была составлена Питером Марком Роже в 1805 г. и опубликована в 1852 г. Позже версия, изданная в 1911 г., была компьютеризирована Л. Джоном Олдом и Д. Левенштейном и выложена в свободный доступ на сайте Project Gutenberg[3]. Словарь разбивает все слова английского языка на 1044 семантических групп[4], входящих в 40 смысловых категорий. Они, в свою очередь, принадлежат 6 семантическим классам:

1) Words Expressing Abstract Relations (слова, выражающие абстрактные отношения);
2) Words Relating to Space (слова, относящиеся к пространству);
3) Words Relating to the Intellectual Faculties (слова, относящиеся к интеллектуальным способностям);

[3] Roget's Thesaurus by Peter Mark Roget // Project Gutenberg [Электронный ресурс]. URL: https://www.gutenberg.org/ebooks/22 (дата обращения: 12.01.2022).
[4] Изначально Тезаурус Роже включал ровно 1000 смысловых групп, однако в более поздних версиях стали добавляться разделения групп, в итоге используемое здесь 8-е издание содержит уже 1044 конечных категорий — семантических групп.

4) Words Relating to the Sentient and Moral Powers (слова, относящиеся к чувствам и морали);
5) Words Relating to the Voluntary Powers (слова, относящиеся к добровольным полномочиям);
6) Words Relating to Matter (слова, относящиеся к материи).

Фактически каждая конечная семантическая группа в Тезаурусе Роже — это группа синонимов, передающих один и тот же смысл.

В данной работе мы разработали методический подход для анализа структуры ценностей группы людей с использованием Тезауруса Роже. Для демонстрации возможностей был создан автоматизированный алгоритм и опробован для анализа жизненных ценностей в группе российской молодежи. В частности, были проверены две гипотезы:

1) женщины и мужчины на уровне смыслов имеют очень похожую структуру ценностей;
2) ценности группы, выражаемые разными частями речи, в значительной мере отличаются между собой по смыслу.

Методика

Источник данных: ассоциации молодых людей с «чем-то самым важным в жизни». Исходные данные для анализа ценностей были получены путем опроса школьников, студентов и молодых ученых (очно, данные собирались до пандемии). Респондентов просили написать «три слова — существительное, прилагательное и глагол, которые ассоциируются с чем-то самым важным в жизни». Такой опрос проводили в 27 разных группах, в 4 городах в период 2018–2019 гг. Более подробно технология проведения опросов описана в [Андреюк и др. 2020]. Для целей данной работы использовали 639 анкет (все, которые удовлетворяли по качеству записей). 367 участников указали свой пол как женский, 263 указали пол как мужской, 9 человек отказались указывать пол. Возраст составил 19,04 года +/-3.33 (стандартное отклонение), разброс 13–38.

Массив слов переводился с русского на английский с помощью автоматического сервиса Deepl.com, и полученный массив английских слов преобразовывался в массив кодов Роже.

Преобразование массива слов в массив кодов семантических групп Тезауруса Роже. Авторами был написан скрипт на языке Python. Программа выложена в открытый доступ и может быть использована для широкого круга исследований. Для этого необходимо скачать файловый архив .zip, который находится по ссылке https://github.com/chameleon-lizard/Roget, с последующей его распаковкой. Сама программа запускается через файл gui.exe в папке gui, после чего открывается окно с четырьмя квадрантами (Рисунок 1). Окно Input отображает исходные данные, которые нужно ввести в файл input.txt, находящийся в папке Roget-main. После нажатия на кнопку Vectorize результаты отображаются в двух квадрантах Output. И верхний, и нижний квадранты показывают код, название принадлежащей ему семантической группы, встречаемость данного кода в тестируемом массиве слов. Верхний квадрант показывает, сколько раз данная семантическая группа встретилась в массиве в абсолютных значениях, нижний квадрант показывает частоту кода/семантической группы в процентах от общего числа кодов. Экспорт данных автоматически осуществляется в файлы, создаваемые в каждой сессии в папке Roget-main, после работы программы данные

в них автоматически обновляются. Коды с абсолютным количеством находятся в файле vector.json, с процентной частотой — в frequences.json. Кроме этого, в файл non-recognized.txt выгружаются все слова, которым не было найдено соответствий в тезаурусе.

Наиболее удобный для дальнейшей работы формат данных — вектор, состоящий из 1044 значений. Номер позиции — это номер кода семантической группы (коды выстроены по возрастанию), значение — частота встречаемости данного кода в виде десятичной дроби (без процентов). Именно такие векторы использовались во всех последующих операциях (Рисунок 2).

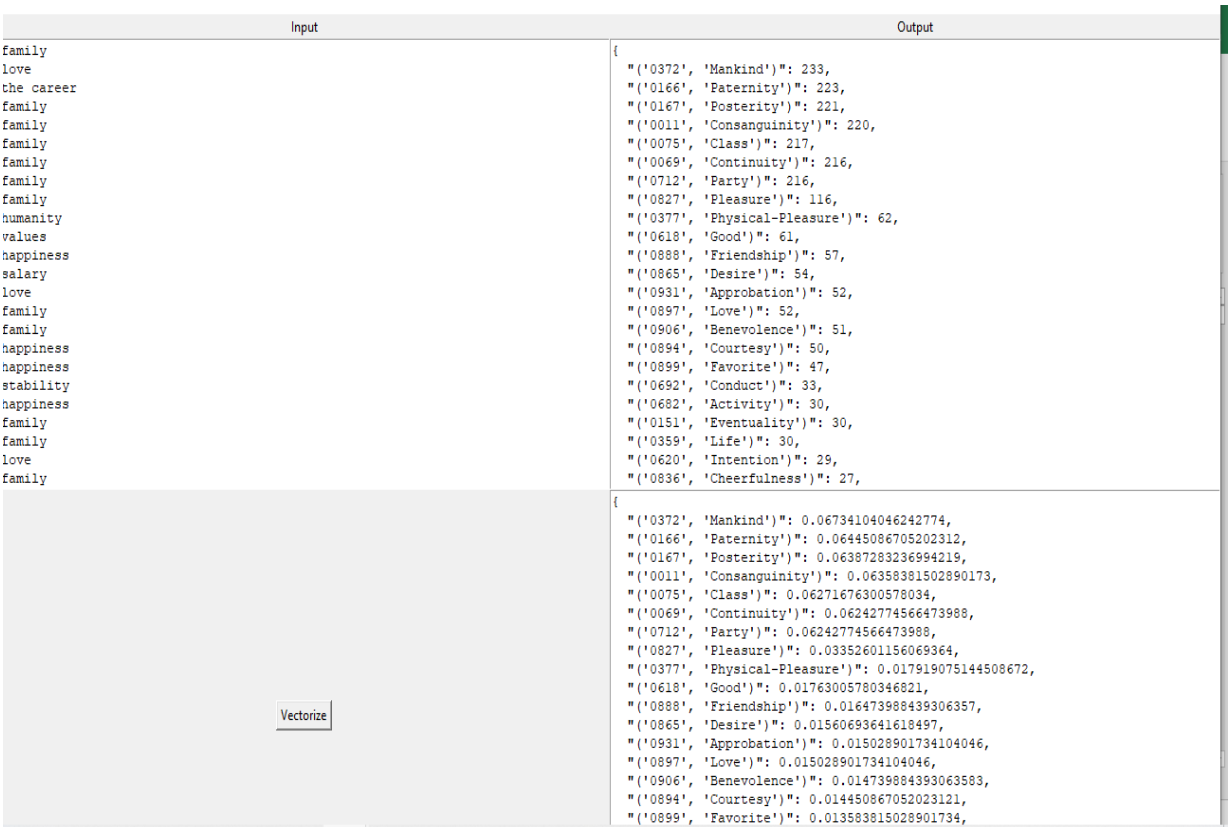


Рисунок 1. Интерфейс программы: пример вывода результатов анализа массива слов[5]

Операции сравнения смысловых векторов

Измерение векторного расстояния — косинусная близость. В данной работе применялась наиболее распространенная формула для определения косинусной близости (коэффициент Отиаи [Ochiai 1957]):

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}};$$

Оценка различий в частотах семантических групп для двух векторов. Для определения достоверности различий частот в двух массивах для данной семантической группы использовался метод парного сравнения по критерию Хи-квадрат. Поскольку анализу подлежат частоты кодов из

[5] Изображение получено авторами при работе программного продукта; сам программный продукт свободно открыт для некоммерческого использования и доступен для скачивания по ссылке https://github.com/chameleon-lizard/Roget.

одного массива, тестируемые события не являются независимыми. В этом случае может возникать проблема ошибок при множественных сравнениях.

Было проверено влияние трех распространенных поправок. Поправка Бонферрони, подход к коррекции FWER — Family-Wise Error Rate, позволяет контролировать групповую вероятность ошибки первого рода [Hochberg 1988]. В нашем случае алгоритм учета поправки сводится к уменьшению уровня α в m раз ($Pi \leq \frac{\alpha}{m}$; общепринятый уровень значимости для Хи-квадрат α = 0.05), где m — количество позиций вектора, в нашем случае m = 1044.

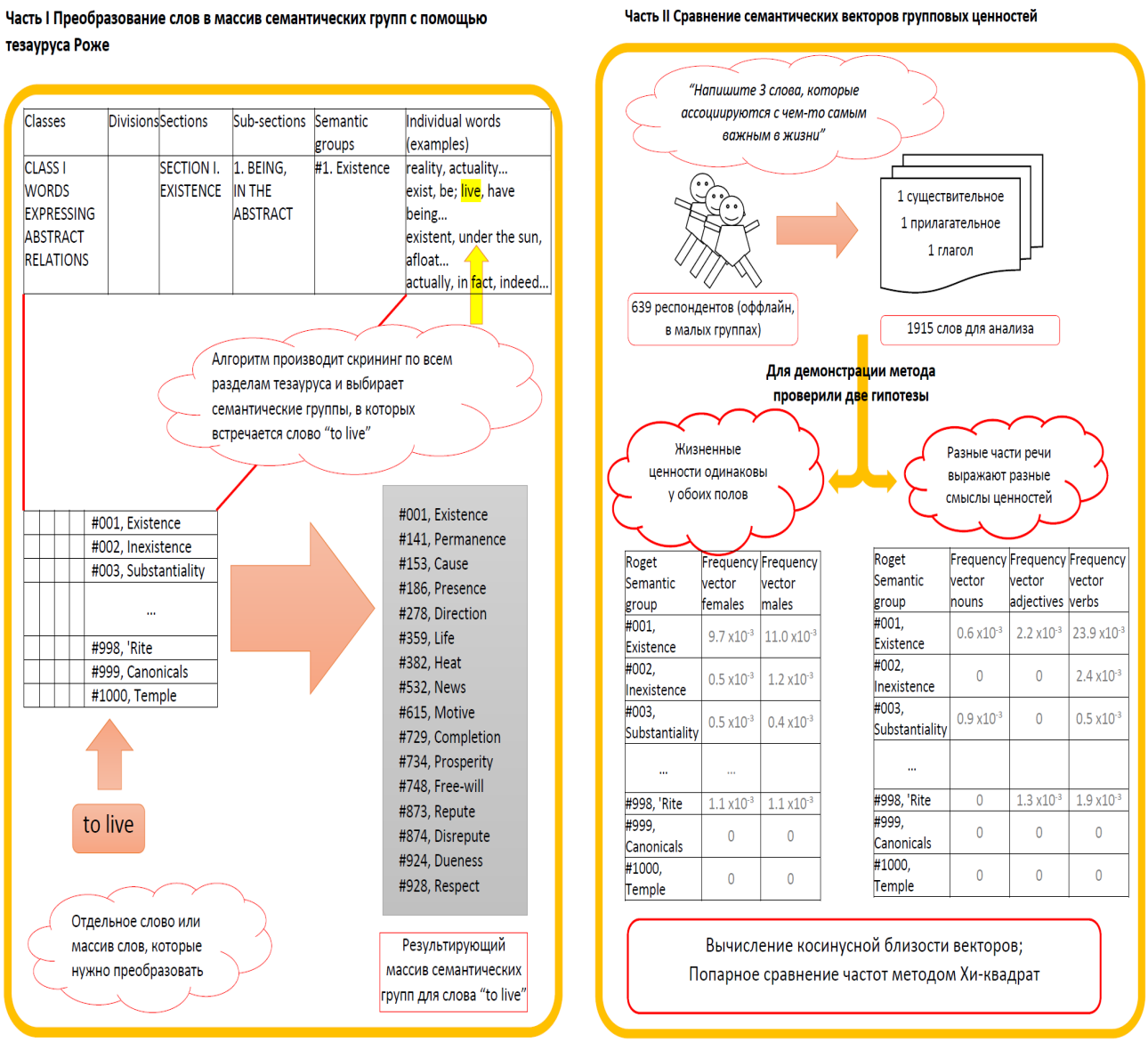
Более сложный вариант такого подхода к коррекции представляет поправка Холма-Бонферрони [Holm 1979]. Алгоритм коррекции предлагает сортировать все гипотезы по возрастанию их p-value, затем для каждой гипотезы проверить неравенство:

$$Pi < \frac{\alpha}{m - i}.$$

Кроме этого, существует подход FDP — False Discovery Proportion, который предписывает контролировать долю ложных отклонений гипотез. Представителем этого подхода была выбрана поправка Беньямини-Хохберга [Benjamini, Hochberg 1995]. Ее алгоритм предлагает пошагово увеличивать в i раз уровень значимости, скорректированный на число множественных сравнений, где i — порядковый номер сравнения, после сортировки сравнений по возрастанию значения p:

$$\alpha_1 = \frac{\alpha}{m}, \ \ldots, \ \alpha_i = \frac{i\alpha}{m}, \ \ldots, \ \alpha_m = \alpha.$$

Важно отметить, что справедливость и эффективность поправок на множественное сравнение неоднократно оспаривалась в отношении анализа экспериментальных данных. Диапазон мнений расходится от полного отрицания целесообразности поправок [Rothman 1990] до аккуратной рекомендации использовать FWER-поправки для надежного отсечения в автоматических обработках больших массивов (как самый строгий фильтр), а поправки FDP и полное отсутствие поправок — как диапазон для поиска возможных событий-кандидатов для дальнейших исследований [Noble 2009].



Classes	Divisions	Sections	Sub-sections	Semantic groups	Individual words (examples)
CLASS I WORDS EXPRESSING ABSTRACT RELATIONS		SECTION I. EXISTENCE	1. BEING, IN THE ABSTRACT	#1. Existence	reality, actuality... exist, be; live, have being... existent, under the sun, afloat... actually, in fact, indeed...

#001, Existence
#002, Inexistence
#003, Substantiality
...
#998, 'Rite
#999, Canonicals
#1000, Temple

#001, Existence
#141, Permanence
#153, Cause
#186, Presence
#278, Direction
#359, Life
#382, Heat
#532, News
#615, Motive
#729, Completion
#734, Prosperity
#748, Free-will
#873, Repute
#874, Disrepute
#924, Dueness
#928, Respect

Roget Semantic group	Frequency vector females	Frequency vector males
#001, Existence	9.7×10^{-3}	11.0×10^{-3}
#002, Inexistence	0.5×10^{-3}	1.2×10^{-3}
#003, Substantiality	0.5×10^{-3}	0.4×10^{-3}
...	...	
#998, 'Rite	1.1×10^{-3}	1.1×10^{-3}
#999, Canonicals	0	0
#1000, Temple	0	0

Roget Semantic group	Frequency vector nouns	Frequency vector adjectives	Frequency vector verbs
#001, Existence	0.6×10^{-3}	2.2×10^{-3}	23.9×10^{-3}
#002, Inexistence	0	0	2.4×10^{-3}
#003, Substantiality	0.9×10^{-3}	0	0.5×10^{-3}
...			
#998, 'Rite	0	1.3×10^{-3}	1.9×10^{-3}
#999, Canonicals	0	0	0
#1000, Temple	0	0	0

Рисунок 2. Принцип работы программного продукта для автоматического преобразования слов в массивы семантических групп по Роже (Часть I) и общая схема эксперимента, показывающего возможности метода (Часть II)[6]

Результаты

Количественный анализ ценностных смыслов в гендерно-однородных подгруппах. Если сравнить семантическое ядро у представителей двух полов, окажется, что рейтинг наиболее частотных слов у них практически совпадает (Таблица 1).

[6] Составлено авторами.

Таблица 1. Топ-10 наиболее частых слов, используемых мужчинами и женщинами для выражения ассоциаций «с чем-то самым важным в жизни»[7]

	Мужчины	Женщины
1	семья (10.0%)	семья (13.6%)
2	любить (5.3%)	любить (9.9%)
3	счастливый (3.1%)	счастливый (6.3%)
4	жить (2.5%)	жить (3.8%)
5	работа (2.2%)	счастье (3.5%)
6	счастье (2.1%)	работа (2.1%)
7	радоваться (1.7%)	любимый (1.9%)
8	красивый (1.6%)	развиваться (1.3%)
9	интересный (1.6%)	радоваться (1.3%)
10	жизнь (1.6%)	интересный (1.3%)

Первые 4 слова в рейтинге одинаковые (в общей выборке слов на эти 4 слова приходится 28% частот), в остальных словах заметна значительная доля совпадений, интуитивно чувствуется, что по смыслу эти слова также близки. Эту близость можно измерить количественно.

Сравнение векторов частот ценностных смыслов по формуле косинусной близости позволяет оценить степень общего совпадения смыслов в двух массивах слов. В частности, для представителей двух полов косинусная близость оказывается исключительно высокой: 0,9664. Это означает, что, хотя слова часто используются разные, смысловое содержание жизненных ценностей у женской и мужской половин почти идентичное. Тем самым подтверждается гипотеза о равенстве полов в отношении жизненных ценностей.

Однако векторизация позволяет проанализировать также и различия в смыслах, даже если они ничтожно малы, как в случае с гендерными подгруппами. В Таблице 2 приведены те смысловые группы Тезауруса Роже, частота встречаемости в которых максимально различается у представителей двух полов (парное сравнение Хи-квадрат).

Таблица 2. Частоты семантических групп Роже, по которым наиболее сильно выражено различие между мужчинами и женщинами в отношении жизненных ценностей[8]

	Мужчины, %	Женщины, %	значение p *
#023 Agreement	0.81	1.43	0.0038
#134 Occasion	0.65	1.20	0.0045
#242 Symmetry	0.32	0.05	0.0059
#578 Elegance	0.67	1.20	0.0074
#655 Decease	0	0.16	0.0027
#827 Pleasure	3.02	4.48	0.0002
#845 Beauty	0.65	0.19	0.0017
#888 Friendship	1.48	2.25	0.0062
#897 Love	1.75	2.60	0.0049
#899 Favorite	1.27	2.14	0.0010

[7] Составлено авторами. В скобках указаны частоты данного слова в процентах от всех слов в данной группе анкет.
[8] Составлено авторами. Примечание: * — различия достоверны при $p<0,05$, однако достоверность отсутствует во всех приведенных здесь парах, если применить поправку на множественное сравнение (любую из трех, описанных в разделе Методы). В случае кода #827 Pleasure абсолютное количество в подгруппе мужчин составило 112 раз, а в подгруппе женщин — 253 раза. Если бы соотношение было хотя бы 110 к 253, различия были бы достоверны с учетом поправки Беньямини-Хохберга.

В целом весь спектр различий между представителями разных полов можно показать на диаграмме Манхэттен, когда по горизонтальной оси отложены все категории, а по вертикальной оси указан отрицательный логарифм p с указанием отсекающего барьера на уровне p<0,05 (см. Рисунок 3, где отсекающая линия показана красным; все точки, которые выше, показывают достоверную разницу в смысловых частотах без учета поправки на множественные сравнения; зеленая линия отсекает уровень значимости с учетом поправки Бонферрони — самой жесткой из трех примененных здесь поправок на множественные сравнения). На рисунке наглядно видно, что ни одно из различий нельзя признать достоверным с учетом поправки на множественные сравнения. Тем не менее для наиболее отличающихся смысловых групп можно говорить о четко выраженной тенденции.

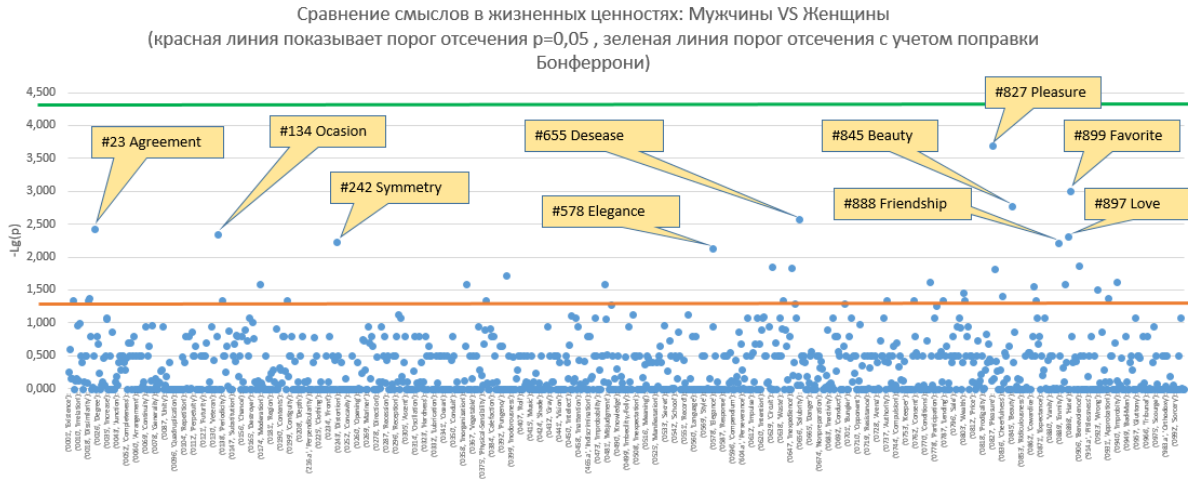


Рисунок 3. Визуализация отличий в частотах отдельных смысловых групп с распределением по длине вектора[9]

Сравнение ценностей, передаваемых в разных частях речи. В ходе преобразования массивов отдельных частей речи, в которых представители целевой группы выражали «самое важное в жизни», в частоты встречаемости семантических групп Роже были получены три вектора. Их попарное сравнение дает возможность количественно сопоставить степень близости смыслов, выражаемых разными частями речи (Таблица 3). Видно, что, в отличие от гендерных подгрупп, подгруппы, сформированные из разных частей речи, различаются между собой значительно, что подтверждает и вторую гипотезу.

Таблица 3. Косинусная близость между векторами частот встречаемости семантических групп Роже в ценностях, выражаемых разными частями речи[10]

Сравниваемые части речи	Косинусная близость семантических векторов
Существительные vs Прилагательные	0.3135
Существительные vs Глаголы	0.2646
Прилагательные vs Глаголы	0.3757

[9] Составлено авторами. По горизонтальной оси в порядке возрастания отложены номера семантических групп Роже, и для каждой указано значение p в виде отрицательного логарифма. Группы с наименьшим значением p (лежащие максимально высоко на графике) отмечены метками. Красная линия отсекает уровень достоверности различий между подгруппами без учета поправки Бонферрони (p=0.05), зеленая — с учетом поправки. Точные значения p для семантических групп, по которым отмечены наиболее выраженные различия между полами, а также точные значения частот для них, перечислены в Таблице 2.

[10] Составлено авторами.

Если посмотреть, как часто разные части речи встречаются в каждом из 6 классов (категории высшего уровня в Тезаурусе Роже), то можно отметить различия на уровне крупных блоков смысловых категорий (Таблица 4): абстрактные категории почти вдвое чаще выражаются с помощью существительных, в то время как глаголы безусловно доминируют при выражении ценностей в пространственных категориях, а глаголы и прилагательные значительно превосходят существительные в частоте использования при выражении чувственных и моральных категорий.

Таблица 4. Распределение семантических групп, связанных с разными частями речи, при выражении свободных ассоциаций «с чем-то самым важным в жизни», по классам Тезауруса Роже[11]

Класс Тезауруса Роже	Существительные	Прилагательные	Глаголы
Words Expressing Abstract Relations	38%	21%	17%
Words Relating to Space	3%	5%	10%
Words Relating to the Intellectual Faculties	6%	13%	12%
Words Relating to the Sentient and Moral Powers	20%	33%	33%
Words Relating to the Voluntary Powers	21%	21%	21%
Words Relating to Matter	12%	7%	8%

Более детальный анализ смысловых различий между ценностями, выражаемыми с помощью разных частей речи, можно провести на уровне представленности отдельных секций тезауруса. Так, в отношении существительных на первом месте по сумме частот находится секция «Причинно-следственная связь» (Causation). В нее входят 14% от всех семантических групп, полученных после преобразования существительных. Для прилагательных секция-лидер — это «Личные пристрастия» (Personal Affections) с суммой частот 18%. Глаголами, как оказалось, наиболее часто выражают ценности, связанные с симпатией (Sympathetic Affections). Списки топ-10 наиболее часто представленных секций в каждой из трех частей речи собраны в Таблице 5.

Таблица 5. Топ-10 наиболее представленных секций Тезауруса Роже, полученных после преобразования массивов слов из трех разных частей речи[12]

Рейтинг	Существительные		Прилагательные		Глаголы	
1	Causation	14%	Personal Affections	18%	Sympathetic Affections	14%
2	Order	13%	Prospective Volition	7%	Personal Affections	12%
3	Organic Matter	11%	Organic Matter	6%	Organic Matter	7%
4	Sympathetic Affections	8%	Sympathetic Affections	6%	Motion	6%
5	Personal Affections	8%	Relation	6%	Prospective Volition	6%
6	Relation	7%	Means of Communicating Ideas	5%	Causation	5%
7	Antagonism	7%	Time	5%	Voluntary Action	5%
8	Prospective Volition	4%	Moral Affections	5%	Change	4%
9	Moral Affections	3%	Possessive Relations	4%	Modes of Communication	4%
10	Voluntary Action	3%	Quantity	3%	Moral Affections	4%

[11] Составлено авторами.
[12] Составлено авторами.

Обсуждение

Выше были показаны два примера количественного анализа смыслов, лежащих в структуре ценностей представителей некоторой группы молодых людей. Первый подход, описанный в первом разделе Результатов, показывает, как можно искать и находить минорные, точечные оттенки смыслов, по которым различаются две группы, которые в целом очень похожи. Молодые женщины и молодые мужчины имеют различия в частотной диаграмме слов, которые ассоциируются «с чем-то самым важным для жизни» (Таблица 1), однако перевод слов в наборы семантических групп приводит к тому, что косинусная близость двух векторов с частотами оказывается близка к 1, то есть на уровне смыслов можно утверждать, что мужчины и женщины ценят практически одно и то же. Вместе с тем у мужчин есть более частые отсылки к ценностям в области абстрактных понятий (симметрия), а женщины чаще называют ценности в области отношений — дружба, любовь, согласие — и удовольствий (Таблица 2).

Второй подход, который был применен для анализа массивов слов, сильно отличающихся по частотам отдельных семантических групп, описан во втором разделе Результатов. Он позволяет посмотреть крупными блоками, в каких зонах смыслового спектра преимущественно сосредоточены смыслы в одной и в другой выборке.

Предлагаемый здесь автоматизированный метод анализа является развитием двух других работ, выполненных на данных опросов групп молодых людей в отношении их ценностей. В работе [Андреюк и др. 2020] слова-ассоциации «с чем-то самым важным для профессионального развития человека» вручную, методом согласования экспертных оценок распределялись по 6 смысловым пакетам, связанным с получением знаний, деятельностью, упорством, креативностью, коммуникациями и морально-этическими категориями. Такая смысловая разметка позволила разделить 27 групп на 4 кластера, в которых было разное соотношение представленности каждого из 6 смысловых пакетов. Существенным недостатком работы следует считать ручной алгоритм распределения слов: из-за высокой трудоемкости такой анализ невозможно проводить массово, кроме этого, субъективный принцип экспертных суждений практически не поддается автоматизации.

В работе [Gerasimenko et al. 2021] массив ассоциаций в отношении жизненных ценностей (тот же, что использовался в данной работе) был выстроен по убыванию частот встречаемости. При этом верхние 26 слов составили примерно 80% от всех частот. Набор из этих 26 слов с учетом их индивидуальных частот использовали в качестве эталона для анализа семантического ядра текстов с описанием миссии и ценностей 6 люксовых брендов. Каждое слово из семантического ядра текста сопоставляли с «эталоном»: если слово встречалось в эталонном наборе, его частота перемножалась с частотой эталонного слова и полученная величина учитывалась с коэффициентом 1. Если слово не встречалось буквально, но имело очень близкий смысловой аналог в эталонном наборе слов, произведение частот учитывалось с коэффициентом 0.2. Все остальные слова из семантического ядра, которые не имели соответствий в эталонном списке, не учитывались вовсе. Таким образом, для каждого бренда формировался вектор из 26 позиций (по количеству слов в эталонном списке). Расчет косинусной близости между векторами позволил вычислить степень соответствия идеологии конкретного бренда ценностям данной конкретной целевой группы, сравнить бренды попарно между собой, а также построить нечто вроде штрих-кода идей бренда — степень выраженности каждой из 26 позиций вектора.

Этот подход гораздо в большей степени приближает к возможности автоматического количественного сравнения ценностей группы и отдельных наборов идей, благодаря векторизации и последующей работе с типовыми векторами. Единственным ограничением оказывается определение смысловой близости для конкретного слова (которая учитывается с коэффициентом 0.2) — ее наличие или отсутствие по-прежнему субъективно.

В данной статье было преодолено последнее препятствие на пути к автоматизации смыслового анализа ценностей благодаря использованию Тезауруса Роже. Предложенный здесь алгоритм, реализованный в программе на языке Python, полностью автоматически учитывает все оттенки смысловых отношений для каждого из слов в массиве. На выходе получается вектор с размерностью 1044, представляющий собой частоты встречаемости каждой из 1044 семантических групп Роже в исследуемом массиве слов.

Разумеется, предлагаемый подход содержит и ограничения. Прежде всего они связаны с языковыми переходами. Здесь использовался автоматический переводчик с русского на английский для того, чтобы полностью исключить субъективность в передаче смыслов. Однако сам переводчик может вносить (и наверняка вносит) определенные смысловые искажения.

Еще один источник возможных ошибок связан с тем, что используемой здесь программе удается идентифицировать в тезаурусе не все слова. Алгоритм сохраняет в отдельном файле все «непосчитанные» слова, поэтому исследователь может осознано принять решение, как с ними поступить. В данной работе мы игнорировали неразмеченные слова из соображений максимальной автоматизации, однако констатировали, что их количество ни разу не превышало 10% от исследуемого массива.

Заключение

Человечество уже довольно давно умеет делать семантический анализ на уровне слов методами дистрибутивной семантики. Здесь мы вплотную подошли к тому, чтобы автоматизированный количественный анализ можно было выполнять на уровне смыслов. Мы не первые, кто использовал Тезаурус Роже для задач семантического анализа (см., например, [Jarmasz, Szpakowicz 2012]). Среди прочего, количественный анализ кодов Роже применялся для выявления смысловых изменений в больших социальных группах [TenHouten 2016; Klingenstein et al. 2014]. Однако, насколько нам известно, до сих пор никто не использовал векторный подход к анализу частот семантических групп Роже в отношении ценностей. А сочетание автоматизации, использования типовых векторов для сравнений смыслов и вербализация ценностей группы методом свободных ассоциаций с оцениваемым предметом, на наш взгляд, откроют дорогу для поточных исследований структуры групповых ценностей. Результаты таких исследований, в свою очередь, могут существенно изменить наши представления о взаимодействии потребителей и брендов, студентов и разработчиков учебных курсов, участников больших мультикультурных проектов друг с другом и во многих других сферах контактов людей с идеями.

Список литературы:

Андреюк Д.С., Петрунин Ю.Ю., Храбровская В.Д. Метод кластеризации групп молодежи на основании ценностных смыслов в отношении профессионального развития и жизни в целом // Государственное управление. Электронный вестник. 2020. № 83. С. 221–242. DOI: 10.24411/2070-1381-2020-10117

Баева Л.В. Ценности изменяющегося мира: Экзистенциальная аксиология истории. Астрахань: Изд-во АГУ, 2004.

Митрофанова О.А. Измерение семантических расстояний как проблема прикладной лингвистики // Структурная и прикладная лингвистика. 2007. № 7. С. 92–101.

Человек в условиях глобальных рисков: социально-психологический анализ / под ред. Т.А. Нестика, А.Л. Журавлева. М: Изд-во «Институт психологии РАН», 2020.

Benjamini Y., Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing // Journal of the Royal Statistical Society: Series B (Methodological). 1995. Vol. 57. № 1. P. 289–300. DOI: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Gerasimenko V., Andreyuk D., Kurkova D. Approach for Management of Brand Positioning: Quantification of Value Matching between Brand and Target Audience // Polish Journal of Management Studies. 2021. Vol. 24. P.96–111. DOI: 10.17512/pjms.2021.24.1.06

Hochberg Y. A Sharper Bonferroni Procedure for Multiple Tests of Significance // Biometrika. 1988. Vol. 75. Is. 4. P. 800–802. DOI: https://doi.org/10.2307/2336325

Holm S. A Simple Sequentially Rejective Multiple Test Procedure // Scandinavian Journal of Statistics. 1979. Vol. 6. Is. 2. P. 65–70.

Jarmasz M., Szpakowicz S. Roget's Thesaurus and Semantic Similarity // Cornell University arXiv. 2012. DOI: https://doi.org/10.48550/arXiv.1204.0245

Klingenstein S., Hitchcock T., DeDeo S. The Civilizing Process in London's Old Bailey // Proceedings of the National Academy of Sciences. 2014. Vol. 111. Is. 26. P. 9419–9424. DOI: https://doi.org/10.1073/pnas.1405984111

Noble W.S. How Does Multiple Testing Correction Work? // Nature Biotechnology. 2009. Vol. 27. P. 1135–1137. DOI: https://doi.org/10.1038/nbt1209-1135

Ochiai A. Zoogeographical Studies on the Soleoid Fishes Found Japan and Its Neighboring Regions // Bulletin of the Japanese Society of Scientific Fischeries. 1957. Vol. 22. Is. 9. P. 526–530.

Roget P.M. Roget's Thesaurus of English Words and Phrases. Austin: MICRA, Inc., 1991.

Rothman K.J. No Adjustments Are Needed for Multiple Comparisons // Epidemiology. 1990. Vol. 1. Is. 1. P. 43–46. DOI: 10.1097/00001648-199001000-00010

Sahlgren M. The Distributional Hypothesis. From Context to Meaning // Rivista di Linguistica. 2008. Vol. 20. Is. 1. P. 33–53.

TenHouten W.D. Neurosociology // Journal of Social and Evolutionary Systems. 1997. Vol. 20. Is. 1. P. 7–37. DOI: https://doi.org/10.1016/S1061-7361(97)90027-8

TenHouten W.D. The Emotions of Powerlessness // Journal of Political Power. 2016. Vol. 9. Is. 1. P. 83–121. DOI: http://dx.doi.org/10.1080/2158379X.2016.1149308

References:

Andreyuk D.S., Petrunin Yu.Yu., Khrabrovskaya V.D. (2020) [Youth Groups Clustering Method Based on the Meanings of Value in Relation to Professional Development and Life in General. *Gosudarstvennoye upravleniye. Elektronnyy vestnik.* № 83. P. 221–242. DOI: 10.24411/2070-1381-2020-10117

Bayeva L.V. (2004) *Tsennosti izmenyayushchegosya mira: Ekzistentsial'naya aksiologiya istorii* [Values of the changing world: Existence axiology of the history]. Astrakhan': Izd-vo AGU.

Mitrofanova O.A. (2007) Estimating Semantic Distance as a Problem of Applied Linguistics. *Strukturnaya i prikladnaya lingvistika.* № 7. P. 92–101.

Nestik T.A., Zhuravlev A.L. (eds.) (2020) *Chelovek v usloviyakh global'nykh riskov: sotsial'no-psikhologicheskiy analiz* [Human under global risks: Social and psychological analysis]. Moscow: Izd-vo «Institut psikhologii RAN».

Benjamini Y., Hochberg Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological).* Vol. 57. № 1. P. 289–300. DOI: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Gerasimenko V., Andreyuk D., Kurkova D. (2021) Approach for Management of Brand Positioning: Quantification of Value Matching between Brand and Target Audience. *Polish Journal of Management Studies.* Vol. 24. P. 96–111. DOI: 10.17512/pjms.2021.24.1.06

Hochberg Y. (1988) A Sharper Bonferroni Procedure for Multiple Tests of Significance. *Biometrika*. Vol. 75. Is. 4. P. 800–802. DOI: https://doi.org/10.2307/2336325

Holm S. (1979) A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics.* Vol. 6. Is. 2. P. 65–70.

Jarmasz M., Szpakowicz S. (2012) Roget's Thesaurus and Semantic Similarity. *Cornell University arXiv*. DOI: https://doi.org/10.48550/arXiv.1204.0245

Klingenstein S., Hitchcock T., DeDeo S. (2014) The Civilizing Process in London's Old Bailey. *Proceedings of the National Academy of Sciences.* Vol. 111. Is. 26. P. 9419–9424. DOI: https://doi.org/10.1073/pnas.1405984111

Noble W.S. (2009) How Does Multiple Testing Correction Work? *Nature Biotechnology.* Vol. 27. P. 1135–1137. DOI: https://doi.org/10.1038/nbt1209-1135

Sahlgren M. (2008) The Distributional Hypothesis. From Context to Meaning. *Rivista di Linguistica*. Vol. 20. Is. 1. P. 33–53.

Roget P.M. (1991) *Roget's Thesaurus* of English Words and Phrases. Austin: MICRA, Inc.

Ochiai A. (1957) Zoogeographical Studies on the Soleoid Fishes Found Japan and Its Neighboring Regions. *Bulletin of the Japanese Society of Scientific Fischeries.* Vol. 22. Is. 9. P. 526–530.

Rothman K.J. (1990) No Adjustments Are Needed for Multiple Comparisons. *Epidemiology*. Vol. 1. Is. 1. P. 43–46. DOI: 10.1097/00001648-199001000-00010

TenHouten W.D. (1997) Neurosociology. *Journal of Social and Evolutionary Systems*. Vol. 20. Is. 1. P. 7–37. DOI: https://doi.org/10.1016/S1061-7361(97)90027-8

TenHouten W.D. (2016) The Emotions of Powerlessness. *Journal of Political Power*. Vol. 9. Is. 1. P. 83–121. DOI: http://dx.doi.org/10.1080/2158379X.2016.1149308

Дата поступления/Received: 24.01.2022