
ОБЩАЯ ПСИХОЛОГИЯ GENERAL PSYCHOLOGY

Методы компьютерной лингвистики и обработки естественного языка: возможности и ограничения для задач психологии личности

Кузьмина А.А.

*Национальный исследовательский университет «Высшая школа экономики» (ФГАОУ ВО «НИУ ВШЭ»),
г. Москва, Российская Федерация*

ORCID: <https://orcid.org/0000-0003-0794-9131>, e-mail: kuzmina.alice@gmail.com

Лифшиц М.А.

Независимый исследователь, г. Нью-Йорк, США

ORCID: <https://orcid.org/0000-0003-0079-0244>, e-mail: lifsh22m@ntholyoke.edu

Костенко В.Ю.

*Национальный исследовательский университет «Высшая школа экономики» (ФГАОУ ВО «НИУ ВШЭ»),
г. Москва, Российская Федерация*

ORCID: <https://orcid.org/0000-0002-5612-3857>, e-mail: vasily.kostenko@gmail.com

Использование современных методов компьютерной лингвистики в психологических исследованиях открывает новые возможности, как для изучения вопросов личности и языка, так и для разработки методов психодиагностики. В данной статье рассматриваются основные возможные направления таких исследований, а также неочевидные нюансы, которые важны при их планировании. Максимально задействовать возможности методов компьютерной лингвистики позволит учет особенностей самих методов, системы языка, источников текстов и выборки их авторов, а также уровня теоретической проработки. Каждый из пунктов подробно рассмотрен на примерах уже проведенных исследований. Обзор не является исчерпывающим, однако позволяет создать общую картину методов для дальнейшего поиска решений конкретных исследовательских задач.

Ключевые слова: компьютерная лингвистика, обработка естественного языка, психология личности, анализ текстовых данных.

Финансирование. Исследование выполнено при финансовой поддержке факультета социальных наук Национального исследовательского университета «Высшая школа экономики».

Для цитаты: Кузьмина А.А., Лифшиц М.А., Костенко В.Ю. Методы компьютерной лингвистики и обработки естественного языка: возможности и ограничения для задач психологии личности [Электронный ресурс] // Современная зарубежная психология. 2022. Том 11. № 1. С. 104—115. DOI: <https://doi.org/10.17759/jmfp.2022110110>

Methods of Computational Linguistics and Natural Language Processing: Opportunities and Limitations for Personality Psychology Tasks

Alisa A. Kuzmina

National Research University Higher School of Economics, Moscow, Russia

ORCID: <https://orcid.org/0000-0003-0794-9131>, e-mail: kuzmina.alice@gmail.com

Maria A. Lifshits

Independent researcher, New York, USA

ORCID: <https://orcid.org/0000-0003-0079-0244>, e-mail: lifsh22m@ntholyoke.edu

Vasily Yu. Kostenko

National Research University Higher School of Economics, Moscow, Russia

ORCID: <https://orcid.org/0000-0002-5612-3857>, e-mail: vasily.kostenko@gmail.com

The use of modern methods of computational linguistics in psychological research opens up new possibilities both for the study of personality and language and for the development of psychodiagnostics methods. This article discusses the main possible directions of such research, as well as non-obvious nuances that are important in their planning. Maximum use of the methods of computational linguistics will allow to consider the characteristics of the methods themselves, the language system, sources of texts and a sample of their authors, as well as the level of theoretical development. Each of the points will be considered in detail on the examples of studies already conducted. This review is not exhaustive but allows to create a general picture for the further search for solutions to specific research problems.

Keywords: computational linguistics, natural language processing, personality psychology, textual data analysis.

Funding. This research is supported by the Faculty of Social Sciences, HSE University.

For citation: Kuzmina A.A., Lifshits M.A., Kostenko V.Yu. Methods of Computational Linguistics and Natural Language Processing: Opportunities and Limitations for Personality Psychology Tasks. *Sovremennaya zarubezhnaya psikhologiya = Journal of Modern Foreign Psychology*, 2022. Vol. 11, no. 1, pp. 104—115. DOI: <https://doi.org/10.17759/jmfp.2022110110> (In Russ.).

Введение

Запрос психологии на особые методы обработки текстовых данных исходит из самого факта активного использования этого типа данных в исследованиях. В настоящее время текст является частью исследований, привлекающих методы интервью, дневниковые записи, эссе, вербальные проективные методики, а также использующих публикации как источник данных для исторического и методологического анализа. При «ручной» обработке таких данных поднимаются вопросы ограниченности ресурсов исследователя, субъективности экспертов, размера выборки или обрабатываемого объема данных.

Существует мнение, что исследовательский проект Вальтера Вайнтрауба по изучению естественного языка методом подсчета слов и фраз [52] не был активно поддержан психологическим сообществом именно из-за трудоемкости и ограничений такой обработки данных [30].

Обращение к языку и лингвистическим характеристикам расширяет теоретический потенциал психологии. Так, базой для создания факторных моделей личности стало исследование словарей английского языка [4]. Р. Кеттел выделил из 4500 прилагательных 16 личностных факторов, на основе которых в дальнейшем была разработана пятифакторная модель личности — «Большая пятерка» [26]. Таким образом, анализ лингвистических единиц позволил разработать ряд моделей личности, которые значительно повлияли на развитие исследований устойчивых психологических черт.

Данная работа ставит своей задачей описать: 1) какими возможностями и ограничениями обладает перспектива привлечения методов компьютерной лингвистики в психологии; 2) каковы основные методы компьютерной лингвистики, которые уже были эффективно применены в теоретических, методических и эмпирических исследованиях в психологии личности. Особое внимание мы уделяем возможностям использования методов для русского языка, а также важным аспектам дизайна таких исследований — от формирования выборки до интерпретации результатов.

Компьютерная лингвистика и обработка естественного языка

Для понимания особенностей работы с методами обработки текстов, а также их сильных и слабых сторон важно определить поле их возникновения и применения.

Лингвистика в общем ее понимании — в первую очередь наука о языке, описывающая и изучающая его развитие и состояние, фокусирующаяся на языковой системе и языковой норме. Компьютерная же лингвистика занимается, в первую очередь, применением и изобретением вычислительных методов для решения задач лингвистики. В англоязычной литературе в качестве конкретных примеров основных задач компьютерной лингвистики чаще всего упоминаются машинный перевод, извлечение информации из текста и проблема коммуникации человека и компьютера [27]. Русскоязычная же литература определяет задачи несколько иначе, в первую очередь обозначая вопросы о компьютерном словообразовательном корпусе русского языка, статическом анализе синтаксических форм текста, формализацию делового разговора [2].

В данной статье мы фокусируемся на универсальном определении задач этой науки — в первую очередь на задаче извлечения информации из текста. Именно этот вектор развития методов расширяет возможности психологии, позволяя ученым работать с различным текстовым материалом на новом уровне доступных возможностей.

Задача компьютерной лингвистики по извлечению информации из текста и компьютерная лингвистика в целом напрямую соприкасаются с такой областью, как обработка естественного языка (Natural Language Processing, NLP).

Несмотря на то, что эти области часто считаются взаимозаменяемыми, это не соответствует действительному положению вещей. Граница между ними постепенно стирается, и задачи и цели становятся ближе друг к другу; однако главное различие этих двух областей состоит в задачах, которые перед ними стоят.

Компьютерная лингвистика в основном фокусируется на вопросах лингвистики и ответах на них в понятной человеку форме, тогда как обработка естественного языка до последнего времени была посвящена в первую очередь прикладной стороне вопроса, опираясь в этом на компьютерные науки, и мало внимания уделяла изучению теории языка [42].

Машинное обучение — третья область, часто упоминаемая вместе с компьютерной лингвистикой и обработкой естественного языка — появляется изначально отдельно от них и в первую очередь является набором методов, которые используются для решения определенных задач (рис. 1). Сочетая продвинутые статистические методы с теорией и возрастающей вычислительной мощностью компьютеров, машинное обучение оказывается способно моделировать закономерности, которые человек усмотреть не способен. Поэтому в последние годы, и особенно с распространением нейронных сетей, машинное обучение оказывается важной частью исследований, как в компьютерной лингвистике, так и в обработке естественного языка [13] (рис. 1).



Рис. 1. Взаимоотношения между компьютерной лингвистикой, обработкой естественного языка и машинным обучением

Основные методы анализа текстов и способы их применения

Главным достижением современных методов обработки текстовых данных является их разнообразие, в том числе в степени сложности применения. Для проведения успешного исследования необходимы навыки программирования на языке Python™, применения алгоритмов и структур данных и умение работать с технической документацией. Такие знания позволяют эффективно подбирать и применять методы

исследования (например, понимать, какой тип нейронной сети лучше всего подойдет для решения поставленной задачи), а также осмысленно интерпретировать полученные результаты. Однако применение отдельных методов возможно и при полном отсутствии навыков написания кода с использованием готовых удобных программ.

В данном разделе будут освещены основные методы обработки и анализа текстов по мере их усложнения с уточнением того, какие требования предъявляются к исследователю или его команде при переходе на следующий уровень сложности.

В табл. 1 приведены основные методы и инструменты для анализа и работы с текстовыми данными, которые уже активно использовались в психологических исследованиях или потенциально могли бы быть использованы для определенных исследовательских вопросов. Для каждого из методов и инструментов приводятся требования, необходимые навыки для применения, ссылка на техническую документацию и примеры использования для решения конкретных задач. Ниже мы разберем более подробно те методы, которые могут оказаться полезными для самого широкого круга исследователей.

Самым распространенным методом анализа текстов в психологии является LIWC (англ. Linguistic Inquiry and Word Count [46]). Метод представляет собой компьютерную программу, которая подсчитывает в текстовых файлах количество слов, относящихся к заданным категориям. Всего таких категорий слов в последней версии программы 2015 года более 50 [46]. Они состоят из грамматических категорий (например, части речи) и психологических лингвистических параметров (например, эмоциональные слова и слова, связанные с достижениями), которые выделялись группой экспертов на базе английского языка. LIWC широко используется для изучения различных феноменов на многих языках.

Основная критика метода относится к сложности интерпретации результатов в связи с отсутствием в алгоритме учета значения сочетаний слов. Так, предложения «Я никогда не был менее счастливым» и «Я самый счастливый человек на свете» могут быть закодированы программой как содержащие одинаковую долю слов с позитивными эмоциями [30]. Однако метод направлен именно на выявление тех особенностей использования языка, которые «ускользают» при чтении с пониманием смысла написанного.

Важно отметить особенности использования LIWC для русского языка. Отечественными исследователями отмечается серьезное ограничение русскоязычной версии словаря, которая была создана путем прямого перевода с английского [15]. В настоящее время предпринимаются попытки создания нового русскоязычного словаря [36]. Кроме того, подвергается сомнению эффективность использования для русского языка самого алгоритма программы [15]. Это связано с тем, что LIWC основан на анализе словарных основ слова,

Т а б л и ц а 1

Методы и инструменты компьютерной лингвистики применительно к задачам психологии личности

Метод	Требования	Использование	Тип
LIWC [46]	Программное обеспечение	Подсчет слов определенной категории	Метод
Тематическое моделирование [5]	Библиотеки Python™ и R, готовые модели [53], программное обеспечение	Выделение тем в тексте	Метод
Sentiment Analysis	Библиотека Python™, готовые модели [18], программное обеспечение	Анализ преобладающей эмоциональной составляющей	Метод
Анализ семантической связности текста [14]	Библиотека Python™, готовые модели	Анализ семантической связности	Метод
Оценка семантической сложности текста [8]	Библиотека Python™, готовые модели	Оценка семантической сложности	Метод
Классификация текстов [45]	Библиотека Python™, готовые модели	Разделение текстов по заданным категориям	Метод
PyMorphy [31]	Знание языка Python™	Лематизация, определение части речи и словоформ	Библиотека
Stylo [22]	Знание языка R	Идентификация автора, определение стилистики текста	Библиотека
Omnia Russica [43], Ruscorpora [1]	Умение работать с корпусом	Источник данных	Корпус
Slovet [16]	Знание языка Python™	Морфологический анализ, синтаксис, распознавание понятий	Библиотека
Freeling [24]	Знание языка C++ или умение работать с командной строкой	Снятие омонимии, распознавание понятий, определение части речи	Библиотека
Gensim [41]	Знание языка Python™	Обработка текстов, тематическое моделирование, использование Word2Vec [3]	Библиотека
Google Cloud NLP [10]	Библиотека, готовые модели, программное обеспечение	Синтаксический анализ, определение эмоциональной составляющей, синтаксический анализ, создание моделей	Программное обеспечение, библиотека
Voyant tools [17]	Программное обеспечение	Облака слов, анализ текста, выделение главных тем и слов	Программное обеспечение
NLTK [9]	Библиотека	Методы естественной обработки языка	Библиотека

что полностью соответствует принадлежности английского к аналитической группе языков. Это языки, в которых отношения между словами во многом передаются через служебные слова [28]. Однако русский язык относится к синтетическим языкам, в которых важную роль играют зависимые морфемы (суффиксы, приставки и т. д.), способные полностью изменять словарные значения изначального слова.

Наличие в LIWC жесткого словаря относит его к методам с закрытым словарем (closed vocabulary). Примером методов с открытым словарем (open vocabulary), т. е. словарь которых создается на базе конкретных текстов, является тематическое моделирование.

Тематическое моделирование (англ. topic modeling) — это метод машинного обучения, основанный на кластеризации ключевых слов по близости их употребления. Метод позволяет разделить тексты на группы по темам и характерным для них термам. Примером термов для темы «спорт» могут быть такие слова, как «игрок», «счет» и «пенальти». Тексты, содержащие такие термы, с некоторой вероятностью относятся к одной группе. Так алгоритм анализирует каждый текст

из множества и выдает заданное количество тем, на которые эти тексты могут быть распределены.

Тематическое моделирование может быть использовано для автоматической классификации новостей или в рекомендательных системах. В случае же использования метода в исследовательских целях существенным ограничением является вероятностная основа результатов тематического моделирования, которые зачастую не поддаются интерпретации.

В связи с этим ограничением для реализации тематического моделирования мы рекомендуем использовать метод неотрицательной матричной факторизации (англ. non-negative matrix factorization, NMF; [35]), который превосходит, с точки зрения возможности интерпретации тем, более популярный метод латентного распределения Дирихле (англ. latent Dirichle allocation, LDA [12]).

Реализация методов тематического моделирования может осуществляться как на языках программирования с использованием специальных библиотек, т. е. сборников готовых подпрограмм, созданных для узких задач (например, Gensim (наиболее широко применя-

емая библиотека) или Scikit-learn для Python™ и Topicmodels для языка R), так и при помощи готовых решений с помощью таких программ и платформ, как MonkeyLearn™, Google Cloud NLP™, Aylien™, Meaning Cloud™ и BigML™.

Методы тематического моделирования успешно использовались в психологии в исследованиях Большой пятерки в целом [6] и черты открытости опыту — в частности [19], самоповреждающего поведения [50], публикациях о делирии [33] и многих других.

Как было отмечено выше, LIWC позволяет производить не только лексический, но и морфологический анализ, который также реализуется в программе с основой на заданный словарь. Альтернативой для анализа текстов на русском и украинском языках может служить **морфологический анализатор PyMorphu** [31], который представляет собой библиотеку для языка программирования Python™. Библиотека может быть использована как для лемматизации (то есть для процесса приведения слова к лемме, его словарной форме) для дальнейшего анализа другими методами, так и как самостоятельный метод морфологического анализа. Анализатор позволяет извлекать морфологические характеристики каждого слова: часть речи, падеж, род и т. д.

Примечательно, что анализатор позволяет определить характеристики как для существующих словарных слов, так и, например, для выдуманного слова (например, слово «бутявковедами» библиотека определяет как одушевленное существительное множественного числа, творительного падежа, мужского рода). Становится возможной работа с текстами с необычной лексикой, которая не входит в используемый библиотекой базовый словарь OpenCorpora. В психологических публикациях морфологический анализатор PyMorphu встречается, например, в исследованиях «Темных черт» [15] и субъективного благополучия [48].

Важно отметить, что омонимия в данном методе не снята полностью. К некоторым словам предлагаются несколько вариантов морфологического разбора с указанием вероятности правдоподобия каждого из них на основе частотности. Например, слово «стали» может быть как глаголом, так и существительным. При этом глагол является более общеупотребительным и такой разбор имеет больший вес.

Методы анализа текстов делают возможной автоматическую идентификацию авторства (automatic authorship identification). Из текста может быть извлечена такая информация об авторе, как, например, пол [7], возраст [29], родной язык [7], политическая ориентация [37] и даже результаты методики с сомнительной славой в области психодиагностики — теста Майерс-Брикс [34].

Подобные задачи относятся к **задаче классификации текстов**. Для классификации текстов существует множество подходов и готовых решений, но они, к сожалению, не универсальны, поэтому исследователю, вероятно, нужно будет подбирать и применять класси-

катор, ориентирующийся на особенности данных для каждой конкретной задачи. Существующие модели для более общих задач можно найти в библиотеках Tensorflow и PyTorch, а для имплементации собственной модели имеет смысл также использовать библиотеки Python™ Gensim и SciKit.

Возможные источники данных и требования к исследовательским текстам

Достоверность и надежность получаемых результатов во многом зависит от размера и содержания выборки. Ниже мы рассмотрим основные источники текстовых данных, их возможные альтернативы, а также затронем проблему репрезентативности выборки.

Публикации в социальных сетях являются самым популярным источником данных в психологических исследованиях с задачей классификации текстов. Изучение поведения в социальных сетях в целом и текстовых публикаций в частности, как объектов исследований, относится к области киберпсихологии. Однако отмечается, что данные, например, Facebook и Twitter используются в социальных науках даже в тех случаях, когда исследовательские вопросы напрямую не относятся к использованию социальных сетей [25, с. 11]. В таком случае социальная сеть выступает исключительно как источник данных.

Так, например, активно создаются методы оценки психологических характеристик по текстам в социальных сетях, которые потенциально могли бы конкурировать с методами самоотчета. Примером такой альтернативы является психодиагностическая модель оценки на основе языка (language-based assessment) черт Большой пятерки по данным социальной сети Facebook [6]. Схожим образом стало возможным определение у более 6,5 тысяч русскоговорящих пользователей Facebook черт «Темной триады» [15] и сильных сторон характера у более 4 тысяч англоговорящих пользователей Twitter из разных стран [47].

По мнению ряда исследователей, социальные сети, как источник данных, потенциально способны решить проблему нехватки респондентов в психологических исследованиях [25]. Пользователи социальных сетей ежедневно производят бесчисленное количество данных, которые могут быть проанализированы. Интерпретация и формулирование гипотез для такого рода анализа пересекается с исследовательскими задачами психологии.

Таким образом, междисциплинарное сотрудничество с компьютерной лингвистикой открывает новые возможности как для психологии, так и для других социальных наук, а также для общей науки о данных (data science).

Безусловно, в качестве источников данных для анализа текстов в психологических исследованиях могут использоваться не только публикации и сообщения в социальных сетях. Так, например, к созданию про-

граммы LIWC привел исследовательский проект Джеймса Пеннебейкера по изучению экспрессивного письма: респондентов просили писать в свободной форме о травматических событиях в течение 15—20 минут на протяжении трех—пяти дней [38]. Компьютерный анализ полученных эссе позволил выделить языковые паттерны, которые указывают на различные аспекты физического и психологического здоровья. Меньшее использование когнитивно нагруженных слов, большее количество слов о смерти и местоимений первого лица предсказывали увеличение симптомов посттравматического стрессового расстройства через 6 месяцев [21].

Исследования эссе направлены на изучение различных феноменов. Примером может служить исследование, посвященное сравнению речи преступников с выраженной психопатией и речи преступников без нее. В результате были выделены такие лингвистические маркеры, как высокая отстраненность (например, фраза «вы знаете»), большое количество личных местоимений, редкое упоминание других людей и низкая эмоциональная выразительность, особенно низкое число слов, связанных с тревогой [49]. Также встречаются исследования, направленные на изучение психических расстройств [32], феномена переживания приближения смерти [20], терроризма [51], музыкальных предпочтений [40] и многих других направлений.

В качестве данных могут также использоваться художественные тексты (например, созданные поэтами, совершившими суицид [44]), публикации в газетах [23], бланки проективных методик [39] и многие другие.

Объем текстовых данных может включать не только собственные данные исследователя, но и большие корпуса. Так сервис Google Books на момент 2010 года уже включал в себя 12% всех когда-либо опубликованных книг в оцифрованном виде и доступен для исследователей [30]. Особого внимания отечественных исследователей заслуживает Национальный корпус русского языка (НКРЯ), который включает в себя собрание различных текстов с подробной лингвистической разметкой: «от статьи современного музыкального критика до инструкции по уходу за кактусами, от рассказов Пелевина до справочника по физике» [1].

Множество технических инструментов различной сложности и источников данных разного объема потенциально становится мощным арсеналом психологических исследований и открывает широкие возможности, как для решения уже знакомых вопросов, так и для постановки методологически новых задач.

Однако в эмпирических исследованиях, связанных с анализом текстов, зачастую в недостаточной мере проработан аспект теоретического вклада.

В статье «Использование больших данных для развития теории личности» исследователи отмечают, что обращение к теории важно, как для повышения предсказательной силы методов машинного обучения, так и для интерпретации полученных результатов [11]. При этом нередко акцент ставится именно на предска-

зательной силе [54]. Авторы отмечают, что из-за недостаточной проработки конструктивной валидности методов компьютерной оценки становится затруднительной дальнейшая интерпретация результатов [11, с. 80].

Так, например, непонятно, как объяснить то, что люди с низким уровнем по шкале нейротизма чаще ставят лайки постам про способы заработка денег. Это приводит к тому, что большинство исследований в психологии личности, направленные на анализ текстовых данных, не вносят значительного теоретического вклада.

Таким образом, для того, чтобы исследование с использованием методов компьютерной лингвистики имело теоретический вес, необходимо учесть двойную модель (рис. 2): построить предположения на основе теории, а затем предложить подробную интерпретацию результатов с точки зрения выбранных теоретических позиций.

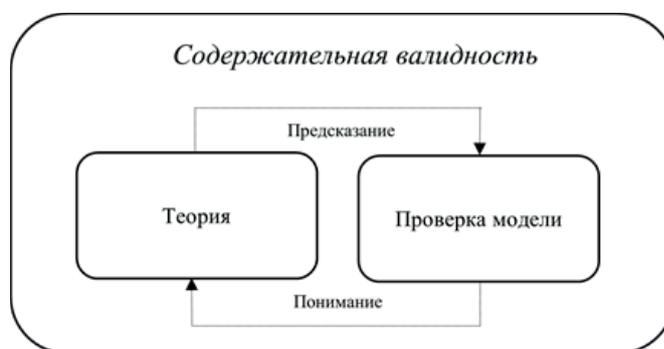


Рис. 2. Содержательная валидность результатов машинного обучения в социальных науках [адаптировано по: 11]

Заключение

Поисковые исследования, эмпирические проверки гипотез, построение теоретических моделей, методологические обзоры — каждое из возможных направлений психологического исследования допускает использование методов компьютерного анализа разнообразных текстовых данных.

Понимание различий и границ компьютерной лингвистики, естественной обработки языка и машинного обучения позволит психологу определить свой междисциплинарный исследовательский интерес относительно каждой из областей.

Проблемы, хорошо знакомые психологам и по другим областям исследований — репрезентативности выборки, адекватности источника данных целям исследования, теоретической проработки и интерпретации результатов — демонстрируют важность глубокого понимания не только технического аспекта, но и изучаемого феномена, а также методологии психологической науки в целом.

Подобные исследования, как правило, проводятся группами исследователей с привлечением специалистов по компьютерным наукам и компьютерной линг-

вистике. Даже базовые знания методов и инструментов в данной области знаний позволят психологу-исследователю строить междисциплинарное взаимодействие и предлагать сложные дизайны исследований.

В представленной статье освещены те инструменты и методы, которые одновременно современны и в достаточной степени проверены в работе с разнообраз-

ными текстовыми данными, что позволяет использовать их как надежную базу.

Безусловно, многие из технологий, которые также могут быть использованы в психологических исследованиях, не вошли в данный обзор. Новые модели и подходы продолжают создаваться и демонстрируют все большую предсказательную эффективность.

Литература

1. Плузьян В.А. Зачем нужен Национальный корпус русского языка? Неформальное введение [Электронный ресурс] // Национальный корпус русского языка: 2003—2005. Результаты и перспективы. М.: Индрик, 2005. С. 6—20. URL: <https://ruscorpora.ru/new/sbornik2005/02plu.pdf> (дата обращения: 03.03.2022).
2. Ясулова Х.С., Шихиев Ш.Б. Прикладные задачи компьютерной лингвистики [Электронный ресурс] // Вестник Социально-педагогического института. 2015. Том 14. № 2. С. 46—49. URL: <https://elibrary.ru/item.asp?id=26629630> (дата обращения: 03.03.2022).
3. Advances in pre-training distributed word representations [Электронный ресурс] / T. Mikolov, E. Grave, P. Bojanowski, C. Puhresch, A. Joulin // arXiv preprint arXiv:1712.09405. 2017. 4 p. URL: <https://arxiv.org/pdf/1712.09405.pdf> (дата обращения: 03.03.2022).
4. Allport G.W., Odbert H.S. Trait-names: A psycho-lexical study // Psychological monographs. 1936. Vol. 47. № 1. i-171. DOI:10.1037/h0093360
5. An analysis of the coherence of descriptors in topic modeling / D. O’Callaghan, D. Greene, J. Carthy, P. Cunningham // Expert Systems with Applications. 2015. Vol. 42. № 13. P. 5645—5657. DOI:10.1016/j.eswa.2015.02.055
6. Automatic personality assessment through social media language / G. Park, H.A. Schwartz, J.C. Eichstaedt, M.L. Kern, M. Kosinski, D.J. Stillwell, L.H. Ungar, M.E.P. Seligman // Journal of Personality and Social Psychology. 2015. Vol. 108. № 6. P. 934—952. DOI:10.1037/pspp0000020
7. Automatically profiling the author of an anonymous text / S. Argamon, M. Koppel, J.W. Pennebaker, J. Schler // Communications of the ACM. 2009. Vol. 52. № 2. P. 119—123. DOI:10.1145/1461928.1461959
8. Besharati M.R., Izadi M. DAST Model: Deciding About Semantic Complexity of Text by DAST Model [Электронный ресурс] // ArXiv. 2019. 40 p. URL: <http://arxiv.org/abs/1908.09080> (дата обращения: 03.03.2022).
9. Bird S., Loper E. NLTK: the natural language toolkit [Электронный ресурс] // COLING ACL 2006 : 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Interactive Presentation Sessions Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions. Stroudsburg, PA: Association for Computational Linguistics (ACL), 2006. P. 69—72. URL: <https://www.aclweb.org/anthology/P04-3031.pdf> (дата обращения: 03.03.2022).
10. Bisong E. Google AutoML: Cloud Natural Language Processing // Building Machine Learning and Deep Learning Models on Google Cloud Platform / E. Bisong. Berkeley, CA: Apress, 2019. P. 599—612. DOI:10.1007/978-1-4842-4470-8_43
11. Bleidorn W., Hopwood C.J., Wright A.G. Using big data to advance personality theory // Current Opinion Behavioral Sciences. 2017. Vol. 18. P. 79—82. DOI:10.1016/j.cobeha.2017.08.004
12. Campbell J.C., Hindle A., Stroulia E. Latent Dirichlet Allocation // The Art and Science of Analyzing Software Data / C. Bird, T. Menzies, T. Zimmermann. Waltham, MA: Elsevier, 2015. P. 139—159. DOI:10.1016/B978-0-12-411519-4.00006-9
13. Clark A., Fox C., Lappin S. The handbook of computational linguistics and natural language processing [Электронный ресурс]. West Sussex, England : Wiley-Blackwell, 2013. 800 p. URL: <https://books.google.ru/books?id=zBmom42eWPcC&lpq=PA3&hl=ru&pg=PA3#v=onepage&q&f=false> (дата обращения: 03.03.2022).
14. Crossley S.A., Kyle K., McNamara D.S. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion // Behavior research methods. 2016. Vol. 48. № 4. P. 1227—1237. DOI:10.3758/s13428-015-0651-7
15. Dark personalities on Facebook: Harmful online behaviors and language / O. Bogolyubova, P. Panicheva, R. Tikhonov, V. Ivanov, Y. Ledovaya // Computers in Human Behavior. 2018. Vol. 78. P. 151—159. DOI:10.1016/j.chb.2017.09.032
16. Deep Learning based NLP modeling for Russian language [Электронный ресурс] / SloVNet // Github. 2020. URL: <https://github.com/natasha/slovnet> (дата обращения: 03.03.2022).
17. Dickerson M. A gentle introduction to text analysis with Voyant tools [Электронный ресурс] // eScholarship. 2018. 22 p. URL: https://escholarship.org/content/qt6jz712sf/supp/Dickerson_TextAnalysisVoyantTools_112018.pdf (дата обращения: 03.03.2022).
18. Dostoevsky: Sentiment analysis library for russian language [Электронный ресурс] / bureaucratic-labs // Github. 2022. URL: <https://github.com/bureaucratic-labs/dostoevsky> (дата обращения: 03.03.2022).

19. Drawing openness to experience from user generated contents: An interpretable data-driven topic modeling approach / Y. Zhang, H. Wei, Y. Ran, Y. Deng, D. Liu // *Expert Systems with Applications*. 2020. Vol. 144. Article ID 113073. 13 p. DOI:10.1016/j.eswa.2019.113073
20. Dying is unexpectedly positive / A. Goranson, R.S. Ritter, A. Waytz, M.I. Norton, K. Gray // *Psychological Science*. 2017. Vol. 28. № 7. P. 988—999. DOI:10.1177/0956797617701186
21. Early linguistic markers of trauma-specific processing predict post-trauma adjustment / B. Kleim, A.B. Horn, R. Kraehenmann, M.R. Mehl, A. Ehlers // *Frontiers in psychiatry*. 2018. Vol. 9. Article ID 645. 7 p. DOI:10.3389/fpsy.2018.00645
22. *Eder M., Rybicki J., Kestemont M.* Stylometry with R: a package for computational text analysis // *The R Journal*. 2016. Vol. 8. № 1. P. 119—121. DOI:10.32614/RJ-2016-007
23. *Ferraro F.R.* Males tend to die, females tend to pass away // *Death studies*. 2019. Vol. 43. № 10. P. 665—667. DOI:10.1080/07481187.2018.1515127
24. FreeLing: An Open-Source Suite of Language Analyzers [Электронный ресурс] / X. Carreras, I. Chao, L. Padro, M. Padro // *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal / Ed. M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silva. Lisbon: European Language Resources Association (ELRA), 2004. P. 239—242. URL: <http://www.lrec-conf.org/proceedings/lrec2004/pdf/271.pdf> (дата обращения: 03.03.2022).
25. Gaining insights from social media language: Methodologies and challenges / M.L. Kern, G. Park, J.C. Eichstaedt, H.A. Schwartz, M. Sap, L.K. Smith, L.H. Ungar // *Psychological methods*. 2016. Vol. 21. № 4. P. 507—525. DOI:10.1037/met0000091
26. *Goldberg L. R.* Language and individual differences: The search for universals in personality lexicons // *Review of personality and social psychology*. 1981. Vol. 2. № 1. P. 141—165.
27. *Grishman R.* Computational linguistics: an introduction [Электронный ресурс]. Cambridge: Cambridge University Press, 1986. 193 p. URL: <https://books.google.ru/books?id=Ar3-TXCXUkC&lpg=PP1&hl=ru&pg=PP1#v=onepage&q&f=false> (дата обращения: 03.03.2022).
28. *Haspelmath M., Michaelis S.M.* Analytic and synthetic: Typological change in varieties of European languages // *Language Variation — European Perspectives VI. Selected papers from the Eighth International Conference on Language Variation in Europe (ICLaVE 8)* / Eds. I. Buchstaller, B. Siebenhaar. Leipzig: John Benjamins Publishing Company, 2017. P. 3—22. DOI:10.1075/silv.19.01has
29. “How Old Do You Think I Am?” A Study of Language and Age in Twitter [Электронный ресурс] / D. Nguyen, R. Gravel, D. Trieschnigg, T. Meder // *Proceedings of the International AAAI Conference on Web and Social Media*. 2013. Vol. 7. № 1. P. 439—448. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/14381> (дата обращения: 03.03.2022).
30. *Ireland M.E., Mehl M.R.* Natural language use as a marker [Электронный ресурс] // *The Oxford handbook of language and social psychology* / Ed. T.M. Holtgraves. New York: Oxford University Press, 2014. P. 201—218. URL: <https://books.google.ru/books?id=I2UJBAAQBAJ&lpg=PP1&hl=ru&pg=PA201#v=onepage&q&f=false> (дата обращения: 03.03.2022).
31. *Korobov M.* Morphological analyzer and generator for Russian and Ukrainian languages // *Analysis of Images, Social Networks and Texts: International Conference on* / Eds. M.Yu. Khachay, N. Konstantinova, A. Panchenko, D. Ignatov, V.G. Labunets. New York: Springer, 2015. P. 320—332. DOI:10.1007/978-3-319-26123-2_31
32. *Lyons M., Aksayli N. D., Brewer G.* Mental distress and language use: Linguistic analysis of discussion forum posts // *Computers in Human Behavior*. 2018. Vol. 87. P. 207—211. DOI:10.1016/j.chb.2018.05.035
33. *McCoy T.H.* Mapping the Delirium Literature Through Probabilistic Topic Modeling and Network Analysis: A Computational Scoping Review // *Psychosomatics*. 2019. Vol. 60. № 2. P. 105—120. DOI:10.1016/j.psym.2018.12.003
34. *Noecker Jr J., Ryan M., Juola P.* Psychological profiling through textual analysis // *Literary and Linguistic Computing*. 2013. Vol. 28. № 3. P. 382—387. DOI:10.1093/lilc/fqs070
35. Novel Algorithm for Non-Negative Matrix Factorization / Tran Dang Hien, Do Van Tuan, Pham Van At, Le Hung Son // *New Mathematics and Natural Computation*. 2015. Vol. 11. № 02. P. 121—133. DOI:10.1142/S1793005715400013
36. *Panicheva P., Litvinova T.* Matching LIWC with Russian Thesauri: An Exploratory Study // *Artificial Intelligence and Natural Language: 9th Conference, AINL 2020: Helsinki, Finland, October 7—9, 2020: Proceedings* / Eds. A. Filchenkov, J. Kauttonen, L. Pivovarova. Cham: Springer, 2020. P. 181—195. DOI:10.1007/978-3-030-59082-6_14
37. *Pennacchiotti M., Popescu A.M.* Democrats, republicans and starbucks aficionados: user classification in twitter // *KDD '11: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York: Association for Computing Machinery, 2011. P. 430—438. DOI:10.1145/2020408.2020477
38. *Pennebaker J.W.* Putting stress into words: Health, linguistic, and therapeutic implications // *Behaviour research and therapy*. 1993. Vol. 31. № 6. P. 539—548. DOI:10.1016/0005-7967(93)90105-4
39. Personality development through natural language / K. Lanning, R.E. Pauletti, L.A. King, D.P. McAdams // *Nature human behaviour*. 2018. Vol. 2. № 5. P. 327—334. DOI:10.1038/s41562-018-0329-0

40. Personality predicts words in favorite songs / L. Qiu, J. Chen, J. Ramsay, J. Lu // *Journal of Research in Personality*. 2019. Vol. 78. P. 25—35. DOI:10.1016/j.jrp.2018.11.004
41. Řehůřek R., Sojka P. Gensim — statistical semantics in Python [Электронный ресурс]. Paris: EuroScipy, 2011. 1 p. URL: <https://www.fi.muni.cz/usr/sojka/posters/rehurek-sojka-scipy2011.pdf> (дата обращения: 03.11.2021).
42. Schubert L. Computational Linguistics [Электронный ресурс] // *The Stanford Encyclopedia of Philosophy Archive*. 2014. URL: <https://plato.stanford.edu/archives/spr2020/entries/computational-linguistics/> (дата обращения: 03.11.2021).
43. Shavrina T.O., Benko V. Omnia russica: even larger russian corpus [Электронный ресурс] // Труды международной конференции «Корпусная лингвистика — 2019 / Под ред. В.П. Захарова. Санкт-Петербург: Издательство Санкт-Петербургского государственного университета, 2019. P. 94—102. URL: https://events.spbu.ru/eventsContent/events/2019/corpora/corp_sborn.pdf (дата обращения: 03.11.2021).
44. Stirman S.W., Pennebaker J.W. Word use in the poetry of suicidal and nonsuicidal poets // *Psychosomatic medicine*. 2001. Vol. 63. № 4. P. 517—522. DOI:10.1097/00006842-200107000-00001
45. Text classification algorithms: A survey / K. Kowsari, K.J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, D. Brown // *Information*. 2019. Vol. 10. № 4. Article ID 150. 68 p. DOI:10.3390/info10040150
46. The development and psychometric properties of LIWC2015 [Электронный ресурс] / J.W. Pennebaker, R.L. Boyd, K. Jordan, K. Blackburn. Austin, TX: University of Texas at Austin, 2015. 26 p. URL: https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf?Sequence=3 (дата обращения: 03.11.2021).
47. The language of character strengths: Predicting morally valued traits on social media / D. Pang, J.C. Eichstaedt, A. Buffone, B. Slaff, W. Ruch, L.H. Ungar // *Journal of personality*. 2020. Vol. 88. № 2. P. 287—306. DOI:10.1111/jopy.12491
48. The Language of Positive Mental Health: Findings From a Sample of Russian Facebook Users / O. Bogolyubova, P. Panicheva, Y. Ledovaya, R. Tikhonov, B. Yaminov // *SAGE Open*. 2020. Vol. 10. № 2. 8 p. DOI:10.1177/2158244020924370
49. The linguistic output of psychopathic offenders during a PCL-R interview / M.T. Le, M. Woodworth, L. Gillman, E. Hutton, R.D. Hare // *Criminal justice and behavior*. 2017. Vol. 44. № 4. P. 551—565. DOI:10.1177/0093854816683423
50. Using Topic Modeling to Detect and Describe Self-Injurious and Related Content on a Large-Scale Digital Platform / P.J. Franz, E.C. Nook, P. Mair, M.K. Nock // *Suicide and Life-Threatening Behavior*. 2020. Vol. 50. № 1. P. 5—18. DOI:10.1111/sltb.12569
51. Vergani M., Bliuc A.M. The language of new terrorism: Differences in psychological dimensions of communication in Dabiq and Inspire // *Journal of Language and Social Psychology*. 2018. Vol. 37. № 5. P. 523—540. DOI:10.1177/0261927X17751011
52. Weintraub W. Verbal behavior: Adaptation and psychopathology. New York: Springer Publishing Company, 1981. 214 p. DOI:10.2307/3790837
53. ‘What is this corpus about?’: using topic modelling to explore a specialised corpus / A. Murakami, P. Thompson, S. Hunston, D. Vajn // *Corpora*. 2017. Vol. 12. № 2. P. 243—277. DOI:10.3366/cor.2017.0118
54. Wright A.G.C. Current directions in personality science and the potential for advances through computing // *IEEE Transactions on Affective Computing*. 2014. Vol. 5. № 3. P. 292—296. DOI:10.1109/TAFFC.2014.2332331

References

1. Plungyan V.A. Zachem nuzhen Natsional’nyi korpus russkogo yazyka? Neformal’noe vvedenie [Why do we need the National Corpus of the Russian Language? informal introduction] [Elektronnyi resurs]. *Natsional’nyi korpus russkogo yazyka: 2003—2005. Rezul’taty i perspektivy [National Corpus of the Russian Language: 2003-2005. Results and prospects]*. Moscow: Indrik, 2005. pp. 6—20. URL: <https://elibrary.ru/item.asp?id=26629630> (дата обращения: 03.03.2022). (In Russ.).
2. Yasulova Kh.S., Shikhiev Sh.B. Prikladnye zadachi komp’yuternoi lingvistiki [Applied Problems of Computational Linguistics] [Elektronnyi resurs]. *Vestnik Sotsial’no-pedagogicheskogo instituta [Bulletin of the Social and Pedagogical Institute]*, 2015. Vol. 14, no. 2, 3 p. URL: <https://cyberleninka.ru/article/n/prikladnye-zadachi-kompyuternoy-lingvistiki> (Accessed 03.03.2022). (In Russ.).
3. Mikolov T., Grave E., Bojanowski P., Puhersch C., Joulin A. Advances in pre-training distributed word representations [Elektronnyi resurs]. *arXiv preprint arXiv:1712.09405*, 2017. 4 p. URL: <https://arxiv.org/pdf/1712.09405.pdf> (Accessed 03.03.2022).
4. Allport G.W., Odbert H.S. Trait-names: A psycho-lexical study. *Psychological monographs*, 1936. Vol. 47, no. 1, i-171. DOI:10.1037/h0093360
5. O’Callaghan D., Greene D., Carthy J., Cunningham P. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 2015. Vol. 42, no. 13, pp. 5645—5657. DOI:10.1016/j.eswa.2015.02.055
6. Park G., Schwartz H.A., Eichstaedt J.C., Kern M.L., Kosinski M., Stillwell D.J., Ungar L.H., Seligman M.E.P. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 2015. Vol. 108, no. 6, pp. 934—952. DOI:10.1037/pspp0000020

7. Argamon S., Koppel M., Pennebaker J.W., Schler J. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 2009. Vol. 52, no. 2, pp. 119—123. DOI:10.1145/1461928.1461959
8. Besharati M.R., Izadi M. DAST Model: Deciding About Semantic Complexity of Text By DAST Model [Elektronnyi resurs]. *ArXiv*, 2019. 40 p. URL: <http://arxiv.org/abs/1908.09080> (Accessed 03.03.2022).
9. Bird S., Loper E. NLTK: the natural language toolkit [Elektronnyi resurs]. *COLING ACL 2006 : 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Interactive Presentation Sessions Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. Stroudsburg, PA: Association for Computational Linguistics (ACL), 2006, pp. 69—72. URL: <https://www.aclweb.org/anthology/P04-3031.pdf> (Accessed 03.03.2022).
10. Bisong E. Google AutoML: Cloud Natural Language Processing. *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Berkeley, CA: Apress, 2019, pp. 599—612. DOI:10.1007/978-1-4842-4470-8_43
11. Bleidorn W., Hopwood C.J., Wright A.G. Using big data to advance personality theory. *Current Opinion Behavioral Sciences*, 2017. Vol. 18, pp. 79—82. DOI:10.1016/j.cobeha.2017.08.004
12. Campbell J.C., Hindle A., Stroulia E. Latent Dirichlet Allocation. In Bird C., Menzies T., Zimmermann T. (eds.), *The Art and Science of Analyzing Software Data*. Waltham, MA: Elsevier, 2015, pp. 139—159. DOI:10.1016/B978-0-12-411519-4.00006-9
13. Clark A., Fox C., Lappin S. The handbook of computational linguistics and natural language processing [Elektronnyi resurs]. West Sussex, England : Wiley-Blackwell, 2013. 800 p. URL: <https://books.google.ru/books?id=zBmom42eWPcC&lpg=PA3&hl=ru&pg=PA3#v=onepage&q&f=false> (Accessed 03.03.2022).
14. Crossley S.A., Kyle K., McNamara D.S. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, 2016. Vol. 48, no. 4, pp. 1227—1237. DOI:10.3758/s13428-015-0651-7
15. Bogolyubova O., Panicheva P., Tikhonov R., Ivanov V., Ledovaya Y. Dark personalities on Facebook: Harmful online behaviors and language. *Computers in Human Behavior*, 2018. Vol. 78, pp. 151—159. DOI:10.1016/j.chb.2017.09.032
16. SloVNet Deep Learning based NLP modeling for Russian language [Elektronnyi resurs]. Github, 2020. URL: <https://github.com/natasha/sloVnet> (Accessed 03.03.2022).
17. Dickerson M. A gentle introduction to text analysis with Voyant tools [Elektronnyi resurs]. *eScholarship*. 2018. 22 p. URL: https://escholarship.org/content/qt6jz712sf/supp/Dickerson_TextAnalysisVoyantTools_112018.pdf (Accessed 03.03.2022).
18. bureaucratic-labs Dostoevsky: Sentiment analysis library for russian language [Elektronnyi resurs]. *Github*, 2022. URL: <https://github.com/bureaucratic-labs/dostoevsky> (Accessed 03.03.2022).
19. Zhang Y., Wei H., Ran Y., Deng Y., Liu D. Drawing openness to experience from user generated contents: An interpretable data-driven topic modeling approach. *Expert Systems with Applications*, 2020. Vol. 144. Article ID 113073, 13 p. DOI:10.1016/j.eswa.2019.113073
20. Goranson A., Ritter R.S., Waytz A., Norton M.I., Gray K. Dying is unexpectedly positive. *Psychological Science*, 2017. Vol. 28, no. 7, pp. 988—999. DOI:10.1177/0956797617701186
21. Kleim B., Horn A.B., Kraehenmann R., Mehl M.R., Ehlers A. Early linguistic markers of trauma-specific processing predict post-trauma adjustment. *Frontiers in psychiatry*, 2018. Vol. 9. Article ID 645, 7 p. DOI:10.3389/fpsy.2018.00645
22. Eder M., Rybicki J., Kestemont M. Stylometry with R: a package for computational text analysis. *The R Journal*, 2016. Vol. 8, no. 1, pp. 119—121. DOI:10.32614/RJ-2016-007
23. Ferraro F.R. Males tend to die, females tend to pass away. *Death studies*, 2019. Vol. 43, no. 10, pp. 665—667. DOI:10.1080/07481187.2018.1515127
24. Carreras X., Chao I., Padro, Padro M. FreeLing: An Open-Source Suite of Language Analyzers [Elektronnyi resurs]. In Lino M.T., Xavier M.F., Ferreira F., Costa R., Silva R. (eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal*. Lisbon: European Language Resources Association (ELRA), 2004, pp. 239—242. URL: <http://www.lrec-conf.org/proceedings/lrec2004/pdf/271.pdf> (Accessed 03.03.2022).
25. Kern M.L., Park G., Eichstaedt J.C., Schwartz H.A., Sap M., Smith L.K., Ungar L.H. Gaining insights from social media language: Methodologies and challenges. *Psychological methods*, 2016. Vol. 21, no. 4, pp. 507—525. DOI:10.1037/met0000091
26. Goldberg L. R. Language and individual differences: The search for universals in personality lexicons. *Review of personality and social psychology*, 1981. Vol. 2, no. 1, pp. 141—165.
27. Grishman R. Computational linguistics: an introduction [Elektronnyi resurs]. Cambridge: Cambridge University Press, 1986. 193 p. URL: <https://books.google.ru/books?id=Ar3-TXCYXUkC&lpg=PP1&hl=ru&pg=PP1#v=onepage&q&f=false> (Accessed 03.03.2022).
28. Haspelmath M., Michaelis S.M. Analytic and synthetic: Typological change in varieties of European languages. In Buchstaller I., Siebenhaar B. (eds.), *Language Variation — European Perspectives VI. Selected papers from the Eighth International Conference on Language Variation in Europe (ICLaVE 8)*. Leipzig: John Benjamins Publishing Company, 2017, pp. 3—22. DOI:10.1075/silv.19.01has

29. Nguyen D., Gravel R., Trieschnigg D., Meder T. “How Old Do You Think I Am?” A Study of Language and Age in Twitter [Elektronnyi resurs]. *Proceedings of the International AAAI Conference on Web and Social Media*, 2013. Vol. 7, no. 1, pp. 439—448. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/14381> (Accessed 03.03.2022).
30. Ireland M.E., Mehl M.R. Natural language use as a marker [Elektronnyi resurs]. In Holtgraves T.M. (eds.), *The Oxford handbook of language and social psychology*. Oxford University Press, 2014, pp. 201—218. URL: <https://books.google.ru/books?id=I2UJBAAAQBAJ&lpg=PP1&hl=ru&pg=PA201#v=onepage&q&f=false> (Accessed 03.03.2022).
31. Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages. In Khachay M.Yu., Konstantinova N., Panchenko A., Ignatov D., Labunets V.G. (eds.), *International Conference on Analysis of Images, Social Networks and Texts*. New York: Springer, 2015, pp. 320—332. DOI:10.1007/978-3-319-26123-2_31
32. Lyons M., Aksayli N. D., Brewer G. Mental distress and language use: Linguistic analysis of discussion forum posts. *Computers in Human Behavior*, 2018. Vol. 87, pp. 207—211. DOI:10.1016/j.chb.2018.05.035
33. McCoy T.H. Mapping the Delirium Literature Through Probabilistic Topic Modeling and Network Analysis: A Computational Scoping Review. *Psychosomatics*, 2019. Vol. 60, no. 2, pp. 105—120. DOI:10.1016/j.psym.2018.12.003
34. Noecker Jr J., Ryan M., Juola P. Psychological profiling through textual analysis. *Literary and Linguistic Computing*, 2013. Vol. 28, no. 3, pp. 382—387. DOI:10.1093/lilc/fqs070
35. Tran Dang Hien, Do Van Tuan, Pham Van At, Le Hung Son Novel Algorithm for Non-Negative Matrix Factorization. *New Mathematics and Natural Computation*, 2015. Vol. 11, no. 02, pp. 121—133. DOI:10.1142/S1793005715400013
36. Panicheva P., Litvinova T. Matching LIWC with Russian Thesauri: An Exploratory Study. In Filchenkov A., Kauttonen J., Pivovarova L. (eds.), *Artificial Intelligence and Natural Language: 9th Conference, AINL 2020: Helsinki, Finland, October 7—9, 2020: Proceedings*. Cham: Springer, 2020, pp. 181—195. DOI:10.1007/978-3-030-59082-6_14
37. Pennacchiotti M., Popescu A.M. Democrats, republicans and starbucks aficionados: user classification in twitter. *KDD '11: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York: Association for Computing Machinery, 2011, pp. 430—438. DOI:10.1145/2020408.2020477
38. Pennebaker J.W. Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour research and therapy*, 1993. Vol. 31, no. 6, pp. 539—548. DOI:10.1016/0005-7967(93)90105-4
39. Lanning K., Pauletti R.E., King L.A., McAdams D.P. Personality development through natural language. *Nature human behavior*, 2018. Vol. 2, no. 5, pp. 327—334. DOI:10.1038/s41562-018-0329-0
40. Qiu L., Chen J., Ramsay J., Lu J. Personality predicts words in favorite songs. *Journal of Research in Personality*, 2019. Vol. 78, pp. 25—35. DOI:10.1016/j.jrp.2018.11.004
41. Řehůřek R., Sojka P. Gensim — statistical semantics in Python [Elektronnyi resurs]. Paris: EuroScipy, 2011. 1 p. URL: <https://www.fi.muni.cz/usr/sojka/posters/rehurek-sojka-scipy2011.pdf> (Accessed 03.11.2021).
42. Schubert L. Computational Linguistics [Elektronnyi resurs]. *The Stanford Encyclopedia of Philosophy Archive*, 2014. URL: <https://plato.stanford.edu/archives/spr2020/entries/computational-linguistics/> (Accessed 03.11.2021).
43. Shavrina T.O., Benko V. Omnia russica: even larger russian corpus [Elektronnyi resurs]. In Zakharova V.P. (eds.), *Trudy mezhdunarodnoi konferentsii «Korpusnaya lingvistika — 2019 [Proceedings of the international conference «Corpus linguistics—2019»]*. Sankt-Peterburg: Izdatel'stvo Sankt-Peterburgskogo gosudarstvennogo universiteta, 2019, pp. 94—102. URL: https://events.spbu.ru/eventsContent/events/2019/corpora/corp_sborn.pdf (Accessed 03.11.2021).
44. Stirman S.W., Pennebaker J.W. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine*, 2001. Vol. 63, no. 4, article ID 150, pp. 517—522. DOI:10.1097/00006842-200107000-00001
45. Kowsari K., Meimandi K.J., Heidarysafa M., Mendu S., Barnes L., Brown D. Text classification algorithms: A survey. *Information*, 2019. Vol. 10, no. 4, 68 p. DOI:10.3390/info10040150
46. Pennebaker J.W., Boyd R.L., Jordan K., Blackburn K. The development and psychometric properties of LIWC2015 [Elektronnyi resurs]. Austin, TX: University of Texas at Austin, 2015. 26 p. URL: https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf?Sequence=3 (Accessed 03.11.2021).
47. Pang D., Eichstaedt J.C., Buffone A., Slaff B., Ruch W., Ungar L.H. The language of character strengths: Predicting morally valued traits on social media. *Journal of personality*, 2020. Vol. 88, no. 2, pp. 287—306. DOI:10.1111/jopy.12491
48. Bogolyubova O., Panicheva P., Ledovaya Y., Tikhonov R., Yaminov B. The Language of Positive Mental Health: Findings From a Sample of Russian Facebook Users. *SAGE Open*, 2020. Vol. 10, no. 2, 8 p. DOI:10.1177/2158244020924370
49. Le M.T., Woodworth M., Gillman L., Hutton E., Hare R.D. The linguistic output of psychopathic offenders during a PCL-R interview. *Criminal justice and behavior*, 2017. Vol. 44, no. 4, pp. 551—565. DOI:10.1177/0093854816683423
50. Franz P.J., Nook E.C., Mair P., Nock M.K. Using Topic Modeling to Detect and Describe Self-Injurious and Related Content on a Large-Scale Digital Platform. *Suicide and Life-Threatening Behavior*, 2020. Vol. 50, no. 1, pp. 5—18. DOI:10.1111/sltb.12569
51. Vergani M., Bliuc A.M. The language of new terrorism: Differences in psychological dimensions of communication in Dabiq and Inspire. *Journal of Language and Social Psychology*, 2018. Vol. 37, no. 5, pp. 523—540. DOI:10.1177/0261927X17751011
52. Weintraub W. Verbal behavior: Adaptation and psychopathology. New York: Springer Publishing Company, 1981. 214 p. DOI:10.2307/3790837

53. Murakami A., Thompson P., Hunston S., Vajn D. ‘What is this corpus about?’: using topic modelling to explore a specialised corpus. *Corpora*, 2017. Vol. 12, no. 2, pp. 243—277. DOI:10.3366/cor.2017.0118

54. Wright A.G.C. Current directions in personality science and the potential for advances through computing. *IEEE Transactions on Affective Computing*, 2014. Vol. 5, no. 3, pp. 292—296. DOI:10.1109/TAFFC.2014.2332331

Информация об авторах

Кузьмина Алиса Андреевна, студентка магистратуры по программе «Позитивная психология», стажер-исследователь Научно-учебной лаборатории лингвистической конфликтологии и современных коммуникативных практик, Национальный исследовательский университет «Высшая школа экономики» (ФГАОУ ВО «НИУ ВШЭ»), г. Москва, Российская Федерация, ORCID: <https://orcid.org/0000-0003-0794-9131>, e-mail: kuzmina.alice@gmail.com

Лифшиц Мария Александровна, независимый исследователь, г. Нью-Йорк, США, ORCID: <https://orcid.org/0000-0003-0079-0244>, e-mail: lifsh22m@mtholyoke.edu

Костенко Василий Юрьевич, кандидат психологических наук, старший научный сотрудник Международной лаборатории позитивной психологии личности и мотивации, Национальный исследовательский университет «Высшая школа экономики» (ФГАОУ ВО «НИУ ВШЭ»), Москва, Российская Федерация, ORCID: 0000-0002-5612-3857, e-mail: vasily.kostenko@gmail.com

Information about the authors

Alisa A. Kuzmina, Student of the Master Program “Positive Psychology”, Research Intern in the Laboratory of Linguistic Conflict Resolution Studies and Contemporary Communicative Practices, National Research University Higher School of Economics, Moscow, Russia, ORCID: <https://orcid.org/0000-0003-0794-9131>, e-mail: kuzmina.alice@gmail.com

Maria A. Lifshits, Independent Researcher, New York, USA, ORCID: <https://orcid.org/0000-0003-0079-0244>, e-mail: lifsh22m@mtholyoke.edu

Vasily Yu. Kostenko, Ph.D (Psychology), Senior research fellow in International Laboratory of Positive Psychology of Personality and Motivation, National Research University Higher School of Economics, Moscow, Russia, ORCID: <https://orcid.org/0000-0002-5612-3857>, e-mail: vasily.kostenko@gmail.com

Получена 10.03.2021

Received 10.03.2021

Принята в печать 13.11.2021

Accepted 13.11.2021