Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 515–549
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-29475

Research article

Russian dictionary with concreteness/abstractness indices

Valery D. SOLOVYEV[1] ✉, Yulia A. VOLSKAYA[1], Mariia I. ANDREEVA[1,2] and Artem A. ZAIKIN[1]

[1]Kazan (Volga region) Federal University, Kazan, Russia
[2]Kazan State Medical University, Kazan, Russia
✉maki.solovyev@mail.ru

Abstract
The demand for a Russian dictionary with indices of abstractness/concreteness of words has been expressed in a number of areas including linguistics, psychology, neurophysiology and cognitive studies focused on imaging concepts in human cognitive systems. Although dictionaries of abstractness/concreteness were compiled for a number of languages, Russian has been recently viewed as an under-resourced language for the lack of one. The Laboratory of Quantitative Linguistics of Kazan Federal University has implemented two methods of compiling dictionaries of abstract/concrete words, i.e. respondents survey and extrapolation of human estimates with the help of an original computer program. In this article, we provide a detailed description of the methodology used for assessing abstractness/concreteness of words by native Russian respondents, as well as control algorithms validating the survey quality. The implementation of the methodology has enabled us to create a Russian dictionary (1500 words) with indices of concreteness/abstractness of words, including those missing in the Russian Semantic Dictionary by N.Yu. Shvedova (1998). We have also created three versions of a machine dictionary of abstractness/concreteness based on the extrapolation of the respondents' ratings. The third, most accurate version contains 22,000 words and has been compiled with the use of a modern deep learning technology of neural networks. The paper provides statistical characteristics (histograms of the distribution of ratings, dispersion, etc.) of both the machine dictionary and the dictionary obtained by interviewing informants. The quality of the machine dictionary was validated on a test set of words by means of contrasting machine and human evaluations with the latter viewed as more credible. The purpose of the paper is to give a detailed description of the methodology employed to create a concrete/abstract dictionary, as well as to demonstrate the methodology of its application in theoretical and applied research on concrete examples. The paper shows the practical use of this vocabulary in six case studies: predicting the complexity of school textbooks as a function of the share of abstract words; comparing abstractness indices of Russian-English equivalents; assessing concreteness/abstractness of polysemantic words; contrasting ratings of different age groups of respondents; contrasting ratings of respondents with different levels of education; analyzing concepts of "concreteness" and "specificity".
Keywords: *concreteness, abstractness, digital dictionary, Russian, academic texts*

For citation:
Solovyev, Valery D., Yulia A. Volskaya, Mariia I. Andreeva & Artem A. Zaikin. 2022. Russian dictionary with concreteness/abstractness indices. *Russian Journal of Linguistics* 26 (2). 515–549. https://doi.org/10.22363/2687-0088-29475

Научная статья

Словарь русского языка с индексами конкретности/абстрактности

В.Д. СОЛОВЬЕВ[1] ✉, Ю.А. ВОЛЬСКАЯ[1], М.И. АНДРЕЕВА[1,2], А.А. ЗАИКИН[1]

[1]Казанский (Приволжский) федеральный университет, Казань, Россия
[2]Казанский государственный медицинский университет, Казань, Россия
✉maki.solovyev@mail.ru

Аннотация.
Для целого ряда исследований в лингвистике, психологии, нейрофизиологии, посвященных репрезентации концептов в когнитивной системе человека, требуется словарь с численными оценками степени конкретности/абстрактности слов. Такие словари созданы для нескольких языков, но до последнего времени не было словаря для русского языка. В лаборатории квантитативной лингвистики Казанского федерального университета подготовлено несколько вариантов такого рода словаря для русского языка. При их создании использованы две методологии: опрос респондентов и разработка компьютерных программ для экстраполяции человеческих оценок. В статье подробно описана методология оценки абстрактности/конкретности слов респондентами-носителями русского языка, а также способы контроля качества их ответов. Применение данной методологии позволило создать словарь русского языка (1500 слов) с указанием индексов конкретности/абстрактности слов, в том числе отсутствующих в Русском семантическом словаре Н.Ю. Шведовой (1998). В нашей лаборатории созданы также три версии машинного словаря абстрактности/конкретности, полученные экстраполяцией оценок респондентов. Последняя версия словаря (22 тыс. слов), составлена с применением современной технологии глубокого обучения нейронных сетей и является наиболее точной. Приведены статистические характеристики (гистограммы распределения оценок, дисперсия и др.) и машинного словаря, и словаря, полученного опросом информантов. Оценка качества машинного словаря осуществлена на тестовом множестве слов путем сопоставлением машинных оценок с человеческими. Цель данной статьи – дать подробное описание методологии создания словаря конкретности/абстрактности, а также на конкретных примерах продемонстрировать методику его применения в теоретических и прикладных исследованиях. В статье показано практическое использование данного словаря в шести конкретных исследованиях: определение сложности текстов по доле абстрактных слов (на примере школьных учебников), сравнение оценок слов и их переводных эквивалентов в английском языке, оценки конкретности/абстрактности многозначных слов, сравнение оценок разных возрастных групп респондентов, сравнение оценок респондентов с разным уровнем образования, сравнение концепций «конкретность» и «специфичность».

Ключевые слова: *конкретность, абстрактность, электронный словарь, русский язык, учебные тексты*

Для цитирования:
Соловьев В.Д., Вольская Ю.А., Андреева М.И., Заикин А.А. Словарь русского языка с индексами конкретности/абстрактности. *Russian Journal of Linguistics.* 2022. Т. 26. № 2. С. 515–549. https://doi.org/10.22363/2687-0088-29475

1. Введение

Категория абстрактности/конкретности уже десятилетия находится в центре внимания когнитивных исследований. Проблема представления конкретных и абстрактных объектов в мозгу человека представляет собой серьезный вызов всей когнитивной науке (Borghi et al. 2017). Современный подход к ее изучению начинается с фундаментальной работы (Paivio 1965). Основной подход к определению этих понятий представлен в работе (Spreen & Shulz 1966). Конкретные понятия – те, которые воспринимаются органами чувств. Примеры конкретных слов – *кошка, стул, гора*. Абстрактные понятия не воспринимаются органами чувств. Например, *ответственность, взаимоотношения, непонимание*. Схожие трактовки встречаются во многих исследованиях. Так, в работе (Schmid 2000) приводится такое определение: «abstract nouns are those nouns whose denotata are not part of the concrete physical world and cannot be seen or touched» («денотаты абстрактных существительных не принадлежат физическому миру, т.е. их нельзя увидеть или дотронуться до них»[1]). Однако данные определения сильно упрощают ситуацию, давая характеристики прототипов конкретности и абстрактности. В действительности экспериментальные исследования показали, что изучаемая категория – континуум, но не дихотомия (Mkrtychian et al. 2019). В связи с этим весьма сложно дать совершенно строгое чисто лингвистическое определение этих понятий, которое позволило бы любое слово однозначно квалифицировать как конкретное или абстрактное.

Для поддержки вышеуказанных когнитивных исследований требуются словари с индексами, характеризующими степень конкретности/абстрактности слов. Обычно словарь создается методом опроса носителей языка, которым предлагается выставить рейтинг конкретности/абстрактности заданных слов, кроме этого применяются методы машинного обучения для расширения словарей путем экстраполяции уже имеющихся рейтингов на другие слова.

Статья подводит итоги первого этапа исследований в этом направлении лаборатории квантитативной лингвистики КФУ и обобщает опыт построения первого для русского языка словаря рейтингов конкретности/абстрактности, а также результаты первых исследований на его основе. Словарь свободно доступен по адресу (ENA, April 17, 2022)[2].

2. Обзор литературы

Исследования категории конкретности/абстрактности ведутся широким фронтом от психологии и психолингвистики до нейрофизиологии и

[1] Здесь и далее перевод выполнен авторами статьи.
[2] https://kpfu.ru/tehnologiya-sozdaniya-semanticheskih-elektronnyh.html.

медицины. Опубликованы тысячи статей, свежие обзоры можно найти в (Mkrtychian et al. 2019, Solovyev 2021). В нейрофизиологии изучался вопрос локализации понятий абстрактности/конкретности. Во многих экспериментах с помощью техники нейровизуализации было показано, что конкретные и абстрактные слова репрезентируются в разных нейроанатомических структурах мозга.

В психологических исследованиях установлен так называемый «эффект конкретности», демонстрирующий большую легкость обработки конкретных слов в человеческом сознании. Конкретные слова лучше запоминаются (Schwanenflugel et al. 1992), лучше распознаются (Fliessbach et al. 2006), быстрее читаются (Schwanenflugel & Shoben 1983), быстрее усваиваются (Mestres-Missé et al. 2014). Для таких слов легче написать словарные толкования, и они будут более детальными (Sadoski 1997). Респонденты легче продуцируют ассоциации в ответ на конкретные слова-стимулы (de Groot 1989).

В современной науке предложены две основные теории репрезентаций конкретной/абстрактной лексики: двойного кодирования (dual coding theory) (Paivio 1990) и доступного контекста (context availability theory) (Schwanenflugel & Shoben 1983). Теория двойного кодирования постулирует существование 2-х систем памяти: образной и словесной, причем образная система, в отличие от словесной, обеспечивает кодирование только конкретной информации. Согласно теории доступного контекста конкретные и абстрактные слова различаются количеством и силой ассоциативных связей. В русле теории доступного контекста было показано, что конкретные слова активируют более широкий вербальный контекст и поэтому обрабатываются быстрее, но не получают доступа к системе обработки изображений. В целом, исследователи соглашаются в отношении способов репрезентации конкретных, но не абстрактных понятий. Когнитивная лингвистика предлагает свой подход (Kousta et al. 2011) к репрезентации абстрактных понятий: согласно гипотезе воплощенности (embodied abstract semantics), эмоциональный опыт имеет критическое значение для репрезентации и обработки абстрактных слов.

Понятия конкретности и абстрактности сами по себе являются предметами изучения в лингвистике достаточно давно, однако в последнее время с появлением больших корпусов текстов и обширных лексических онтологий появились принципиально новые идеи исследований и результаты. К числу наиболее интересных можно отнести следующие. В работе (Snefjella et al. 2019) показано, что с течением времени степень конкретности слов возрастает. В статье (Reilly & Desai 2017) описано, что плотность множества семантически близких слов выше для конкретных слов, нежели для абстрактных. В работе (Naumann et al. 2018) замечено, что в корпусах текстов абстрактные слова чаще встречаются вместе с абстрактными, а конкретные – с конкретными. В работе (Ivanov & Solovyev 2021) проведено сопоставление категорий конкретности и специфичности, показано их существенное различие.

Исследования конкретности/абстрактности имеют различные прикладные аспекты. В медицине абстрактные слова играют важную роль в ходе терапии больных с афазией (Dallin et al. 2020). Доля абстрактных слов является одним из значимых показателей сложности текстов (Sadoski et al. 2001, McNamara et al. 2014). Этот параметр включен в доступные онлайн-пакеты расчета сложности текстов (Coh-Metrix). Автоматически оцененная доля абстрактных слов вместе с другими параметрами может быть использована в педагогике для оценки сложности текстов с целью адекватного выбора образовательных материалов.

Для проведения психологических, нейрофизиологических экспериментов нужны списки слов с оценками степени их конкретности/абстрактности. Далее в статье как синоним слова *оценка* будет использоваться и слово *рейтинг*. Оценки получаются методом опроса носителей языка, в результате которого составляется словарь с рейтингами абстрактности/конкретности слов. Для английского языка первый крупный словарь такого рода создан в 1981 г. (Coltheart 1981). Он содержит почти 4 тыс. слов и свободно доступен в составе психолингвистической базы данных MRC (ENA, April 17, 2022)[3]. Позднее был создан словарь, включающий почти 40 тыс. слов (Brysbaert et al. 2014a). Каждое слово получает не менее 25 оценок респондентов по 5-балльной шкале, которые усредняются. Кроме английского языка сравнимый по объему словарь создан лишь для нидерландского (Brysbaert et al. 2014b). Очевидной проблемой является большая трудоемкость составления подобных словарей. Для немецкого языка словарь (Maximilian & Walde 2016) содержит лишь 4 тыс. слов. Недавно опубликована база данных с рейтингами конкретности/абстрактности для хорватского языка на 6 тыс. слов (Peti-Stantić et al. 2021). Аналогичные словари созданы для итальянского (Vergallito et al. 2020), китайского (Yao et al. 2017) и ряда других языков.

В связи с большой трудоемкостью построения словаря путем проведения опросов актуальной является задача создания компьютерных словарей методом автоматической экстраполяции человеческих оценок, полученных на небольшом множестве слов, на большее множество. Основная идея экстраполяции человеческих оценок на ранее неоцененные слова состоит в использовании векторной семантики слов (Mikolov et al. 2013), построенной на базе большого корпуса текстов, и получении новых оценок на основе семантической близости слов в построенном семантическом пространстве. Таким образом, необходимым условием создания в определенном языке компьютерного словаря является существование большого корпуса текстов, на основе которого можно строить векторную семантику.

Принципиально важной является оценка качества машинных словарей. Они оцениваются путем сравнения со словарями, созданными на основе опросов, с вычислением коэффициента корреляции двух словарей, чаще всего по Спирмену. К настоящему времени лучший достигнутый результат –

[3] https://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm

машинный словарь для английского языка работы (Charbonnier & Wartena 2019), он имеет коэффициент корреляции со словарем на основе опросов 0,900. Словарь создан с использованием технологии fastText (Joulin et al. 2016) для построения семантического пространства и SVM (Cristianini & Shawe-Taylor 2000) в качестве классификатора. Экстраполяция человеческих оценок в данном словаре осуществляется путем кросс-валидации на 40-тысячном словаре (Brysbaert et al. 2014a). В (Brysbaert, Warriner & Kuperman 2014a) было проведено сопоставление двух словарей на основе опроса респондентов, и оказалось, что коэффициент корреляции между ними равен 0,919. Т.е. результат 0,900 – почти предельно возможный.

3. Словарь с индексами конкретности/абстрактности для русского языка

3.1. Построение словаря

Словарь с индексами конкретности/абстрактности для русского языка, включающий 1,5 тыс. слов, создан методом опроса респондентов в Казанском федеральном университете. Для оценки взяты наиболее частотные существительные из известного частотного словаря О.Н. Ляшевской, С.А. Шарова (Ляшевская & Шаров 2009). Слова предлагались респондентам в виде анкет (гугл-форм) по 50 слов в каждой. Мы считаем, что наш опрос проведен более тщательно по сравнению с опросом для английского языка. Дело в том, что анкеты для английского содержали по 300 слов (Brysbaert et al. 2014a, 2014b). Естественно ожидать, что к концу столь длинного списка слов концентрация внимания респондентов падает, и доля ошибочных оценок должна возрастать.

В проведенном исследовании не было установлено ограничение времени на заполнение анкет респондентами. Для оценки конкретности/абстрактности слов авторы руководствовались исследованием (Laming 2004) и использовали 5-балльную шкалу Ликерта, где 1 маркирует высокую степень конкретности, а 5 – высокую степень абстрактности. Исследования, проведенные на материале английского языка (Colheart 1981), основывались на 7-балльной шкале от 1 (абстрактные) до 7 (конкретные). Полученные оценки слов масштабировались с коэффициентом 100. Таким образом, была получена шкала от 100 до 700. Для возможности последующего сравнения данных обоих языков произведена перенормировка наших данных по формуле y = 100 * (1.5 * (5-x)) +1), где x – значение оценки конкретности для русского языка (Solovyev et al. 2019b). Таким образом рассчитаны другие значения рейтинга, удобные для сопоставления с английским рейтингом: наивысшее значение конкретности маркировалось 700 единицами, наивысшее значение абстрактности – 100. В дальнейшем в разных исследованиях мы использовали различные шкалы – и от 1 до 5, и от 100 до 700.

Опрос был разделен на две части. В первой (Solovyev et al. 2019a) в качестве респондентов участвовали около 400 студентов (от 17 до 25 лет) очной

формы обучения Казанского федерального университета и около 300 студентов Белорусского государственного педагогического университета, носителей русского языка (Zhuravkina et al. 2020). В этой части получены оценки для 1000 слов, для каждого слова – не менее 40 оценок.

Во второй части опроса (500 слов) (Вольская 2020) использовалась система Яндекс.Толóка, в экспериментах могли принять участие все желающие. Во второй части для каждого слова опрашивалось 60 человек. В первой части какие-либо дополнительные инструкции участникам не давались. Однако в аналогичном построении словарей для английского языка респонденты подробно инструктировались (Brysbaert et al. 2014a, 2014b). Поэтому во второй части перед прохождением опроса мы также давали респондентам детальные инструкции, причем максимально близкие к приведенным в зарубежных работах. В них были даны подробные описательные определения абстрактных и конкретных слов с примерами. Определения в духе тех, что приведены во введении, опирались на возможность восприятия слов органами чувств. Далее подчеркивалось, что некоторые слова могут сочетать в себе как признаки конкретности, так и абстрактности; приводилось описание принципа оценки слов (указывалось соответствие числовых значений степени проявления абстрактности). Приведем фрагмент пояснений: «слово “любовь” более абстрактно, так как означает некое отвлеченное понятие, которое лишено физической очерченности, а вот слово “стол” – более конкретно, это реальный предмет, который можно потрогать, увидеть и т.д».

При использовании системы Яндекс.Толóка респондентам было необходимо указать возраст, пол, уровень образования (среднее, средне специальное, неоконченное высшее, высшее, высшее филологическое). После этого пользователям открывался доступ к оценке лексических единиц. Система предоставляет возможность отбирать респондентов по уровню образования, возрасту, родному языку, их квалификации, судя по предыдущей работе в Яндекс.Толóка. Мы выделили две возрастные группы: от 18 до 30 и от 31 до 55 лет. Допускались лишь те респонденты, для которых русский язык является родным. Также данный опрос могли проходить только лучшие исполнители сервера, число которых составляет 20% от общего количества зарегистрированных на Яндекс.Толóка участников.

В рамках первой части опроса для перепроверки оценок для 100 слов получены дополнительные оценки других участников эксперимента. Коэффициент корреляции для этих двух независимых оценок оказался равен 0,879. Аналогичное сравнение двух вышеупомянутых экспериментов для английского дало коэффициент корреляции 0,919. Несколько более низкий результат у нас можно объяснить тем, что в этой части эксперимента мы, в отличие от опросов для английского языка, не давали респондентам четких определений конкретности/абстрактности. В итоге впервые создан словарь слов русского языка с численными оценками конкретности/абстрактности, полученными опросом респондентов.

3.2. Коррекция исходных данных. Статистика

В ходе визуальной проверки оценок при проведении первого опроса был выявлен ряд недобросовестных респондентов, например таких, которые оценили все слова одним и тем же баллом. В связи с этим возникла проблема очистки собранных данных от мусора. На основе работы (Chandola et al. 2009) реализованы 5 способов очистки данных.

1. Расчет автокорреляции оценок респондента. Слишком высокий уровень автокорреляции первого порядка указывает на несерьезное или по меньшей мере недостаточно вдумчивое отношение к эксперименту. Удаляются ответы респондентов, у которых оценки выходят за рамки стандартного распределения.

2. Расстояние от вектора оценок респондента до вектора средних оценок. Расстояние измерялось по манхэттенской метрике (Black 2019). Респонденты, оценки которых слишком отличались от средних, исключались из исследования.

3. Совпадение результатов двух и более респондентов. Если у двух респондентов оценки полностью совпадали, то результаты одного из них отбрасываются.

4. Алгоритм иерархической кластеризации с одиночным связыванием применялся к множеству векторов оценок респондентов. Далее выделялись кластеры, слишком далеко отстоящие от остальных, и также удалялись.

5. В каждом опросе (50 слов, оценки не менее 40 респондентов) отбиралось одно слово с наименьшей средней оценкой и одно слово с наибольшей оценкой. Удаляются результаты тех респондентов, которые оценили слово обратным образом – 5 баллами или 1 баллом соответственно (что могло быть связано с простой ошибкой в полюсах семантического дифференциала).

При создании словаря реализован жесткий подход, при котором удалялись не только явные выбросы, но и все сомнительные случаи. В итоге удалено около четверти всех результатов. Таким образом мы получили около 30 оценок для каждого слова. Отметим, что при создании словаря для английского языка для каждого слова исходно предполагалось получить не менее 30 оценок, однако после аналогичного отбрасывания оценок недобросовестных респондентов для ряда слов количество оценок уменьшилось до 25 (Brysbaert et al. 2014a, 2014b). Гистограмма данных, оставшихся после удаления ошибочных, приведена на рис. 1. Большинство оценок приходится на интервал от 1,4 до 3,6 с некоторым преобладанием оценивания слов как конкретных. Среднее значение равно 2,5. Выделяются три пика: скорее конкретных слов, скорее абстрактных и промежуточных.

Разности оценок до и после очистки распределены по нормальному закону (рис. 2). Большинство разностей по абсолютной величине не превышает 0,2. p-значение критерия Шапиро–Уилка равно 0,576. Коэффициент корреляции Пирсона между средними оценками исходными и очищенными составил 0,978. Таким образом, очистка практически не повлияла на конечный результат.

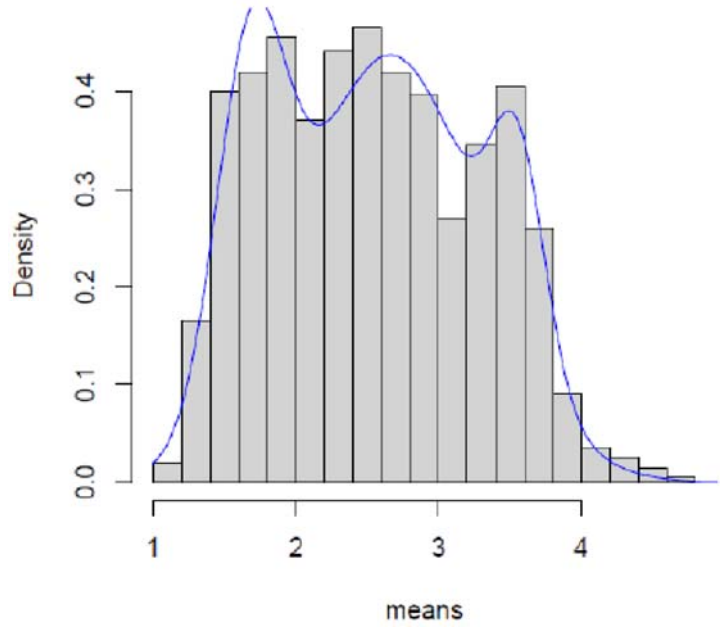

Рис. 1. Гистограмма распределения оценок / Fig. 1. Histogram of ratings distribution

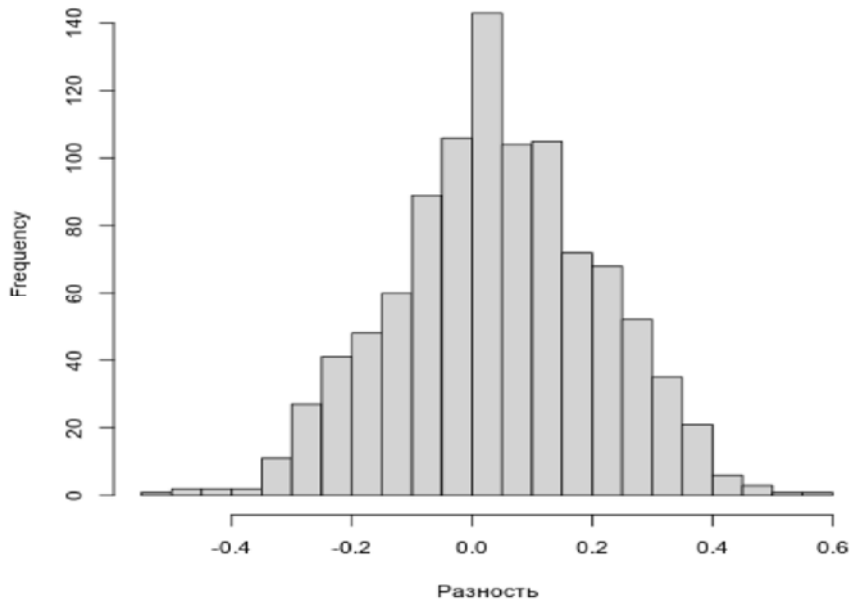

Рис. 2. Распределение разностей оценок до и после очистки /
Fig. 2. Distribution of ratings difference prior to and after filtration

На рис. 3 приведена гистограмма распределения дисперсии оценок. Для большинства слов дисперсия находится в пределах от 0,8 до 1,4. Мы провели сравнение дисперсии оценок для конкретных и абстрактных слов. Для выделения конкретных и абстрактных слов все слова упорядочены по рейтингам

и разделены на три равные по величине части, причем часть с промежуточными рейтингами не рассматривается. Средняя дисперсия конкретных слов равна 0,9, абстрактных – 1,15. Для абстрактных слов оценки респондентов имеют больший разброс, т.е. респонденты чаще оценивают их по-разному. Это хорошо согласуется с результатом недавнего психологического исследования (Wang & Bi 2021), показавшем, что респонденты указывают больше значений абстрактных слов, чем конкретных.

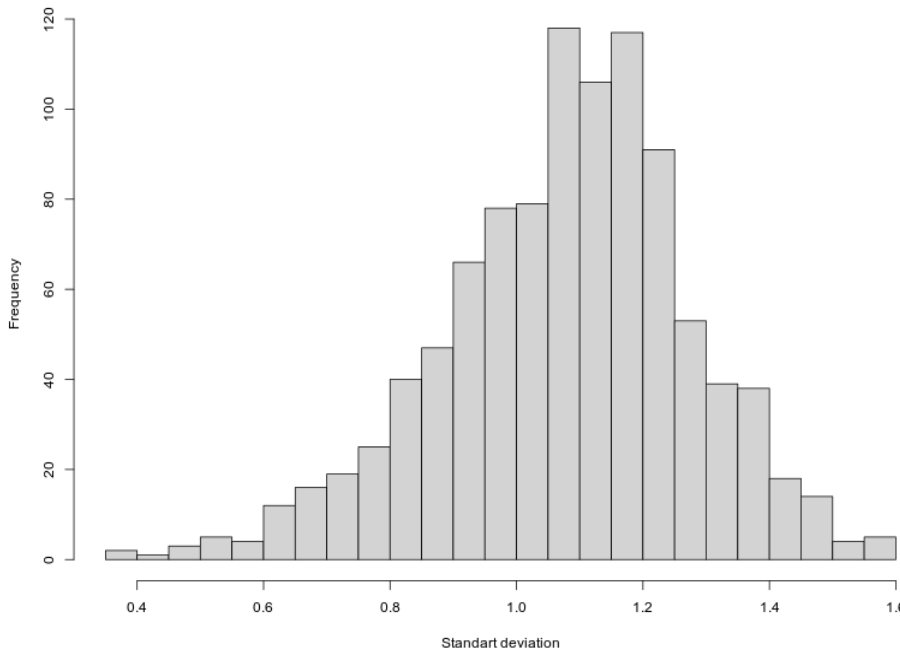

Рис. 3. Гистограмма распределения дисперсии оценок
Fig. 3. Histogram of distribution of ratings dispersion

На рис. 4 приведен график рассеяния с наложенной линией, скользящего среднего. Стандартное отклонение уменьшается при малых и больших значениях конкретности и достигает максимума для приблизительно средних значений.

На рис. S1 в Приложении приведен график распределения оценок, упорядоченных по величине. Характерные распределения оценок для типичных конкретного и абстрактного слова приведены на рис. S2 в Приложении.

В ходе второго опроса в настройках Яндекс.Толóка применялась отложенная приемка, что позволяло оценить ответы пользователей в соответствии с критериями контроля качества и в случае необходимости отклонить ответы тех из них, которые нарушали установленные правила. Для контроля качества прохождения опросов использовались следующие критерии, поддерживаемые сервисом Яндекс.Толóка и аналогичные используемым в работах (Brysbaert et al. 2014a, 2014b):

1) в каждый список было включено 10 контрольных слов. Это наиболее частотные единицы, которые уже были оценены ранее и которые демонстрируют весь диапазон проявления степени конкретности/абстрактности: *дверь, рука, книга, машина, место, слово, часть, сила, возможность, отношение*. Если при анализе ответов была обнаружена слабая корреляция между оценками контрольных слов данного пользователя со средними оценками, полученными в ходе первого опроса, ответы данного пользователя отклонялись;

2) не принимались ответы с единообразием оценок;

3) если пользователь выполнял задание быстрее, чем за установленное минимальное время – 4 минуты, то его ответы отклонялись автоматически.

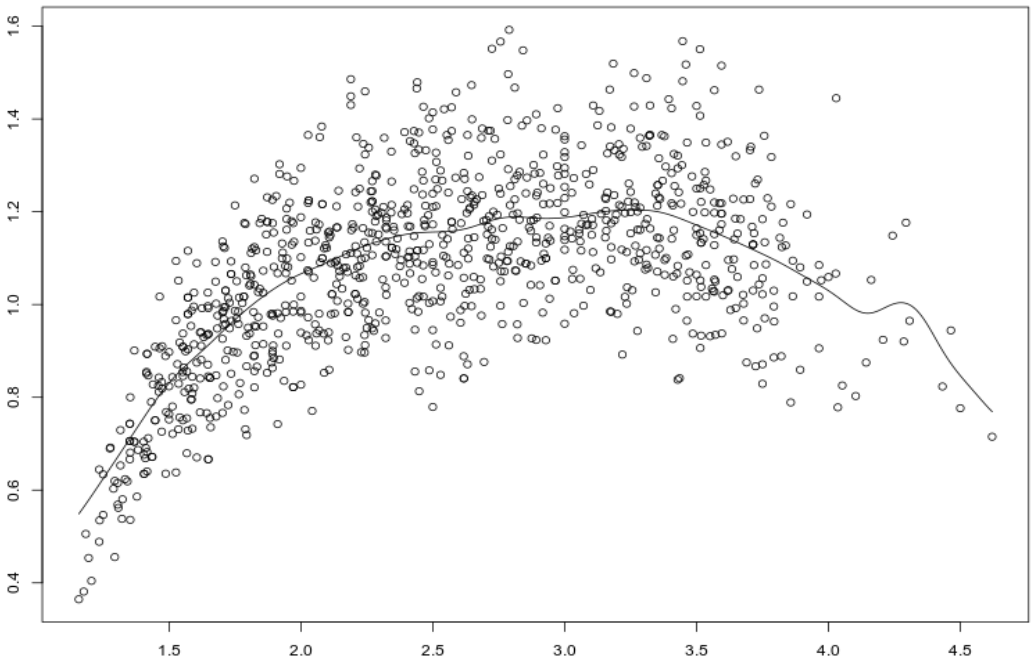

Рис. 4. Распределение дисперсий с наложенной линией скользящего среднего
Fig. 4. Dispersions distribution with superimposed moving average line

3.3. Машинный словарь

Для увеличения количества слов с рейтингами абстрактности/конкретности создан также компьютерный словарь. Причем он существует в трех вариантах. Первый – словарь словоформ (а не лемм) (Solovyev et al. 2019b), содержащий 88 тыс. словоформ существительных и прилагательных, созданный на материале корпуса Google Books Ngram (https://books.google.com/ngrams). При его составлении реализован оригинальный метод, основанный на идее, что конкретные слова встречаются в текстах чаще вместе с конкретными, а абстрактные – вместе с абстрактными. Второй вариант компьютерного словаря создан по технологии

word2vec, модель fastText (ENA, April 17, 2022)[4]. Он содержит 64 тыс. слов (лемм) (Solovyev et al. 2020a). Третий вариант – машинный словарь на 22 тыс. слов, построенный на технологии глубокого обучения, модель BERT (Devlin et al. 2018). Все варианты машинных словарей доступны по адресу (ENA, April 17, 2022)[5].

Оценка качества словарей показала, что наиболее высокий уровень корреляции между машинным и рейтингом респондентов у третьего словаря, созданного при помощи технологии BERT – 0.81 по Спирмену. Это заметно ниже результата, зафиксированного для английского языка – 0,90. Вероятно, это связано с тем, что качество русскоязычной версии BERT ниже англоязычной. В работе (Peti-Stantić et al. 2021) в аналогичном исследовании для хорватского языка полученный результат также оказался заметно хуже итогов сопоставления англоязычных словарей. Однако если ограничиться построением рейтингов высокочастотных слов (встречающихся в Google Books Ngram не менее 1 млн раз), то, как показано в работе (Solovyev et al. 2020b), точность предсказания рейтингов значительно возрастает, примерно до 0,86–0,87 по Спирмену.

4. Анализ словаря

В этом разделе статьи мы проанализируем полученные на основе ответов респондентов данные. Оценки будут рассмотрены под различными углами зрения.

4.1. Сопоставление со словообразовательным критерием

Одним из известных признаков абстрактности слова является наличие в нем определенных суффиксов. К ним относятся следующие: -изм, -аж, -итет (м.р.); -б-а, -от-а, -изн-а, -ин-а, -иц-а, -ура, -к-а, -аци-я, -н-я, -отн-я, -щин-а, -чин-а, -ость, -есть, -ность, -емость, -имость (ж.р., 3 скл.); -ие, -ье (-ьё), -ние, -нье (-ньё), -тие, -тье (-тьё), -ств-о, -еств-о, -тельств-о, -овств-о (ср.р.) (Виноградов 2001).

Из 1500 слов словаря были отобраны 150 слов, которые были оценены респондентами как наиболее абстрактные, и 150 слов, оцененные как наиболее конкретные. Это слова с рейтингами от 3,5 до 5 и от 1 до 2 соответственно. В результате оказалось, что у 94 слов (примерно две трети) с абстрактным значением такие суффиксы присутствуют. Таким образом, наличие суффикса абстрактности является хорошим критерием определения абстрактности слова, но все же он не охватывает примерно треть абстрактных слов.

Из 150 слов с наибольшим индексом конкретности (по данным наших опросов) у 19 слов присутствуют суффиксы абстрактности. Данные имена существительные классифицируем в 4 группы. Во-первых, суффикс -ени-е

[4] https://fasttext.cc/

[5] https://kpfu.ru/tehnologiya-sozdaniya-semanticheskih-elektronnyh.html.

обнаружен в морфемной структуре слов «*стихотворение*» и «*растение*». Однако данный суффикс указывает на абстрактность существительных только в том случае, если оно имеет процессуальное значение и образовано от глагольных основ. В данном случае существительное «*растение*» по словообразовательным признакам нельзя отнести к ЛГР абстрактных существительных. Слово «*стихотворение*» можно трактовать как спорный случай, хотя Семантическим словарем под редакцией Н.Ю. Шведовой (Шведова 1998) оно трактуется как абстрактное.

Во-вторых, в составе двух единиц: *лекарство* и *агентство* – выделяется суффикс -ств-о, указывающий на абстрактность, если производное существительное образовано от имени прилагательного. Указанные лексемы образованы от имен существительных. В-третьих, в одном слове, *прокуратура*, выделяется суффикс -ур-а. В данном случае суффикс указывает на собирательное значение рассматриваемого существительного: «система органов, осуществляющих от имени государства высший надзор за соблюдением законодательства».

В-четвертых, суффикс -к-а выявлен у 14 существительных. Однако данный суффикс может указывать на абстрактность только в том случае, если существительное со значением «действие» образовано от глагола. Из 14 слов только 2 удовлетворяют данному критерию: это существительные «*поставка*» и «*разведка*». В Семантическом словаре Н.Ю. Шведовой слово «*разведка*» в двух значениях трактуется как конкретное и в одном значении – как абстрактное. Существительные «*улыбка*» и «*записка*» также образованы от глаголов, но имеют значение «результат действия, которое указано производящей основой». 10 единиц являются производными от основ имен существительных.

Итак, из 19 слов с высокой степенью конкретности, имеющих суффиксы абстрактности, только два слова по словообразовательным признакам можно отнести к действительно абстрактным существительным (*поставка, стихотворение*).

4.2. Сравнение с данными словаря Н.Ю. Шведовой

Одним из немногих словарей русского языка, содержащим информацию о конкретности/абстрактности лексем, является Семантический словарь под редакцией Н. Ю. Шведовой. Его первый и второй тома посвящены конкретным существительным, третий – абстрактным. Следует отметить, что отглагольные существительные не включены в опубликованную часть словаря. В Семантическом словаре 115 слов из 150, рассмотренных в предыдущем разделе, также классифицируются как абстрактные существительные. 22 слова имеют и абстрактные, и конкретные значения, т.е. присутствуют и во втором, и в третьем томах. 13 слов в словаре Н.Ю. Шведовой отсутствуют (*проведение, осуществление, распространение, ведение, обслуживание, выполнение,*

изучение, основное, принятие, рассмотрение, снижение, увеличение, эффективность). Все они, кроме адъективного существительного *основное*, являются отглагольными существительными.

Из 150 слов с высокими значениями рейтинга конкретности, 144 слова по словарю Н.Ю. Шведовой во всех, либо в некоторых своих значениях являются конкретными 6 слов считаются абстрактными во всех своих значениях: *стихотворение, матч, улыбка, поставка, неделя, надпись.* Если обратиться к сенсорному критерию, предполагающему, что конкретные сущности воспринимаются органами чувств, то отнесение слов *улыбка* и *надпись* к абстрактным можно оспаривать. Таким образом, следует отметить очень высокую степень согласия словаря Шведовой с результатами опроса респондентов. Лишь в 1,3% случаев (*матч, неделя, поставка, стихотворение*) решение Н.Ю. Шведовой и оценки респондентов расходятся.

4.3. Сравнение оценок, полученных в результате опросов респондентов двух возрастных групп

В зарубежных исследованиях возраст респондентов никак не учитывался. Представляется интересным выяснить, есть ли заметные расхождения в оценках конкретности/абстрактности респондентами разных возрастов. Как указывалось ранее, вся выборка респондентов в нашем исследовании разделена на две группы двух возрастных категорий – от 18 до 30 лет (первая группа) и от 31 до 55 лет (вторая группа). В обеих группах оценивались одни и те же слова и респонденты находились в равных условиях. В ходе анализа полученных данных значительных расхождений между ответами обнаружено не было. Коэффициент корреляции Спирмена между оценками обеих групп является очень высоким – 0,933.

Следующая диаграмма (рис. 5) наглядно демонстрирует высокую степень корреляции оценок двух возрастных групп. На диаграмме точками представлены слова, по оси X размещены оценки второй группы, по оси Y – первой.

Разница между оценками варьируется от –1,4 до 1,2. Наибольшая отрицательная разница (от –0,5 до –1,4) обнаружена между оценками 25 слов, приведенных в табл. 1, получивших по ответам респондентов первой возрастной группы оценки от 1,5 до 3,9, а по оценкам респондентов второй группы – от 2,1 до 4,4.

Наибольшая положительная разница выявлена между оценками 20 слов (табл. 2). По ответам респондентов первой возрастной группы данные слова получили рейтинги сот 1,76 до 4,3, по ответам пользователей второй группы – от 1,26 до 3,56.

Среди ответов респондентов первой группы выявлены три лексемы с высокой степенью конкретности (от 1,76 до 2,36), 9 слов со срединным значением (от 2,5 до 3,13) и 8 слов с высокой степенью абстрактности (от 3,46

до 4,3). Среди ответов респондентов второй группы выявлено 11 лексем с высокой степенью конкретности (от 1,26 до 2,4), 8 слов со срединным значением (от 2,46 до 3,46) и слово с высокой степенью абстрактности (3,56). В целом установлено, что возраст (в рассмотренном нами диапазоне) не оказывает заметного влияния на оценку конкретности/абстрактности.

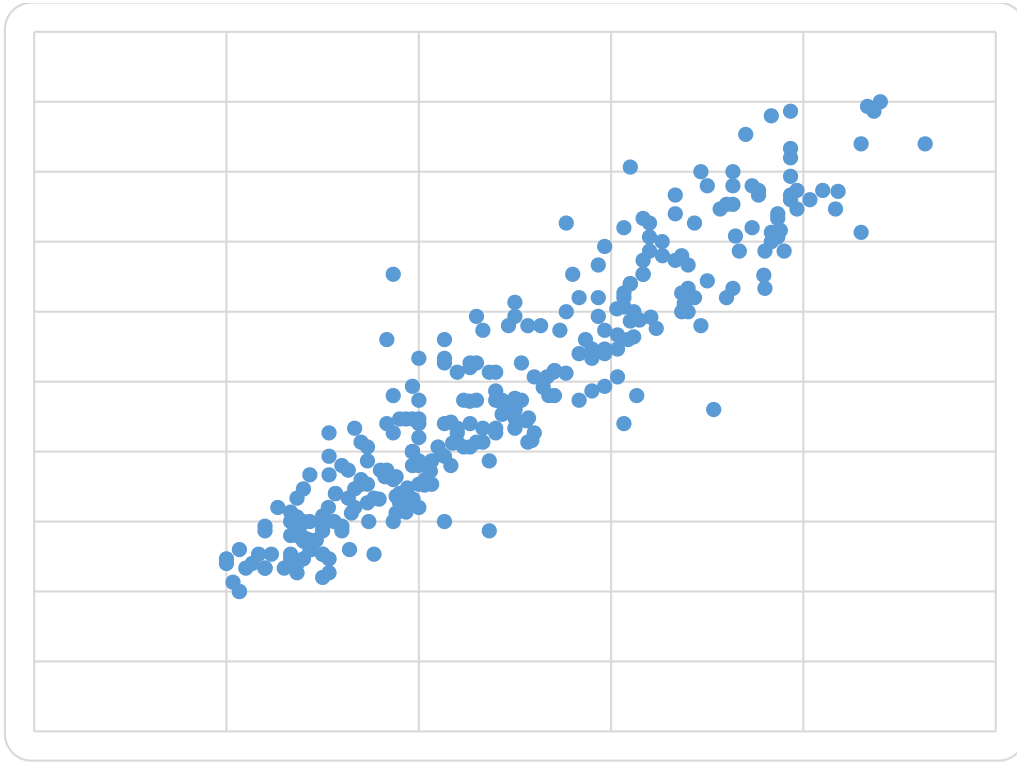

Рис. 5. Диаграмма оценок по двум возрастным группам /
Fig. 5. Ratings plot based on two age groups

Таблица 1. **Слова с наибольшей отрицательной разницей оценок**

Слово	Оценки первой группы	Оценки второй группы	Слово	Оценки первой группы	Оценки второй группы
разведка	1,533	2,133	критерий	1,867	3,267
отчет	1,667	2,167	охота	2,8	3,267
узел	1,867	2,4	дар	2,967	3,467
агент	1,967	2,467	нагрузка	3,067	3,6
указ	2,133	2,633	тариф	2,767	3,633
лекция	2	2,667	методика	3,167	3,667
знакомство	2,133	2,667	намерение	3,333	3,833
интервью	1,833	2,8	глупость	3,467	4
справка	2,133	2,8	жалоба	3,1	4,033
наказание	2,333	2,867	концентрация	3,7	4,267
свадьба	2,3	2,967	возможность	3,833	4,4
воскресенье	2,5	2,967	страдание	3,933	4,433
статистика	2,5	3,067			

Table 1. **Words with major negative difference of ratings**

Word	Group 1 ratings	Group 2 ratings	Word	Group 1 ratings	Group 2 ratings
scouting	1,533	2,133	criterion	1,867	3,267
report	1,667	2,167	hunt	2,8	3,267
knot	1,867	2,4	gift	2,967	3,467
agent	1,967	2,467	exertion	3,067	3,6
decree	2,133	2,633	rate	2,767	3,633
lecture	2	2,667	method	3,167	3,667
acquaintance	2,133	2,667	intention	3,333	3,833
interview	1,833	2,8	dullness	3,467	4
inquiry	2,133	2,8	complaint	3,1	4,033
ordeal	2,333	2,867	concentration	3,7	4,267
wedding	2,3	2,967	opportunity	3,833	4,4
Sunday	2,5	2,967	suffering	3,933	4,433
statistics	2,5	3,067			

Таблица 2. **Слова с наибольшей положительной разницей оценок**

Слово	Оценки первой группы	Оценки второй группы	Слово	Оценки первой группы	Оценки второй группы
шар	1,767	1,267	добыча	2,900	2,433
туман	2,367	1,433	питание	2,967	2,467
салон	2,133	1,500	символ	3,033	2,533
съемка	2,567	2,067	оборот	3,467	2,900
задание	2,588	2,080	перемена	3,600	3,100
совещание	2,600	2,133	карьера	3,633	3,167
раздел	3,067	2,200	секрет	3,800	3,167
темп	3,533	2,300	напряжение	3,794	3,260
свидетельство	2,833	2,367	осуществление	3,900	3,433
стандарт	3,133	2,400	восстановление	4,300	3,567

Table 2. **Words with major positive difference of ratings**

Word	Group 1 ratings	Group 2 ratings	Word	Group 1 ratings	Group 2 ratings
ball	1,767	1,267	prey	2,900	2,433
fog	2,367	1,433	nourishment	2,967	2,467
hall	2,133	1,500	symbol	3,033	2,533
filming	2,567	2,067	turn	3,467	2,900
task	2,588	2,080	change	3,600	3,100
meeting	2,600	2,133	career	3,633	3,167
section	3,067	2,200	secret	3,800	3,167
tempo	3,533	2,300	strain	3,794	3,260
certificate	2,833	2,367	implementation	3,900	3,433
standard	3,133	2,400	restoration	4,300	3,567

4.4. Сравнение оценок, полученных в результате опросов респондентов с разным уровнем образования

В ходе сбора данных посредством сервиса Яндекс.Толóка сохранялись сведения об уровне образования респондентов. Они были использованы для проверки гипотезы о том, что уровень образования может влиять на

вариативность оценок: чем больше значений слова известно респондентам данной группы, тем больше вариативность выбранных оценок по этому слову. Таким образом, предполагалось, что чем выше уровень образования, тем больше значений слова известно пользователю, следовательно, разброс оценок будет шире в группе респондентов с высшим образованием.

Нами был проведен сравнительный анализ оценок пользователей со средним специальным и средним общим образованием (группа 1) с оценками пользователей с высшим и неоконченным высшим образованием (группа 2). На 300 словах второй части опроса было рассчитано среднеквадратическое отклонение оценок респондентов первой и второй группы. У половины слов отклонение оказалось больше у первой группы, у другой половины – у второй. Для первой группы среднее квадратичное отклонение равно 1,039, для второй – 1,046. На рис. 6 приведена диаграмма дисперсии, у которой по оси X размещено среднее отклонение слов у первой группы, по оси Y – у второй. За исключением нескольких выбросов, все остальные точки укладываются вдоль главной диагонали, коэффициент корреляции Пирсона – 0,687. Таким образом, сколько-нибудь значительного различия в дисперсии оценок в зависимости от уровня образования не выявлено.

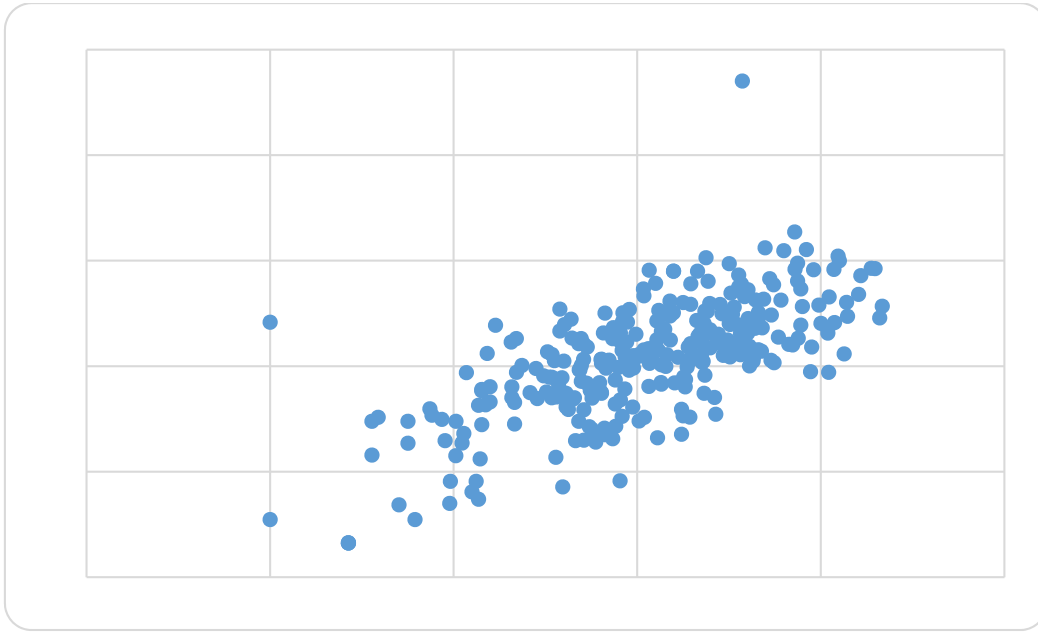

Рис. 6. Дисперсия оценок респондентов по уровням образования /
Fig. 6. Ratings dispersion based on respondents' education

4.5. Сопоставление рейтингов конкретности/абстрактности русских и английских слов

Наличие словарей с рейтингами конкретности/абстрактности для разных языков позволяет провести исследование того, в какой мере концепция конкретности/абстрактности является языково-специфической. В работе

(Solovyev, Ivanov & Akhtyamov 2019a) впервые проведено такое межъязыковое сопоставление: слова из нашего словаря сопоставлены с их английскими эквивалентами и сравнены соответствующие рейтинги. В исследовании использована американская психолингвистическая база MRC (Coltheart 1981), в которой был осуществлен поиск англоязычных аналогов исследуемых русских слов. Межъязыковое сопоставление проведено для 770 слов (из 1000 слов первого опроса). 230 слов из нашего словаря не вошли в сопоставительное исследование по следующим причинам. 1) Отсутствие эквивалентов в английском словаре. Это не только слова, обозначающие этнокультурные реалии, такие как: *милиционер, дача*, но также и названия месяцев, дней недели, по какой-то причине не включенных в MRC. 2) Вторую группу слов составили однозначные слова, которым в английском языке соответствуют разные понятия. Например, *монастырь – monastery* (букв. мужской монастырь) и *convent* (букв. женский монастырь). Рейтинги конкретности/абстрактности русских слов по вышеприведенной формуле конвертированы в формат базы MRC: от 100 (абстрактные) до 700 (конкретные). Фрагмент сопоставления представлен в табл. 3.

Таблица 3. **Рейтинги русских слов и их английских эквивалентов**

#	Слово (рус)	Рейтинг (рус.)	Рейтинг (англ.)	Разница рейтингов	Слово (англ.)
1	сила	340	339	1	strength
2	дерево	606	604	2	tree
3	эффект	288	295	7	effect
	…	…	…	…	…
771	администрация	599	231	268	administration

Table 3. **Russian-English ratings**

#	Russian word	Rating (Rus.)	Rating (Eng.)	Rating difference	English word
1	sila	340	339	1	strength
2	derevo	606	604	2	tree
3	effekt	288	295	7	effect
	…	…	…	…	…
771	administratcija	559	331	268	administration

Коэффициент корреляции Пирсона между рейтингами конкретности/абстрактности русских и английских слов следует признать высоким, он составил 0,78 (Evans 1996). По итогам сопоставления высокая степень различий (выше 67%) рейтингов обнаружена у 46 существительных. При сравнении разницы оценок абстрактных и конкретных слов обнаружено, что большее различие характерно для абстрактных слов. Для проведения такого сравнения слова русского словаря разбиты на 3 равные по величине группы – наиболее конкретных, наиболее абстрактных и слов с промежуточными рейтингами. Для наиболее конкретных слов средняя разница в оценках составила 47 единиц, а для наиболее абстрактных – 56.

Обсудим возможные причины большой разницы между рейтингами конкретности/абстрактности у некоторых пар переводных эквивалентов на

примере слова ***администрация***. В русском языке ***администрация*** имеет только значения: органы управления и должностные лица, возглавляющие организацию (Кузнецов 2006). В то же время в английском, кроме этого слово ***administration*** имеет еще и значение “the activities that are done in order to plan, organize and run a business, school or other institution”, указанное в Oxford Learner’s Dictionaries (ENA, April 17, 2022)[6].

Например: *the day-to-day administration of a company* (там же). Это значение соответствует русскому ***администрирование***. Таким образом, даже у казалось бы точных переводных эквивалентов вполне возможны различия в значениях, причем значения могут различаться именно в аспекте конкретности/абстрактности. Примеры: *Я пошел в администрацию* и *Администрацию университета пора полностью менять* указывают на вполне конкретные значения людей и места. В тоже время ***администрирование*** относится к весьма абстрактному процессу.

Данное исследование позволило сформулировать два основных вывода. Во-первых, русские и английские рейтинги конкретности/абстрактности рассмотренных слов преимущественно расположены в одном и том же сегменте шкалы и во многих случаях весьма близки. Это указывает на то, что концепция конкретности/абстрактности в значительной степени является языковонезависимой, по меньшей мере в пределах культуры западной цивилизации. Во-вторых, важную роль в этой концепции имеет языково-специфический компонент, определяемый разницей культур.

4.6. Многозначные слова

При составлении словарей, подобных нашему, особой проблемой является многозначность слов (Volskaya et al. 2020). Ясно, что разные значения слов вполне могут иметь разные индексы. Однако ранее эта проблема игнорировалась. Нами впервые (Andreeva et al. 2020) предпринята попытка присвоения индексов отдельным значениям слов. С этой целью был проведен отдельный эксперимент. Для каждого заведомо многозначного слова для простоты выбиралось два его разных значения, одно из которых является конкретным, а другое – абстрактным. Значения брались по словарю (Малый академический словарь 1981). Для обоих значений подбирались контексты, в которых реализуются эти значения. Контекст задавался словосочетанием из двух (редко 3–4) слов. Словосочетания составлялись так, чтобы их частотность (по НКРЯ) была примерно одинаковой. В анкеты для оценки включались именно такие словосочетания, в итоге отобраны 206 слов (из 1000). В анкетах словосочетания были сгруппированы по 30 слов (60 сочетаний). Респондентами явились 280 носителей русского языка в возрасте от 18 до 60 лет. Рейтинги нормированы к диапазону 100–700.

[6] https://www.oxfordlearnersdictionaries.com/definition/english/administration?q=administration

Рейтинги конкретности/абстрактности отдельных значений были сопоставлены дважды: (1) друг с другом и (2) с рейтингами слов, оцененных ранее как единое целое. Например, для слова ***дорога*** мы сопоставили (1) оценки двух значений, реализованных в сочетаниях «*проселочная* ***дорога***» (192) и «*собраться в* ***дорогу***» (475); (2) рейтинги обоих этих значений с общей оценкой слова «***дорога***» (199). Как мы видим, в данном конкретном случае два рейтинга, т.е. «*проселочная* ***дорога***» и «***дорога***», близки (192 против 199), в то время как рейтинг сочетания «*собраться в* ***дорогу***» значительно отличается. Первое может свидетельствовать о том, что при восприятии слова ***дорога*** носители языка прежде всего визуализируют физическую дорогу, вроде *проселочной* ***дороги***, а не более абстрактные значения этого слова, такие как "путешествие". Рис. 7 представляет различия в рейтингах многозначных слов, позволяющие оценить степень их разброса на шкале оценок. Средняя разница двух оценок – 204.

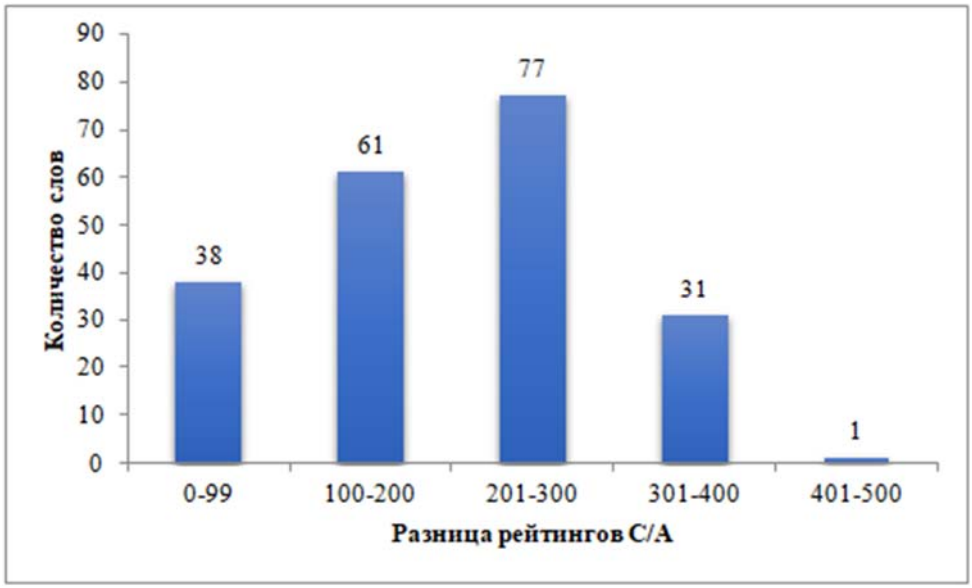

Рис. 7. Разница рейтингов в словосочетаниях / Fig. 7. Ratings difference in word combinations

Максимальная разница в рейтингах (более 400 единиц) обнаруживается, в частности, в слове ***поворот***, определяемом как: 1) «*место, где дорога поворачивает, отклоняется в сторону*»; 2) «*полное изменение в развитии чего-либо*» (Малый академический словарь 1981). Рейтинги сочетаний «***поворот*** *налево от дома*» (129) и «***поворот*** *судьбы*» (540) указывают на различие между оценками респондентов конкретных и абстрактных значений слова. Рейтинг слова ***поворот*** в целом, без разделения на значения, равен 464, т.е. ближе ко второму абстрактному значению. Полученные результаты показывают, что имеет смысл выделять отдельные значения многозначных слов и оценивать степень их конкретности/абстрактности в словосочетаниях, иллюстрирующих только один смысл.

4.7. Сравнение с данными машинного словаря

После создания машинных словарей и до их использования целесообразно проанализировать характер машинных оценок, провести количественное и качественное сравнение их с человеческими. Сопоставим оценки третьей версии машинного словаря с оценками респондентов 1300 слов в нашем словаре. Разница между машинными и человеческими оценками варьируется в диапазоне от 1,28 до 2,1. Отрицательная разница означает, что машинный рейтинг оказался меньше. Большая часть слов (916 из 1300) получила оценки с небольшой разницей в интервале от –0,4 до 0,8 (рис. 8).

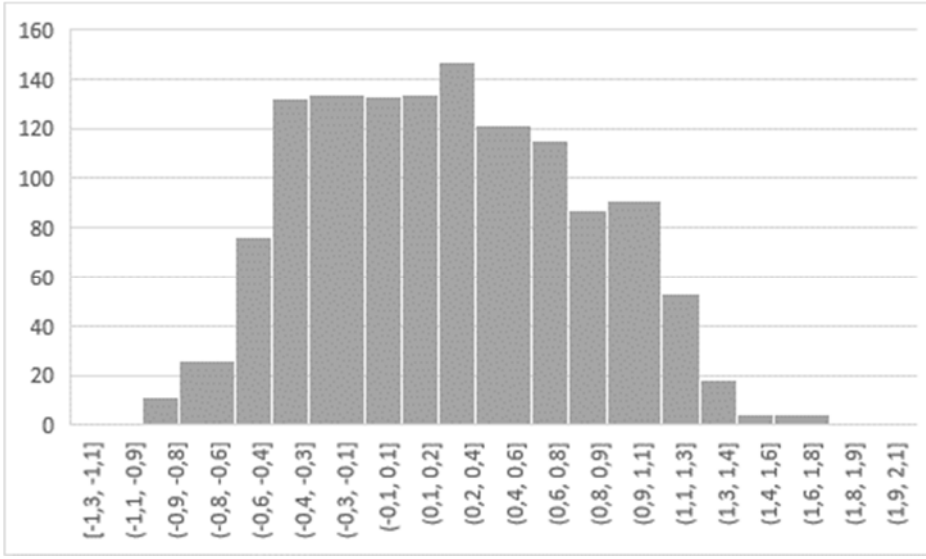

Рис. 8. Разница в оценках / Fig. 8. Ratings' difference

Рассмотрим слова с наибольшей разницей в оценках. Это 114 слов, которые были оценены с разницей от –0,5 до –1,28 (рис. 9), и 155 слов с разницей от 1 до 2,1 (рис. 10). Слова с наибольшей отрицательной разницей – это, как правило, существительные, которые по данным машинного словаря оценивались как более конкретные (с рейтингом меньше 2), а по оценкам респондентов – как менее конкретные. Слова, которые получили наибольшую положительную разницу в оценках, по данным машинного словаря, являются, как правило, более абстрактными (с рейтингом больше 4).

Важно отметить тенденцию, которая обнаруживается при анализе. По оценкам респондентов, большая часть слов получила срединное значение от 2,5 до 3,4 (496 слов), выявлено всего 19 слов со степенью, близкой к 5, и всего 96 слов со степенью от 1 до 1,5, т.е. большая часть опрашиваемых при прохождении опроса не выбирала на шкале крайние значения – 1 или 5. Однако, по данным машинного словаря, единиц со срединными оценками выявлено меньше – 378 существительных, а единиц со степенью, приближенной к крайним значениям, напротив, обнаружено больше, а именно 242 лексемы со степенью от 4 до 5, 211 – со степенью от 1 до 1,5 (рис. 11).

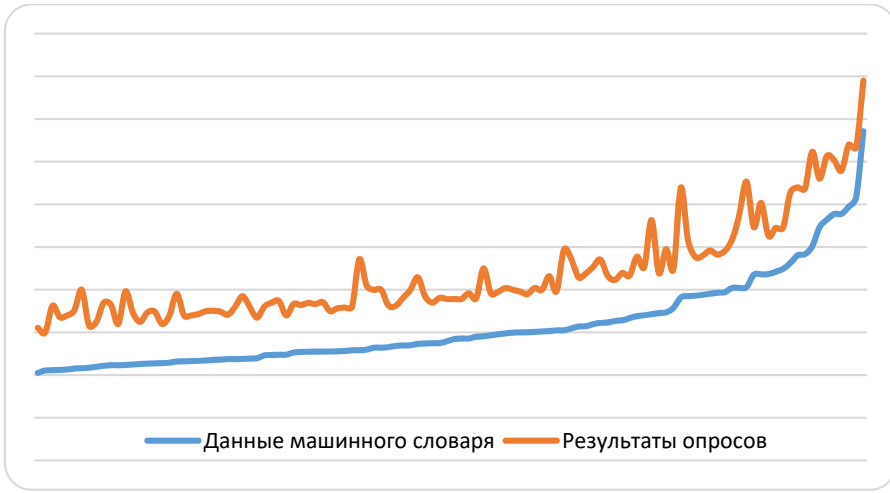

Рис. 9. Отрицательная разница между данными машинного словаря и оценками респондентов /
Fig. 9. Negative difference between machine dictionary and survey results data

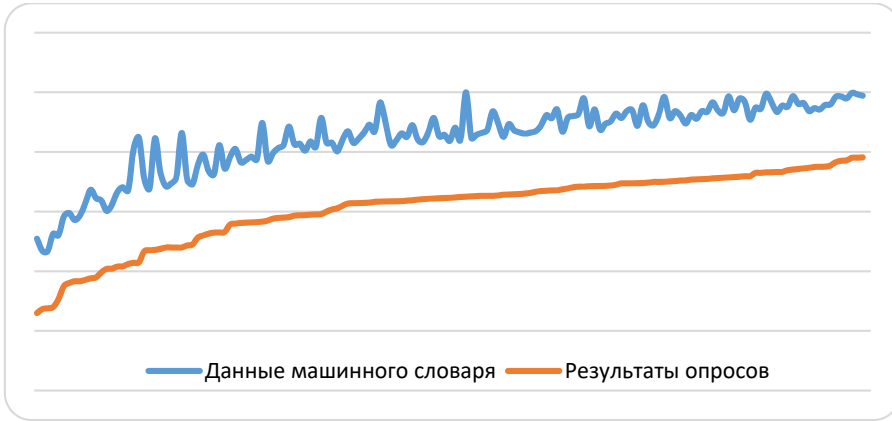

Рис. 10. Положительная разница между данными машинного словаря и оценками респондентов /
Fig. 10. Positive difference between machine dictionary and survey results data

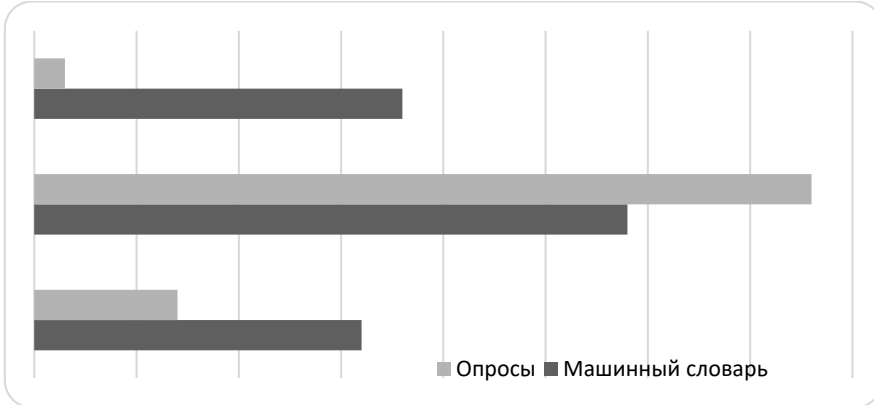

Рис. 11. Количество оценок с крайними и средними значениями /
Fig. 11. Ratings with extreme and mean values

Подобная ситуация означает, что респонденты не склонны к резким оценкам, что является следствием учета ими определенного собственного опыта (Pasquale et al. 2010) либо редких значений многозначных слов, в некоторых из которых слово можно расценивать как абстрактное, а в других – как конкретное. Это приводит к сдвигу оценок к середине шкалы. Такие редкие значения могут быть не представлены в должной мере в корпусах текстов, на которых обучаются нейронные сети.

5. Лингвистические исследования: конкретность vs специфичность

В лингвистике используется семантическая категория, близкая к конкретности, – специфичность. Кажется, что между этими категориями есть корреляция, и поэтому их не всегда различают. Скажем, понятие «*диван*» более специфическое, чем понятие «*мебель*», и одновременно слово *диван* более конкретное, чем *мебель.* Возникает вопрос: в какой мере эти две категории коррелируют? Первое подобное исследование на эмпирическом материале для английского языка было проведено в работе (Bolognesi et al. 2020). В ней показано, что корреляция есть, но умеренная – 0,361 по Спирмену. В нашей работе (Ivanov & Solovyev 2021) ставятся те же цели, что и в указанной публикации, но исследование проводится для русского языка, при этом, естественно, меняются используемые внешние лингвистические ресурсы. Исследование ограничено именами существительными для обеспечения сопоставимости с работой (Bolognesi et al. 2020), а также в связи с тем, что именно для существительных иерархические отношения описаны наиболее подробно.

Категория специфичности/общности интуитивно представляется достаточно понятной. Важный вклад в ее изучение внесли классические работы Рош (Rosch 1975). После создания тезауруса WordNet (Fellbaum 1998) степень специфичности/общности обычно оценивается по положению единицы в иерархии тезауруса в тезаурусе WordNet (Devitt & Vogel 2004). Структура WordNet и ее релевантность лингвистическим фактам представлена в (Miller 1998). Чем понятие, представленное синсетом (синонимическим рядом) WordNet, ближе к нижним уровням тезауруса, тем оно более специфично. Это можно автоматически оценить количественно. Для этого мы используем формулу, предложенную в (Bolognesi 2020): рейтинг специфичности – $(1 + d)/D$, где d – общее число гиперонимов (прямых и непрямых) целевого слова и D – максимальное расстояние от листьев до вершины иерархии. Для WordNet эта величина равна 20. В используемом нами тезаурусе русского языка RuThes (Лукашевич 2011) D = 13. Тезаурус RuThes (ENA, April 17, 2022)[7] содержит более 31,5 тыс. понятий, 111,5 тыс. различных текстовых входов (слов и выражений русского языка). Рейтинг специфичности стандартизирован – приведен к 5-балльной шкале.

[7] http://www.labinform.ru/pub/ruthes/index.htm

Значения конкретности и специфичности всех рассматриваемых нами 14294 слов (общих для RuThes и словаря конкретности) русского языка приведены в файле Concreteness Ratings in RuThes на сайте проекта «Технологии создания семантических электронных словарей» (ENA, April 2017, 2022)[8]. Коэффициент корреляции Спирмена между рейтингами конкретности и специфичности оказался равен 0,264, Пирсона – 0,256 ($p < 0{,}001$). Для английского языка коэффициенты корреляции – 0,361 и 0,354 соответственно (Bolognesi, Burgers & Caselli 2020).

Установленный нами коэффициент корреляции, хотя и является положительным и статистически значимым, классифицируется как слабый (Evans 1996). Таким образом, на материале русского языка подтверждается качественный результат работы (Bolognesi, Burgers & Caselli 2020), что указывает на независимость параметров «конкретность» и «специфичность» и необходимость их самостоятельного изучения. Специфические концепты могут быть как конкретными, так и абстрактными. Рис. 12 визуализирует распределение слов по параметрам конкретность-специфичность. По рисунку видно отсутствие явной корреляции между этими двумя параметрами.

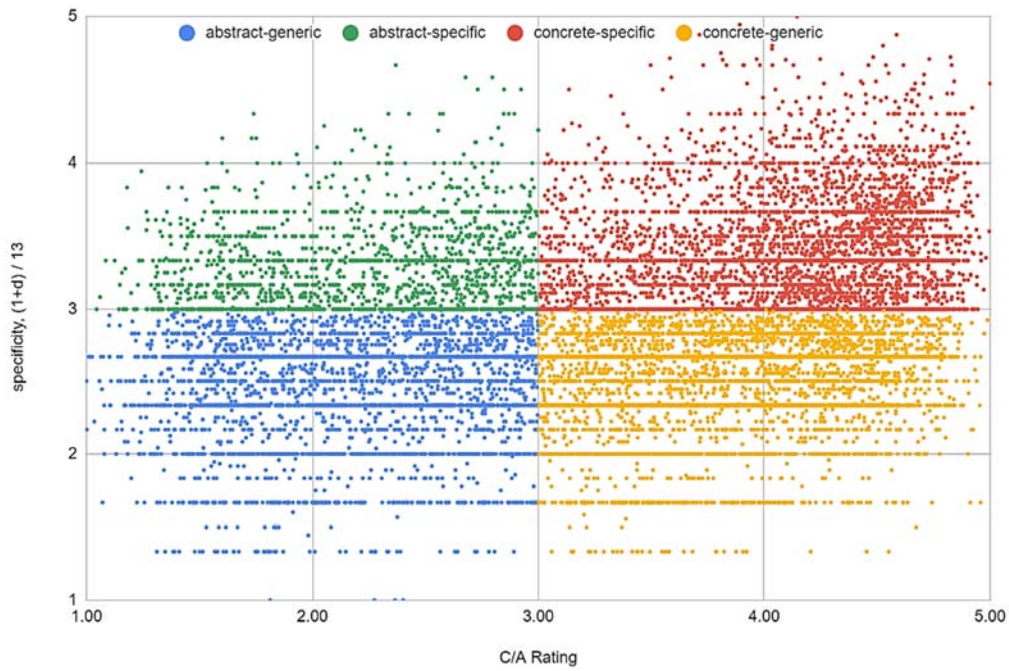

Рис. 12. Распределение слов в двумерном пространстве конкретность-специфичность (Ivanov & Solovyev 2021) /
Fig. 12. Word distribution in two-dimensional space of concreteness-specificity (Ivanov & Solovyev 2021)

[8] https://kpfu.ru/tehnologiya-sozdaniya-semanticheskih-elektronnyh.html

Сопоставление категорий специфичности и конкретности имеет важные импликации для когнитивной науки. В (Bolognesi et al. 2020) выдвинуто предположение, что специфичность отражает характер структурирования мира языком, в то время как конкретность – структурирования мира сознанием в широком смысле, включая перцептивный уровень. Различие этих двух категорий (низкая степень корреляции) является аргументом против «сильной» версии гипотезы лингвистической относительности Сепира–Уорфа, предполагающей, что язык определяет мышление.

В следующем разделе мы перейдем к возможным практическим приложениям словаря и опишем одно уже реализованное его применение – к анализу сложности текстов.

6. Приложения словаря: сложность текстов

Доля абстрактных слов как признак сложности текста рассматривалась рядом исследователей (Taylor & Weir 2012, Fisher et al. 2017). Корреляция между абстрактностью и сложностью текста была также продемонстрирована в работах российских ученых, проводивших исследование на материале русскоязычных текстов (Криони и др. 2008, Solovyev et al. 2019b).

В работе (Solovyev et al. 2020b) изучалась вариативность рейтингов в образовательных текстах. Сопоставлялись учебники: (1) для начальных и старших классов, (2) для гуманитарных и естественнонаучных дисциплин, а также (3) тексты-оригиналы и пересказы. Материал исследования составили Учебный корпус русского языка (УКРЯ) и Корпус пересказов (КП). УКРЯ включает 74 учебника общим объемом 3 млн словоупотреблений (табл. 4). В УКРЯ вошли учебники 2006–2020 гг. выпуска по гуманитарным и естественнонаучным дисциплинам.

Таблица 4. **Учебный корпус русского языка (УКРЯ)**

Класс	Количество словоупотреблений		
	Естественнонаучные дисциплины	Гуманитарные дисциплины	Итого
1	21304	4757	26061
2	29284	28235	57519
3	53565	-	53565
4	51489	24621	76110
5	102467	19527	121994
6	-	159664	159664
7	75205	111788	186993
8	-	273251	273251
9	88335	390821	479156
10	207271	656072	863343
11	-	436322	436322
Итого	628920	2105058	2733978

Table 4. **Russian Academic Corpus**

Grade	Number of words		
	Sciences	Humanities	TOTAL
1	21304	4757	26061
2	29284	28235	57519
3	53565	-	53565
4	51489	24621	76110
5	102467	19527	121994
6	-	159664	159664
7	75205	111788	186993
8	-	273251	273251
9	88335	390821	479156
10	207271	656072	863343
11	-	436322	436322
Итого	628920	2105058	2733978

Корпус пересказов (КП) – это результат исследования, направленного на оценку влияния связности текста на его восприятие читателями (McCarthy et al. 2019). В исследовании участвовали 289 респондентов в возрасте 11–12 лет, ученики 5-го класса. Респондентов индивидуально просили прочитать один из учебных текстов, оригинальный текст (ОТ) и модифицированный текст (МТ), оба состояли примерно из 200 слов. Тексты представляли собой фрагменты главы из учебника Боголюбов Л.Н. Обществознание 5 класс. Учебник для средних школ. 3-е издание. Просвещение, 127 (2013). Пересказы текстов транскрибировались экспертами. Общий размер корпуса составляет 6473 словоупотребления, он доступен на сайте проекта «Технологии создания семантических электронных словарей».

Поскольку в корпусе учебных текстов содержится значительно больше слов, чем в нашем словаре, созданном с помощью опроса респондентов, был использован машинный словарь на 88 тыс. словоформ. По техническим причинам в нем рейтинг принимает значение в диапазоне от –5 (наиболее абстрактные) до +5,5 (наиболее конкретные). Рейтинг текста определяется как среднее арифметическое рейтингов всех входящих в него слов. Если слово из текста отсутствует в словаре, то оно не учитывается. Средние индексы абстрактности текстов учебников, представленные в табл. 5, подсчитаны при помощи онлайн-инструмента RuLingva (Solovyev et al. 2020b), использующего созданный словарь.

Средний индекс указывает на следующее: а) самый высокий индекс конкретности демонстрируют тексты по биологии и учебники для начальной школы, при этом конкретность учебников по биологии для средней школы является самой высокой – +0,49, что даже выше, чем у текстов начальной школы (+0,34); б) учебники по обществознанию имеют самый высокий уровень абстрактности, а учебники по истории расположены в середине шкалы с оценкой «0»; в) индекс абстрактности растет в классах с 1-го по 11-й. Оценка статистической значимости зависимости индекса абстрактности от класса методом линейной регрессии дает значение $p = 1.69e^{-10}$.

Таблица 5. **Индексы текстов учебников и пересказов**

Дисциплина	Класс	Средний индекс
Начальная школа	1-4	+0,34
Биология	5-7	+0,49
Биология	9-10	+0,15
История	10-11	0
Обществознание	5-8	-0,11
Обществознание	9-11	-0,15
Литература	6-8	+0,08
Литература	9-11	-0,14
Тексты МТ	5	0,12
Тексты ОТ	5	0,17
Пересказы	5	0,27

Table 5. **Ratings of recalls and textbooks**

Subject	Grade	Mean rating
Primary school	1-4	+0,34
Biology	5-7	+0,49
Biology	9-10	+0,15
History	10-11	0
Social studies	5-8	-0,11
Social studies	9-11	-0,15
Literature	6-8	+0,08
Literature	9-11	-0,14
MT Texts	5	0,12
OT Texts	5	0,17
Recalls	5	0,27

Тексты ОТ и МТ, предложенные респондентам для пересказа, имеют близкие средние индексы. Как видно из приведенной выше таблицы, средний индекс для пересказов выше, чем у исходных текстов, что подтверждает вывод: респонденты склонны опускать более абстрактные слова и сохранять конкретные при пересказе. Сравнение показателей пересказов и исходных текстов на основе критерия Стьюдента ($p = 0{,}0003$) подтверждает гипотезу о том, что эта разница статистически значима.

Таким образом, созданный нами инструментарий – словари, в том числе машинные, и программа RuLingva – позволяет рассчитывать уровень абстрактности текстов. Полученные на учебных текстах данные подтверждают гипотезы исследования. Аналогичным образом словарь использовался и в других работах по сложности текстов (Солнышкина и др. 2021, Gizatulina et al. 2020).

7. Заключение

Словари являются важным инструментом междисциплинарных исследований концепции конкретности/абстрактности. Словари с рейтингами конкретности/абстрактности созданы для целого ряда языков. В статье описывается первый словарь такого рода для русского языка. Словарь содержит 1500 наиболее частотных существительных. Подробно описана методология создания, которая может быть полезной для создания словарей конкретности/абстрактности для других языков, равно как и для составления других семантических словарей. Методология создания включает многоуровневую систему очистки данных, впервые последовательно примененную в таком масштабе. Выполнено тщательное исследование влияния на результат различных аспектов создания словаря, таких как возраст и уровень образования респондентов. Проведен отдельный эксперимент по оценке многозначных слов. Для других языков подобные исследования не проводились. Нами предложена и опробована методика оценки рейтингов для отдельных значений слов, которая может быть рекомендована к использованию при

построении словарей рейтингов для других языков. Приведено сравнение наших данных с характеристикой конкретности/абстрактности в Русском семантическом словаре Н.Ю. Шведовой, с известными критериями абстрактности.

В дополнение к словарю, созданному путем опроса респондентов, составлен также компьютерный словарь значительно больших размеров, в котором рейтинги конкретности/абстрактности получены путем экстраполяции имеющихся рейтингов респондентов с применением наиболее современных технологий глубокого обучения нейронных сетей. Показано, что качество компьютерных словарей и созданных людьми вполне сопоставимо. Проведенный качественный и количественный анализ данных машинного словаря выявил характер его расхождений с человеческими оценками, что может быть учтено при подборе слов в прикладных исследованиях.

В статье продемонстрированы возможности применения словаря в фундаментальных теоретических исследованиях, направленных на изучение репрезентации в сознании человека таких понятий, как конкретность, абстрактность, специфичность, а также и в прикладных исследований, в частности, для оценки сложности текстов.

В аспекте чисто лингвистических исследований количественно оценена эффективность такого хорошо известного критерия абстрактности, как наличие специфических аффиксов. Проведено сопоставление рейтингов для слов русского языка и их переводных эквивалентов на английском. С одной стороны, имеется высокий уровень корреляции между ними, с другой – выявлены слова с существенным расхождением оценок, что указывает на языковую зависимость категории конкретности/абстрактности. Концепция конкретности близка к концепции специфичности. В связи с этим интересно проанализировать их соотношение. Для русского языка реплицировано исследование, ранее проведенное для английского языка, и показано, что между ними имеется лишь слабая корреляция, поэтому они должны изучаться независимо друг от друга.

Важным приложением словаря конкретности/абстрактности является определение уровня сложности текстов. Чем в тексте больше абстрактных слов, тем он сложнее для восприятия. Поэтому доля абстрактных слов является одним из ключевых параметров, определяющих сложность текстов. С этой точки зрения проанализированы тексты учебников для средней школы.

Подводя итоги серии исследований, можно отметить, что созданные словари имеют высокий уровень качества, достаточный объем и позволяют проводить разнообразные теоретические и прикладные исследования.

Мы планируем проводить дальнейшие исследования в трех направлениях: 1. Повышение качества компьютерных словарей за счет использования более совершенных технологий. 2. Создание словаря с рейтингами позитивности/негативности слов. 3. Лингвистические исследования «эффекта конкретности».

Благодарность

Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета.

REFERENCES / СПИСОК ЛИТЕРАТУРЫ

Andreeva, Mariia, Marina Solnyshkina, Artem Zaikin, Olga Bukach & Radif Zamaletdinov. 2020. Assessment of comparative abstractness: Quantitative approach. *Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020) co-located with 16th International Conference on Computational and Cognitive Linguistics (TEL 2020)*. 132–144.

Black, Paul. 2019. Manhattan distance. *In Dictionary of Algorithms and Data Structures [Online]*. http://www.nist.gov/dads/HTML/manhattanDistance.html. (accessed 19.04.2022)

Bolognesi, Marianna, Burgers Christian & Caselli Tommaso. 2020. On abstraction: Decoupling conceptual concreteness and categorical specificity. *Cognitive Processing* 21 (3). 365–381. DOI: https://doi.org/10.1007/s10339-020-00965-9.

Borghi, Anna M., Ferdinand Binkofski, Cristiano Castelfranchi & Felice Cimatti. 2017. The challenge of abstract concepts. *Psychol. Bull* 143. 263–292.

Brysbaert, Marc, Amy Beth Warriner & Victor Kuperman. 2014a. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46 (3). 904–911.

Brysbaert, Marc, Michaël Stevens, Simon De Deyne, Simon De Deyne & Gert Storms. 2014b. Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica* 150. 80–84. https://doi.org/10.1016/j.actpsy.2014.04.010

Chandola, Varun, Arindam Banerjee & Vipin Kumar. 2009. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)* 41(3). 1–58.

Charbonnier, Jean & Wartena Christian. 2019. Predicting word concreteness and imagery. *In Proceedings of the 13th International Conference on Computational Semantics-Long Papers.* 176–187.

Cristianini, Nello & John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press.

Coltheart, Max. 1981. The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology* 33A. 497–505.

Dallin, J Bailey, Christina Nessler, Kiera N Berggren & Julie L Wambaugh. 2020. An Aphasia treatment for verbs with low concreteness: A pilot study. *American Journal of Speech-Language Pathology* 29 (1). 299–318.

de Groot, Annette M. 1989. Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, and Cognition 15*(5). 824–845. https://doi.org/10.1037/0278-7393.15.5.824

Devitt, Ann & Vogel Carl. 2004. The Topology of WordNet: Some Metrics. *GWC Proceedings*. 106–111.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv*:1810.04805.

Evans, James D. 1996. *Straightforward Statistics for the Behavioral Ssciences*. Brooks/Cole Publishing, Pacific Grove.

Fellbaum, Christiane. 1998. *Wordnet: An Electronic Lexical Database.* MIT Press. Cambridge, Massachusetts.

Fisher, Douglas, Frey Nancy & Lapp Diane. 2016. *Text Complexity: Stretching Readers with Texts and Tasks*. Corwin Press.

Fliessbach, Klaus, Susanne Weis, Peter Klaver, Christian E. Elger & Bernhard Weber. 2006. The effect of word concreteness on recognition memory. *NeuroImage* 32 (3). 1413–1421. https://doi.org/10.1016/j.neuroimage.2006.06.007

Gizatulina, Diana, Farida Ismaeva, Marina Solnyshkina, Ekaterina Martynova & Iskander Yarmakeev. 2020. Fluctuations of text complexity: The case of Basic State Examination in English. *In SHS Web of Conferences* 88. EDP Sciences.

Ivanov, Vladimir & Solovyev Valery. 2021. The Relation of Categories of Concreteness and Specificity: Russian Data. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2021*". URL: http://www.dialog-21.ru/media/5260/ivanovvplussolovyevv049.pdf. (accessed 19.04.2022).

Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou & Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv*:1612.03651.

Kousta, Stavroula-Thaleia, Gabriella Vigliocco, David P Vinson & Mark Andrews. 2011. The representation of abstract words: Why emotion matters. *Exp Psychol Gen. Feb*. 140 (1). 14–34. https://doi.org/10.1037/a0021446.

Krioni, Nikolay K., Alexey D. Nikitin & Anastasiya V. Fillipova. 2008. Avtomatizirovannaya sistema analiza slozhnosti uchebnyh tekstov. *Bulletin of Ufa State Technical University of Aviation* 11. 1 (28). 101–107. (In Russ.)

Kuznecov, Sergey A. 2006. *Bol'shoy Tolkovy Slovar' Russkogo Yazyka.* Norint. (In Russ.)

Laming, Donald. 2004. *Human Judgement: The Eye of the Beholder*. London: Thompson Learning.

Lukashevich, Natilia V. 2011. *Thesauruses in Information Search Tasks.* M.: Izd-vo Moskovskogo universiteta. (In Russ.)

Maximilian, Köper & Sabine Schulte im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 German lemmas. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2595–2598.

McCarthy, Kathryn Soo, Danielle Siobhan Mcnamara, Marina I. Solnyshkina, Fanuza Kh. Tarasova & Roman V. Kupriyanov. 2019. The Russian language test: Towards assessing text comprehension. *Vestnik Volgogradskogo Gosudarstvennogo Universiteta*. Seriiâ˜2, Iazykoznanie; Volgograd 18 (4). 231–247.

McNamara, Danielle, Arthur C. Graesser, Philip M. Mccarthy & Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge, MA: Cambridge University Press.

Mestres-Missé, Anna, Thomas F. Münte & Antoni Rodriguez-Fornells. 2014. Mapping concrete and abstract meanings to new words using verbal contexts. *Second Language Research* 30 (2). 191–223.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546.*

Miller, George A. 1998. Nouns in WordNet. In Christiane Fellbaum (ed.), *Wordnet: An electronic lexical database mit press*. Cambridge, Massachusetts.

Mkrtychian, Nadezhda, Evgeny Blagovechtchenski, Diana Kurmakaeva, Daria Gnedykh, Svetlana Kostromina & Yury Shtyrov. 2019. Concrete vs. Abstract Semantics: From mental representations to functional brain mapping. *Frontiers in Human Neuroscience* 13. 267. https://doi.org/10.3389/fnhum.2019.00267

Naumann, Daniela, Diego Frassinelli & Sabine Schulte im Walde. 2018. Quantitative semantic variation in the contexts of concrete and abstract words. *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, New Orleans, LA*. 76–85.

Paivio, Allan. 1965. Abstractness, imagery, and meaningfulness in paired-associate learning. *Journal of Verbal Learning and Verbal Behaviour* 4. 32–38. https://doi.org/10.1016/s0022-5371(65)80064-0

Paivio, Allan. 1990. *Dual Coding Theory, in Mental Representations: A Dual Coding Approach*. Oxford: Oxford University Press. 53–83. https://doi.org/10.1093/acprof:oso/9780195066661.003.0004

Pasquale, A. Della Rosa, Eleonora Catricalà, Gabriella Vigliocco & Stefano F. Cappa. 2010. Behavior Research Methods Beyond the abstract–concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian. *Behavior Research Methods* 42 (4). 1042–1048. https://doi.org/10.3758/BRM.42.4.1042

Peti-Stantić, Anita, Maja Anđel, Vedrana Gnjidić, Gordana Keresteš, Nikola Ljubešić, Irina Masnikosa, Mirjana Tonković, Jelena Tušek, Jana Willer-Gold & Mateusz-Milan Stanojević. 2021. *The Croatian Psycholinguistic Database: Estimates for 6000 Nouns, Verbs, Adjectives and Adverbs.* 1–18. https://doi.org/10.3758/s13428-020-01533-x

Reilly, Megan, & Rutvik H. Desai. 2017. Effects of semantic neighborhood density in abstract and concrete words*. Cognition* 169. 46–53. https://doi.org/10.1016/j.cognition.2017.08.004

Rosch, Eleanor. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psycholology: General* 104 (3). 192–233.

Sadoski, Mark, Wiliam A. Kealy, E. T. Goetz & Allan Paivio. 1997. Concreteness and imagery effects in the written composition of definitions. *Journal of Educational Psychology* 89(3). 518–526. https://doi.org/10.1037/0022-0663.89.3.518

Sadoski, Mark. 2001. Resolving the effects of concreteness on interest, comprehension, and learning important ideas from text. *Educational Psychology Review* 13(3). 263–281.

Schmid, Hans-J¨org. 2000. *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. *Topics in English Linguistics.* Berlin: Mouton de Gruyter.

Schwanenflugel, Paula J. & Edward J. Shoben. 1983. Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 9 (1). 82–102. https://doi.org/1037/0278-7393.9.1.82

Schwanenflugel, Paula J., Carolyn Akin & Wei-Ming Luh. 1992. Context availability and the recall of abstract and concrete words. *Memory & Cognition* 20 (1). 96–104. https://doi.org/10.3758/bf03208259

Snefjella, Bryor, Michel Généreux & Victor Kuperman. 2019. Historical evolution of concrete and abstract language revisited. *Behavior Research Methods* 51 (4). 1693–1705.

Solnyshkina, Marina I., Radif. R. Zamaletdinov, Ehl'zara Gizzatullina-Gafiyatova, Diana Gizatulina & Maria Begaeva. 2021. Mnogofaktorny analiz slozhnosti teksta. *Inostrannye Yazyki v Shkole*. 28–34. (In Russ.)

Solovyev, Valery D., Vladimir V. Ivanov & Rauf B. Akhtiamov. 2019a. Dictionary of abstract and concrete words of the Russian language: A methodology for creation and application. *Journal of Research in Applied Linguistics* 10. 215–227.

Solovyev, Valery, Mariia Andreeva, Marina Solnyshkina, Radif Zamaletdinov, Andrey Danilov & Dina Gaynutdinova. 2019b. Computing concreteness ratings of Russian and English most frequent words: Contrastive approach. *In the Proceedings of the 12th International Conference on Developments in eSystems Engineering (DeSE)*. 403–408.

Solovyev, Valery D., Vladimir V. Bochkarev & S. V. Khristoforov. 2020a. Generation of a dictionary of abstract/concrete words by a multilayer neural network. *Journal of Physics: Conference Series* 1680 (1). 012046.

Solovyev, Valery, Marina Solnyshkina, Mariia Andreeva, Andrey Danilov & Radif Zamaletdinov. 2020b. Text Complexity and Abstractness: Tools for the Russian

Language. *Proceedings of the International Conference "Internet and Modern Society"*. 75–87.
Solovyev, Valery. 2021. Concreteness/Abstractness Concept: State of the Art. *Advances in Intelligent Systems and Computing* 1358. 275–283.
Spreen, Otfried & Rudolph W. Schulz. 1966. Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. *Journal of Verbal Learning and Verbal Behavior* 5. 459–468.
Taylor, Linda & Weir Cyril J. 2012. *IELTS Collected Papers 2: Research in Reading and Listening Assessment* 2. Cambridge University Press.
Turney, Peter D. & Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37. 141–188.
Vergallito, Alessandra, Marco Alessandro Petilli & Marco Marelli. 2020. Perceptual modality norms for 1,121 Italian words: A comparison with concreteness and imageability scores and an analysis of their impact in word processing tasks. *Behavior Research Methods*. 1–18.
Vinogradov, Victor V. 2001. Russian language (Grammatical studies of a word). *Russian Language*. (In Russ.)
Vol'skaia, Iulia A. 2020. Creating a dictionary of abstract beings in the Russian language: A criterion for selecting vocabulary. *Philology and Culture* 1 (59). 13–17. (In Russ.)
Volskaya, Yulia A., Irina S. Zhuravkina & Alexander P. Lobanov. 2020. Dictionary of abstract the words of the Russian language: Nouns with high numerical measure of abstractness. *International Journal of Criminology and Sociology* 9. 2398–2405.
Wang, X. & Y Bi. 2021. Idiosyncratic tower of Babel: Individual differences in word-meaning representation increase as word abstractness increases. *Psychological Science 32*(10). 1617–1635.
Yao, Zhao, Jia Wu, Yanyan Zhang & Zhenhong Wang. 2017. Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 Chinese words. *Behav Res* 49. 1374–1385. https://doi.org/10.3758/s13428-016-0793-2
Zhuravkina, Irina, Valery Soloviev, Alexander Lobanov & Andrey Danilov. 2020. Comparative analysis of concreteness abstractness of Russian words. *In Conference of Open Innovation Association, FRUCT*. 464–470.

Dictionaies and internet resources / Словари и интернет-ресурсы

Lyashevskay Olga N. & Sharoff S.A. 2009. New Russian frequency dictionary. (In Russ.) http://dict.ruslang.ru/freq.php (accessed 28.12.2021).
Small Academic Dictionary. 1981–1984. (In Russ.) https://gufo.me/dict/mas (accessed 28.05.2021).
Russian National Corpus. (In Russ.) http://www.ruscorpora.ru/ (accessed 28.12.2021).
Russian Semantic Dictionary. 1998. In Shvedova N.Yu. (ed.). 'Azbukovnik' (In Russ.)
RuThes Thesaurus. (In Russ.) http://www.labinform.ru/pub/ruthes/index.htm (accessed 28.12.2021).
Technologies of Compiling Semantic Electronic Dictionaries. (In Russ.) https://kpfu.ru/tehnologiya-sozdaniya-semanticheskih-elektronnyh.html (accessed 28.12.2021).
Cohmetrix. http://cohmetrix.com/ (accessed 28.12.2021).
Corpus of Contemporary American English. https://www.english-corpora.org/coca (accessed 28.05.2021).
Google Books Ngram. https://books.google.com/ngrams (accessed 28.12.2021).
FastText. Library for efficient text classification and representation learning. https://fasttext.cc/ (accessed 28.12.2021).

Приложение / Application

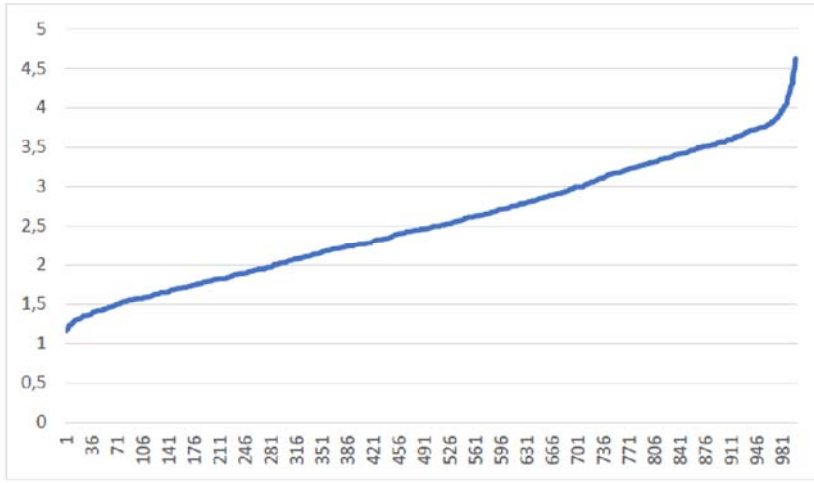

Рис. S1. График распределения оценок, упорядоченных по величине / Fig. S1. Graph of ratings distribution sorted by value

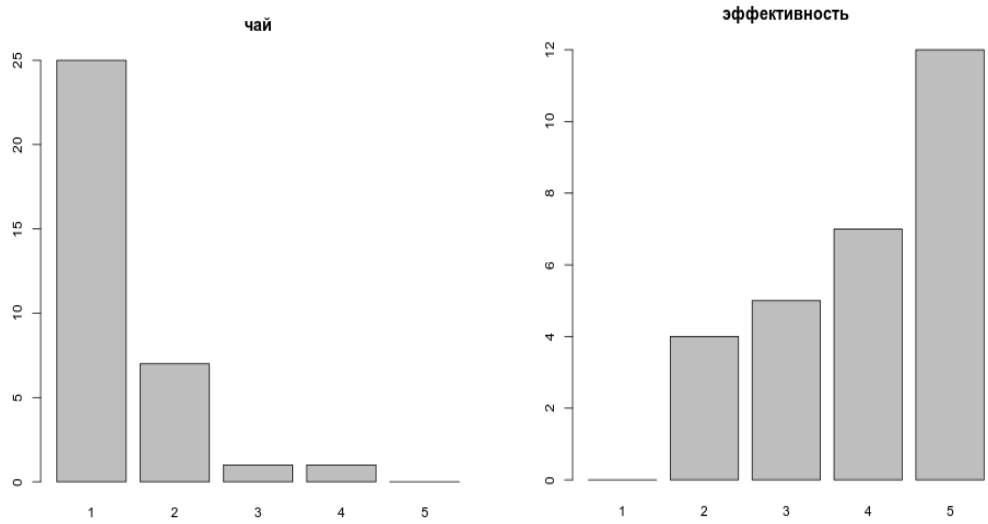

Рис. S2. Характерные распределения оценок для типичных конкретного и абстрактного слова / Fig. S2. Characteristic distributions of typical ratings of concreteness and abstractness of a word

Article history:
Received: 19 November 2021
Accepted: 21 January 2022

Bionotes:
Valery D. SOLOVYEV is Doctor Habil. of Physics and Mathematics, Professor, Chief Researcher of the Text Analytics Research Laboratory at Kazan (Volga Region) Federal University. His research interests embrace cognitive sciences, computer linguistics and text complexity.

Contact information:
Kazan (Volga Region) Federal University
18 Kremlevskaya St., Kazan, 420008, Russia
e-mail: maki.solovyev@mail.ru
ORCID: 0000-0003-4692-2564

Yulia A. VOLSKAYA is Assistant Professor of the Department of Applied and Experimental Linguistics, and Junior Research Fellow of the Neurocognitive Research Laboratory at Kazan (Volga Region) Federal University. Her research interests include applied linguistics and semantics.
Contact information:
Kazan (Volga Region) Federal University
18 Kremlevskaya St., Kazan, 420008, Russia
e-mail: kovaleva95julia@mail.ru
ORCID: 0000-0001-8276-5864

Mariia I. ANDREEVA holds a PhD degree in Philology and is Associate Professor of the Department of Foreign Languages at Kazan State Medical University. She is also Junior Research Fellow of the Text Analytics Research Laboratory at Kazan (Volga Region) Federal University. Her research interests are focused on semantics, sociolinguistics and text analysis.
Contact information:
Kazan State Medical University
49 Butlerov St., Kazan, 420012, Russia
email: lafruta@mail.ru
ORCID: 0000-0002-5760-0934

Artem A. ZAIKIN is Doctor of Physics and Mathematics and Associate Professor of the Department of Mathematical Statistics at Kazan (Volga Region) Federal University. He is also Research Fellow of the Research Laboratory investigating the state and evolution of underground tanks and Junior Research Fellow of TRIZ Modeling Center of Research and Education. His research interests are focused on mathematical statistics.
Contact information:
Kazan (Volga Region) Federal University
18 Kremlevskaya St., Kazan, 420008, Russia
e-mail: kaskrin@gmail.com
ORCID: 0000-0002-5596-3176

Сведения об авторах:
Валерий Дмитриевич СОЛОВЬЕВ – доктор физико-математических наук, профессор, главный научный сотрудник научно-исследовательской лаборатории «Текстовая аналитика» Казанского (Приволжского) федерального университета. Сфера его научных интересов охватывает когнитивную науку, компьютерную лингвистику и сложность текстов.

Контактная информация:
Казанский (Приволжский) федеральный университет
Россия, 420008, г. Казань, ул. Кремлевская, 18
e-mail: maki.solovyev@mail.ru
ORCID: 0000-0003-4692-2564

Юлия Александровна ВОЛЬСКАЯ – ассистент кафедры прикладной и экспериментальной лингвистики, младший научный сотрудник научно-исследовательской лаборатории «Нейрокогнитивные исследования» Казанского (Приволжского) федерального университета. В сферу ее научных интересов входят прикладная лингвистика и семантика.

Контактная информация:
Казанский (Приволжский) федеральный университет
Россия, 420008, г. Казань, ул. Кремлевская, 18
e-mail: kovaleva95julia@mail.ru
ORCID: 0000-0001-8276-5864

Мария Игоревна АНДРЕЕВА – кандидат филологических наук, доцент кафедры иностранных языков Казанского государственного медицинского университета, младший научный сотрудник научно-исследовательской лаборатории «Текстовая аналитика» Казанского (Приволжского) федерального университета. Сфера ее научных интересов включает семантику, социолингвистику и анализ текста.

Контактная информация:
Казанский государственный медицинский университет
Россия, 420012, г. Казань, ул. Бутлерова, 49
e-mail: lafruta@mail.ru
ORCID: 0000-0002-5760-0934

Артем Александрович ЗАИКИН – кандидат физико-математических наук, доцент кафедры математической статистики Казанского (Приволжского) федерального университета, научный сотрудник Научно-исследовательской лаборатории изучения состояния и эволюции подземных резервуаров, младший научный сотрудник научно-образовательного центра «Моделирование ТРИЗ». Основная сфера его научных интересов – математическая статистика.

Контактная информация:
Казанский (Приволжский) федеральный университет
Россия, 420008, г. Казань, ул. Кремлевская, 18
e-mail: kaskrin@gmail.com
ORCID: 0000-0002-5596-3176