Когнитивные технологии

Особенности текста и психологические особенности: опыт эмпирического компьютерного исследования*

С.Н. Ениколопов[I], Ю.М. Кузнецова[II], А.Н. Минин[III], М.Ю. Пенкина[IV], И.В. Смирнов[II], М.А. Станкевич[II], Н.В. Чудова[II]

[I] Научный центр психического здоровья, г.Москва, Россия
[II] Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия
[III] Курганский государственный университет, г.Курган, Россия,
[IV] Московский городской психолого-педагогический университет, г.Москва, Россия

Аннотация. Пилотажное исследование проведено в русле актуального направления, связанного с развитием средств автоматического анализа текста в целях выявления психологических особенностей его автора. В работе рассматриваются возможности использования для получения психодиагностически ценной информации нового метода компьютерной обработки текста – лингвистического анализатора PLATIn, разработанного в ИСА РАН на основе процессора Exactus Expert, и позволяющего проводить психолингвистический и лексико-частотный анализ текстов. Показатели PLATIn, полученные для текстов 160 испытуемых (студенты и взрослые, г. Москва и г. Курган), сопоставлялись с данными их психодиагностического обследования (10 методик). Проведенный корреляционный анализ выявил наличие связей между некоторыми текстовыми и психодиагностическими показателями.
Ключевые слова: *сетевая психодиагностика, личностные особенности, автоматический анализ текста, психолингвистические показатели, лексико-частотный анализ.*
DOI: 10.14357/20790279190308

Введение

Проблема отражения психологических особенностей человека в его речевой деятельности давно является предметом исследования в различных отраслях психологии, однако с развитием средств автоматического лингвистического анализа появляются новые возможности проверки работоспособности ранее введенных психолингвистических показателей и вновь вводимых текстовых характеристик. Бурно развивающаяся в последнее десятилетие область на стыке компьютерной лингвистики и искусственного интеллекта - автоматический анализ текста - дает мощные средства для поиска связей между данными психодиагностики и лингвистического анализа.

В общем виде логика исследования связи между лингвистическими и психологическими переменными заключается в том, что данные традиционной диагностики личностных особенностей (шкал, опросников и тестов) сопоставляются с различными категориями описания текста, как стилевыми, так и содержательными [14].

Мировая тенденция компьютеризации анализа текста заключается в привлечении в качестве материала контента сетевой коммуникации в каче-

* Работа выполнена при поддержке РФФИ (№ 17-29-02247 «Создание методов диагностики распространения фрустрации в сетевых дискуссиях»).

стве источника информации об индивидуальных, личностных и социальных особенностях коммуникаторов. В качестве показателей анализа сетевого контента выступают: частота использования определенных лексических и текстовых единиц, характеристики сетевой активности (количество постов в час в течение дня, количество ретвитов, хэштеги, смайлы, публикация фото и т.п.) и сетевого группирования (количество френдов, подписчиков, опосредованных читателей). Для анализа данных применяются такие алгоритмы, как линейная регрессия или Support Vector Machines (SVM) [15]. Результатом анализа сетевого контента выступает лингвистический профиль – совокупность лингвистических показателей, отражающих психологические особенности автора, и позволяющий объяснять и предсказывать интересы, мнения, поведение и другие индивидуальные различия [13,14].

Поскольку среди многочисленных работ, посвященных вопросам выявления личностных и индивидуальных особенностей по сетевому контенту, в иноязычных изданиях все чаще появляются публикации обобщающего и методологического характера, можно полагать, что на настоящем этапе речь идет о формировании нового научного направления – сетевой психодиагностики, имеющей собственный предмет и методологию, свои способы получения, обработки, валидизации и интерпретации данных, а также свои возможности и ограничения.

В обзоре [16] описываются этапы новых методов психологического исследования, основанных на анализе сетевого текста:

1. Отбор данных. Получение доступа к данным, хранящимся на платформах социальных сетей, с учетом параметров корпоративной политики конфиденциальности. Практикуется также привлечение добровольцев, открывающих доступ к своим текстам и участвующих в психодиагностических опросах. Минимальным количеством, обеспечивающим статистически достоверные результаты, считается материал объемом в 1000 речевых единиц на одного испытуемого и 50 тыс. речевых единиц на группу.
2. Извлечение данных – техническая задача, связанная со специфической формой хранения информации на платформах социальных медиа, может быть решена при сотрудничестве со специалистами в области компьютерных технологий.
3. Подготовка данных – отдельная задача при работе с большим массивом информации. К методам подготовки данных относятся: токенизация (расщепление текста на минимальные значимые фрагменты, стематизация (идентификация основ речевых единиц), маркирование (приписывание высказыванию определенной пометки, учитываемой при дальнейшем анализе) и др.
4. Группирование речевых единиц. Выделенные фрагменты текста (отдельные слова или словосочетания) должны быть представлены численно, например, в виде частоты встречаемости изучаемых категорий. Психологические исследования чаще всего подразумевают использование закрытых словарей, то есть, заранее созданных списков лексем, релевантных конкретным исследовательским целям. Компьютерные науки предлагают метод открытых словарей, основанный на формировании групп слов или символов или определении темы непосредственно на материале анализируемых текстов. Применяются такие техники, как латентный семантический анализ (Latent semantic analysis, LSA), дифференциальный лингвистический анализ (Differential language analysis, DLA), автоматическое определение темы (Automatic topic creation) и др. Одним из преимуществ метода открытых словарей является возможность учета неконвенциональных текстовых единиц и их форм, весьма характерных для контента социальных медиа.
5. Применение средств лексического анализа. Наиболее используемым является инструмент LIWC, который будет описан ниже. Кроме того, применяются средства оценки эмоционального содержания (сантимент-анализа), такие, как SAS (http://www.sas.com /en_us/software/analytic/sentiment-analysis.html), а также оценки частотности тематической лексики, например, General Inquirer (http://www.wjh.harvard.edu/~inquirer/Home.html) или DICTION (http: //www.dictionsoftware.com). Выбор конкретного инструмента зависит от исследовательской задачи и типа анализируемых данных.

Ограничения нового подхода имеют характер технический (необходимость специального компьютерного оборудования и специальных программных средств для хранения и обработки данных), лингвистический (проблемы, связанные с разными видами неопределенности в языке, речи и тексте), процедурный (ошибки при извлечении и анализе данных). Вопросы этического свойства связаны с соблюдением норм конфиденциальности при использовании информации из социальных сетей [16].

Для выявления психологических характеристик, привлекаемых к анализу сетевого контента, применяются такие методики, как DISC

(Dominance, Influence, Stability and Compatibility), MBTI (Myers-Briggs Type Indicator) или методики с более узкой направленностью, соответствующей конкретной теме исследования. Однако репутация наиболее надежного средства закрепилась за опросником Big Five Personality, с помощью которого проведено на настоящий день подавляющее большинство исследований [13,20].

Базовым лингвистическим инструментом, как уже было отмечено, является программа Linguistic Inquiry and Word Count (LIWC), впервые представленная в 1993 г. и с тех пор непрерывно совершенствуемая авторским коллективом под руководством J.W.Pennebaker. Основой LIWC является набор словарей, с помощью которого проводится частотный анализ лексических единиц, принадлежащих к каждому словарю. Кроме того, инструмент позволяет осуществлять сантимент-анализ, подсчитывать количество слов, принадлежащих к различным частям речи, и ссылок на высказывания других людей. Текущая версия LIWC2015 вычисляет более 80 различных показателей [18,19].

Предметом особого внимания со стороны специалистов служит достоверность психологических описаний и предсказаний, получаемых на основе анализа сетевого контента. Один из приемов ее определения заключается в сопоставлении результатов автоматического анализа и экспертных заключений. Например, в работе [21] показано, что некоторые признаки психологического неблагополучия (наличие депрессии) по тексту лучше определяются экспертами, а некоторые («катастрофизации боли») – программой автоматического анализа.

В отличие от мировой практики, в России исследования, направленные на анализ сетевого контента с целью извлечения психодиагностической информации, как считают авторы работы [8] на основании результатов поиска в научной базе РИНЦ, практически отсутствуют. Интересы в области компьютеризации психолингвистического анализа до последнего времени были сосредоточены на обработке офлайн текстов. За последние двадцать лет были осуществлены отечественные разработки, предназначенные для аналитического аннотирования ценностного и оценочного содержания публикаций СМИ [4], изучения лексико-семантических характеристик речи детей и подростков с невротическими расстройствами [6], выявления в текстах риторических отношений («противопоставление», «последовательность», «консеквенция» и т.п.), а на их основе – специфических эмоциональных и личностных особенностей их авторов [7]. Примером средства сантимент-анализа сетевых текстов на русском языке может служить программа SentiScan (http://semanticanalyzer.info/blog/category/анализ-тональности/) или исследование [11]. В Воронежском региональном центре русского языка при ВГУ под руководством Т.А.Литвиновой проводятся работы, в основном воспроизводящие схему зарубежных исследований: психодиагностические данные получаются с применением русскоязычной адаптации «Big Five», лингвистические показатели – с помощью русскоязычной версии LIWC, и дополнительно некоторые психолингвистические (индекс Флеша, Фог-индекс Ганнинга, индекс лексического разнообразия текста и пр.) – с помощью авторского скрипта [9,10].

1. Цель и задачи исследования

Целью проведения настоящего исследования явилась апробация текстового анализатора PLATIn, созданного специалистами по искусственному интеллекту ФИЦ ИУ РАН, в качестве инструмента для выявления психодиагностически значимых характеристик текста. В соответствии с данной целью исследование подразумевало сопоставление данных текстового анализа с данными психодиагностики. Конкретными задачами при этом выступили:

1) выявление связей между психологическими особенностями автора и психолингвистическими показателями текста;
2) выявление связей между психологическими особенностями автора и данными лексико-частотного анализа текста;
3) создание корпуса текстов испытуемых, прошедших психодиагностическое обследование, для последующего использования его в автоматическом анализе текста с применением методов машинного обучения.

2. Методы

В исследовании приняли участие 160 чел.: студенты гуманитарных и технических вузов г. Москвы и г. Кургана и взрослые жители этих городов в возрасте от 32 до 46 лет. Испытуемым было предложено написать эссе на тему «Я, другие, мир» объемом в одну страницу и заполнить 10 опросников: опросник агрессивности Басса-Перри (BPAQ) в адаптации С.Н. Ениколопова и Н.П.Цибульского; опросник конструктивного мышления С.Эпштейна (ОКМ) в адаптации С.Н. Ениколопова и С.В. Лебедева; опросник нарциссических черт личности (НЧЛ) Н.М. Клепиковой, О.А. Шамшиковой; опросник Способы совладающего поведения (ССП) Р.

Лазаруса в адаптации Т.Л. Крюковой и Е.В. Куфтяк; опросник Стиль саморегуляции поведения (ССПМ) В.И. Моросановой; тест жизнестойкости (ТЖ) С. Мадди в адаптации Д.А. Леонтьева и Е.И. Рассказовой; опросник черт характера (ОЧХ) В.М. Русалова и О.Н. Маноловой; Пятифакторный личностный опросник (5PFQ) Р. МакКрае и П. Коста в адаптации А.Б. Хромова, а также русскоязычные версии методик – Personal Need for Structure Thompson, Naccarato, Parker, & Moskowitz (шкала Потребности в структуре, ШПС), Multidimensional scale of anomie Heydari, Davoudi, & Teymoori (Шкала аномии, ША), New Personal Fable Scale Lapsley, Fitzgerald, Rice, & Jackson (опросник Личный миф, ЛМ) в адаптации Ю.М. Кузнецовой.

В качестве инструмента автоматического анализа текстов в исследовании был использован анализатор PLATIn [3], разработанный на основе процессора Exactus Expert [12,17]. Результатами работы анализатора являются текстовые показатели двух типов: 1) показатели частотности в тексте или коллекции текстов лексических единиц, относящихся к определенным тематическим группам слов; 2) психолингвистические показатели.

В лингвистике под тематическими группами слов (ТГС) понимается объединение слов на основе классификации предметов и явлений в соответствии с экстралингвистическим принципом сопряженности с определенной темой [1,5]. ТГС объединяют слова вне зависимости от их частеречной принадлежности, связанные друг с другом различными парадигматическими и синтагматическими отношениями [5]. Данные о составе разнообразных конкретных ТГС содержатся в многочисленных диссертационных работах, а также в специальных тематических и идеографических словарях.

Сформированные в соответствии с идеографическим принципом группирования лексического материала, ТГС в качестве исследовательского средства позволяет оценивать выраженность соответствующих тем в текстах, что может рассматриваться в качестве показателя значимости данной тематики для авторов этих текстов. Для составления ТГС применяется способ сплошной выборки лексических единиц с соответствующей целям исследования семантикой из наиболее полных по объему лексического материала словарей, и привлечение материала из специальных тематических словарей. Благодаря этому ТГС представляет собой относительно полный список существующих в конкретном языке средств выражения, используемых носителями для речевого общения на определенную тему. Один из вариантов анализа текстов на основе ТГС представляют собой упомянутые выше сантимент-аналитические процедуры, основанные на применении предварительно составленных списков эмотивной лексики, частотность которой характеризует тональность анализируемого контента.

В соответствии с целями нашего исследования, направленного на выявление психодиагностического потенциала компьютерного анализа текстов, в лексике русского языка были выделены группы с тематикой психологического неблагополучия, напряжения, фрустрации. Конкретные темы определялись на основе лингвистических исследований, посвященных проблеме языковых средств выражения данных состояний. ТГС были сформированы методом сплошной выборки из материала Русского орфографического словаря Российской академии наук и тематических словарей: Словарь современного русского города (Гайдамак и др.), Юрислингвистический словарь инвективной лексики русского (Голев, Головачева), Большой словарь молодежного сленга (Левикова), Словарь русской брани (Мокиенко, Никитина), Словарь молодежного сленга (Никитина), Большой словарь мата (Плуцер-Сарно). Всего в используемые нами ТГС вошло более 47 тыс. лексических единиц. ТГС создавались с целью идентификации в ходе автоматического анализа текстов лексических единиц, относящихся к следующим темам:

1. ЛЕКСИКА НЕДИФФЕРЕНЦИРУЕМОЙ ПО СМЫСЛУ ЭКСПРЕССИИ, в том числе: Лексика мотивации, деятельности и напряжения; Жаргонная лексика («Молодежный», «Компьютерный», Криминальный жаргон); Обсценная лексика; Безысключительная и усилительная лексика (ок. 13000 ед.);
2. ЛЕКСИКА ОТРИЦАТЕЛЬНОЙ ЭМОЦИОНАЛЬНОЙ ОЦЕНКИ, в том числе: Инвективы; Лексика разрушения и насилия; Лексика страдания; Лексика стенических негативных эмоций; Лексика социальной разобщенности; Лексика протестного поведения (ок. 30000 ед.);
3. ЛЕКСИКА ОТРИЦАТЕЛЬНОЙ РАЦИОНАЛЬНОЙ ОЦЕНКИ, в том числе: Лексика неэффективных ментальных действий и их результатов; Лексика отсутствия у предметов необходимых качеств; Лексика утраты позитивных характеристик; Лексика возникновения негативных характеристик (ок. 4000 ед.). Кроме того, были составлены ТГС, тематика которых соответствует описанным в социологии (ВЦИОМ, Левада-центр) социально-экономическим ПРИЧИНАМ СОЦИАЛЬНОГО СТРЕССА: Законность и правопорядок; Дети и образование; Эконо-

мика; Демография и экология; Преступность; Власти; Силовые структуры; ЖКХ; Социальное неравенство и несправедливость; Катастрофы; Здравоохранение и бесплатная медицина (ок. 3000 ед.). Анализатор PLATIn осуществляет идентификацию в текстах правильных изменяемых форм внесенных в ТГС лексических единиц, а также вероятностную идентификацию их форм, образованных с нарушениями правил русского языка или содержащих орфографические ошибки. Результаты отражаются в виде числовых показателей представленности (частотности) лексики, относящейся к каждой из используемой ТГС.

К выявляемым анализатором психолингвистическим показателям относятся: коэффициент Трейгера (КТ) – отношение количества глаголов к количеству прилагательных; коэффициент опредмеченности действия (КОД) – соотношение количества глаголов к количеству существительных; количество существительных и глаголов по сравнению с прилагательными и наречиями; количество глаголов в страдательном залоге; количество безличных глаголов; количество глаголов прошедшего времени; количество глаголов будущего времени; количество причастий и деепричастий; длина слов; количество местоимений вообще, а также личных местоимений первого лица множественного числа, первого лица единственного числа, третьего лица множественного числа и др.

Для оценки корреляций между текстовыми показателями (всего 44 показателя) и показателями опросников (всего 81 показатель) применялся критерий Спирмена.

3. Результаты и их анализ

Получена 121 статистически значимая ($p \leq 0{,}05$) корреляция между текстовыми и психодиагностическими показателями. Рассмотрим в общих чертах наиболее интересные связи, чтобы проиллюстрировать возможности предлагаемого инструмента.

Выявлено, что все психолингвистические показатели чувствительны к одному или более психологическому параметру из числа измеренных в данном исследовании. Наибольшее количество корреляций отмечено для шкалы аномии (ША). Отсутствие связей с психолингвистическими показателями выявлено для значений теста жизнестойкости (ТЖ), опросника агрессивности (BPAQ) и опросника конструктивного мышления (ОКМ).

Наиболее информативным оказался психолингвистический показатель «Отношение числа существительных и глаголов к количеству прилагательных и наречий». Для него получено 9 значимых корреляций с показателями четырех опросников, в частности, положительная корреляция с характеристиками саморегуляции «настойчивость» и «самоконтроль» (5PFQ), и отрицательная – со стилями саморегуляции «планирование», «программирование», «самостоятельность» (ССПМ). Кроме того, этот текстовый параметр тем выше, чем выше переживание собственной уникальности (и по опроснику нарциссизма НЧЛ, и по опроснику Личный миф), а также чувствителен к повышению чувства вины (5PFQ) и склонности к зависти (НЧЛ).

Обнаружилось также, что из показателей опросника 5PFQ лишь небольшая часть имеет психолингвистические корреляты: 1.4 поиск впечатлений, 1.5 проявление чувства вины, 3.2 настойчивость, 5.1 любопытство, III самоконтроль. Из них показатель по шкале 5.1 имеет максимальное количество связей, показывающих, что чем менее развито в человеке любопытство, тем реже он употребляет местоимения вообще и «мы» в частности, а также причастия, деепричастия и длинные слова, и тем более выражена в его тексте акциональность по сравнению с признаковостью (коэффициент Трейгера).

Показатель частоты лексики с аффективной семантикой в целом оказался немного более чувствителен к психологическим особенностям испытуемых, чем психолингвистические показатели (56 значимых корреляций для ТГС и 42 для психолингвистических показателей). Выявлено, что употребление усилительной и безысключительной лексики характерно для тех, кто склонен к консерватизму и обособленности (5PFQ), но не к аномии (ША) и не обладает эмотивными чертами характера (ОЧХ). Лексика протестного поведения употребляется тем чаще, чем выше показатели тревожности и депрессивности (5PFQ), отчетливее чувство собственной незначительности (ЛМ) и снижена стратегия самоконтроля (ССП). Мягкие инвективы чаще употребляют люди с низкой активностью и высоким самоконтролем (5PFQ), дистимные (ОЧХ) и не имеющие ярких переживаний нарциссического круга (НЧЛ и ЛМ). Лексика страдания чаще встречается в текстах тех, кто склонен к подозрительности (5PFQ) и чувству зависти (НЧЛ), чей уровень саморегуляции невысок (ССПМ), а вера в собственную уникальность не концептуализирована на уровне личного мифа (ЛМ).

Среди показателей по ТГС факторов социальной напряженности наибольшую чувствительность к психологическим особенностям автора проявила лексика с семантикой «Силовые струк-

туры» и «Власть». Первая тема интересна авторам с высокой экстраверсией (5PFQ), с дистимным складом характера (ОЧХ) и необщительностью (5PFQ). Тему власти склонны обсуждать люди, не доверяющие государству (США), а также получившие высокие оценки по шкале тревожности 5PFQ и низкие оценки по шкале тревожности ОЧХ. Последнее наблюдение важно в контексте проблемы границ адекватности использования текстовых характеристик как источника психодиагностических, а не исследовательских выводов.

Заключение

Проведенное пилотажное исследование показало перспективность разработки средств сетевой психодиагностики на основе созданного в ФИЦ ИУ РАН инструмента – текстового анализатора PLATIn. В основу его работы положен принцип оценки значимости для авторов текстов определенной темы, речевое выражение которой определяет выбор соответствующих лексических единиц. Полученный эмпирический материал демонстрирует чувствительность лексико-частотных показателей, получаемых с помощью PLATIn, к одной или более психодиагностических переменных из числа измеренных в данном исследовании. Диагностические возможности PLATIn увеличивает имеющаяся в нем функция вычисления психолингвистических показателей, для которых получены многочисленные корреляции с психодиагностическими переменными. Перспективным направлением развития анализатора является привлечение методов машинного обучения для перехода от использования заранее сформированных закрытых словарей к открытым тематическим спискам лексических единиц, пополняемых в автоматическом режиме на основе учета совместной встречаемости лексем в анализируемых текстах. Подход, основанный на использовании открытых тематических списков, способен обеспечить высокую адаптивность инструмента по отношению к целям исследований, в которых он может применяться.

Литература

1. *Алефиренко Н.* Теория языка: введение в общее языкознание. Волгоград: Перемена. 1987.
2. *Васильев Л.М.* Современная лингвистическая семантика. М.: Высшая школа. 1990. с. 101-103.
3. *Девяткин Д.А., Кузнецова Ю.М., Чудова Н.В., Швец А.В.* Интеллектуальный анализ проявлений вербальной агрессивности в текстах сетевых сообществ // Искусственный интеллект и принятие решений. 2014. № 2. С. 95-109.
4. *Зевахина Т.С., Олейникова Е.Е.* Аналитическое аннотирование текстов СМИ: Ценности и оценки // http://www.dialog-21.ru/digests/dialog2006/materials/pdf/ZevakhinaT.pdf
5. *Зеленецкий А.Л., Новожилова О.В.* Теория немецкого языкознания. М.: Наука. 1982. с. 277.
6. *Корабельникова Е.А., Вейн А.М., Голубев В.Л., Крейнес М.Г.* Психолингвистическое исследование сновидений детей и подростков с невротическими расстройствами // Журнал неврологии и психиатрии. 1999. № 1. С. 18-21.
7. *Корпусное исследование устного русского дискурса* / Под ред. А.А.Кибрика и В.И.Подлесской. М.: Языки славянских культур. 2009.
8. *Ледовая Я.А., Тихонов Р.В., Боголюбова О.Н.* Социальные сети как новая среда для междисциплинарных исследований поведения человека // Вестник СПб. ун-та. Сер. 16. Психология. Педагогика. 2017. Т. 7. Вып. 3. С. 193-210.
9. *Литвинова Т.А., Литвинова О.А., Рыжкова Е.С., Бирюкова Е.Д., Середин П.В., Загоровская О.В.* Исследование влияния пола и психологических характеристик автора на количественные параметры его текста с использованием программы Linguistic Inquiry and Word Count // Научный диалог. 2015. № 12 (48). С. 101-109.
10. *Литвинова Т.А., Середин П.В.* Исследование динамики идиостиля суицидента // Известия ВГПУ. 2017. № 3 (276). С. 150-154.
11. *Русначенко Н.Л., Лукашевич Н.В.* Методы интеграции лексиконов в машинное обучение для систем анализа тональности // Искусственный интеллект и принятие решений. 2017. № 2. С. 78-89.
12. *Тихомиров И.А., Смирнов И.В., Соченков И.В., Девяткин ДА.., Шелманов А.О., Зубарев Д.В., Швец А.В., Лешкин А.В., Суворов Р.Е.* Exactus Expert: Поисково-аналитическая система поддержки научно-технической деятельности // Труды тринадцатой национальной конференции по искусственному интеллекту с международным участием КИИ-2012. Б.: БГТУ. 2012. Том 4. С. 100-108.
13. *Ahmad, N., & Siddique, J.* Personality Assessment using Twitter Tweets // 21st International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France. Procedia Computer Science. 2017. 112. P. 1964–1973.
14. *Caplan, J., Adams, K., & Boyd, R.* Language and personality. The Wiley-Blackwell Encyclopedia of Personality and Individual Differences.

2017. URL: https://www.researchgate.net/publication/315671233.
15. *Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H., & Eichstaedt J.C.* Detecting depression and mental illness on social media: an integrative review // Current Opinion in Behavioral Sciences 2017, 18. P. 43–49.
16. *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. URL: http://dx.doi.org/10.1037/met0000091.
17. *Osipov G., Smirnov I., Tikhomirov I., Shelmanov A.* Relational-Situational Method for Intelligent Search and Analysis of Scientific Publications // Proceedings of the Integrating IR technologies for Professional Search Workshop. Moscow. 2013. P. 57-64.
18. *Pennebaker J, Boyd R, Jordan K, Blackburn K.* The development and psychometric properties of LIWC2015. 2015. URL: https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf.
19. *Pennebaker, J.W.* Mind mapping: Using everyday language to explore social & psychological processes // Procedia Computer Science, 2017, 118: 100–107. URL: https://utexas.influuent.utsystem.edu/en/publications/mind-mapping-using-everyday-language-to-explore-social-amp-psycho.
20. *Tandera, T., Hendro, Suhartono, D., Wongso, R., & Prasetio, Y.L.* Personality Prediction System from Facebook Users // Procedia Computer Science, 2017. 116. P. 604–611.
21. *Ziemer, K.S., & Korkmaz, G.* Using text to predict psychological and physical health: A comparison of human raters and computerized text analysis // Computers in Human Behavior, 2017. 76. P. 122-127.

Ениколопов Сергей Николаевич. ФГБНУ «Научный центр психического здоровья», г. Москва. Заведующий отделом медицинской психологии, кандидат психологических наук, доцент. Количество печатных работ: более 150, в т. ч. 3 монографии. Область научных интересов: психология агрессии, психология юмора, психология девиантного и деликвентного поведения. E-mail: enikolopov@mail.ru

Чудова Наталья Владимировна. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН), г. Москва. Старший научный сотрудник, кандидат психологических наук. Количество печатных работ: более 100 (в т.ч. 3 монографии). Область научных интересов: психология агрессии, психология интернета, картина мира. E-mail: nchudova@gmail.com

Кузнецова Юлия Михайловна. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН), г. Москва. Старший научный сотрудник. кандидат психологических наук. Количество печатных работ: 103 (в т. ч. 4 монографии в соавторстве). Область научных интересов: психосемантика, психодиагностика, психолингвистика, психология агрессивности. E-mail: kuzjum@yandex.ru

Пенкина Марина Юрьевна. Институт экспериментальной психологии Федерального государственного образовательного учреждения высшего образования «Московский государственный психолого-педагогического университет» (ФГБОУ ВО МГППУ), г. Москва. Старший преподаватель кафедры общей психологии. Количество печатных работ: 6. Область научных интересов: психология личности, психология агрессивности, детская психология, когнитивные исследования. E-mail: mpenkina@mail.ru

Станкевич Максим Алексеевич. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН), г. Москва. Инженер. Количество печатных работ: 5. Область научных интересов: обработка естественного языка, машинное обучение, анализ социальных сетей. E-mail: stankevich@isa.ru.

Смирнов Иван Валентинович. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН), г. Москва. Доцент, заведующий отделом, кандидат физико-метематических наук. Количество печатных работ: 75. Область научных интересов: обработка естественного языка, интеллектуальный анализ информации. E-mail: ivs@isa.ru

Минин Алексей Николаевич. Федеральное государственное бюджетное образовательное учреждения высшего образования «Курганский государственный университет» (ФГБОУ ВО «КГУ»),г.Курган. Старший преподаватель кафедры социологии и социальной работы. Количество печатных работ: 8. Область научных интересов : религиоведение, социология религии. E-mail : Aminlex@yandex.ru

Text features and psychological characteristics: an empirical study of computer

S.N. Enikolopov[I], Y.M. Kuznetsova[II], A.N. Minin[III], M.Y. Penkina[IV], I.V. Smirnov[II], M.A. Stankevich[II], N.V. Chudova[II]

[I] Mental Health Research Center, Moscow, Russia

[II] Federal Research Center "Computer Science and Control" of RAS, Moscow, Russia

[III] Kurgan State University, Kurgan, Russia

[IV] Moscow State University of Psychology and Education, Moscow, Russia

Abstract. The work is devoted to the identification of links between the automatically distinguished features of the text and the autor's psychological characteristics. On the basis of the analysis of works carried out in Russia and in the world, and on the results of our study, it is proposed to consider the system of automatic text analysis as research tools for a work of a psychologist with large text corpora. The results of a study conducted using the linguistic analyzer PLATIn, developed on the basis of the processor Exactus Expert in ISA RAS, are presented. The study included psycholinguistic and lexical-frequency analysis of texts created by the subjects (142 people, students and adults, Moscow and Kurgan), who underwent psychodiagnostic examination. The lexical analysis was based on the specially designed dictionaries of emotional topics (about 53 thousand lexical units). The analysis of correlations between the text parameters (12 psycholinguistic and 30 lexical) and the psychodiagnostic parameters (81 scales of 10 questionnaires) was carried out. The study showed the sensitivity of the lexical and psycholinguistic parameters to some psychological characteristics. The conclusion about the expediency of the automatic text analysis in psychological population studies is made.

Keywords: *network psychodiagnostics, automatic text analysis, psycholinguistic indicators, emotive vocabulary, psychological features of author*.

DOI: 10.14357/20790279190308

References

1. *Alefirenko N.* Teoriya yazyka: vvedenie v obshchee yazykoznanie. Volgograd: Peremena. 1987.
2. *Vasil'ev L.M.* Sovremennaya lingvisticheskaya semantika. M.: Vysshaya shkola, 1990. P. 101-103
3. *Devyatkin D.A., Kuznecova YU.M., CHudova N.V., SHvec A.V.* Intellektual'nyj analiz proyavlenij verbal'noj agressivnosti v tekstah setevyh soobshchestv // Iskusstvennyj intellekt i prinyatie reshenij. 2014. № 2. P. 95-109.
4. *Zevahina T.S., Olejnikova E.E.* Analiticheskoe annotirovanie tekstov SMI: Cennosti i ocenki // http://www.dialog-21.ru/digests/dialog2006/materials/pdf/ZevakhinaT.pdf.
5. *Zeleneckij A.L., Novozhilova O.V.* Teoriya nemeckogo yazykoznaniya. M.: Nauka. 1982. P. 277.
6. *Korabel'nikova E.A., Vejn A.M., Golubev V.L., Krejnes M.G.* Psiholingvisticheskoe issledovanie snovidenij detej i podrostkov s nevroticheskimi rasstrojstvami // ZHurnal nevrologii i psihiatrii. 1999. № 1. P. 18-21.
7. *Korpusnoe issledovanie ustnogo russkogo diskursa* / Pod red. A.A.Kibrika i V.I.Podlesskoj. M.: YAzyki slavyanskih kul'tur, 2009.
8. *Ledovaya YA.A., Tihonov R.V., Bogolyubova O.N.* Social'nye seti kak novaya sreda dlya mezhdisciplinarnyh issledovanij povedeniya cheloveka // Vestnik SPb. un-ta. Ser. 16. Psihologiya. Pedagogika. 2017. T. 7. Vyp. 3. P. 193-210
9. *Litvinova T.A., Litvinova O.A., Ryzhkova E.S., Biryukova E.D., Seredin P.V., Zagorovskaya O.V.* Issledovanie vliyaniya pola i psihologicheskih harakteristik avtora na kolichestvennye parametry ego teksta s ispol'zovaniem programmy Linguistic Inquiry and Word Count // Nauchnyj dialog. 2015. № 12 (48). P. 101-109.
10. *Litvinova T.A., Seredin P.V.* Issledovanie dinamiki idiostilya suicidenta // Izvestiya VGPU. 2017. № 3 (276). P. 150-154.
11. *Rusnachenko N.L., Lukashevich N.V.* Metody integracii leksikonov v mashinnoe obuchenie dlya sistem analiza tonal'nosti // Iskusstvennyj intellekt i prinyatie reshenij. 2017. № 2. P. 78-89
12. *Tihomirov I.A., Smirnov I.V., Sochenkov I.V., Devyatkin DA.., SHelmanov A.O., Zubarev D.V., SHvec A.V., Leshkin A.V., Suvorov R.E.* Exactus Expert: Poiskovo-analiticheskaya sistema podderzhki nauchno-tekhnicheskoj deyatel'nosti // Trudy trinadcatoj nacional'noj konferencii po iskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2012. B.: BGTU. 2012. Tom 4. P. 100-108.
13. *Ahmad, N., & Siddique, J.* Personality Assessment using Twitter Tweets // 21st International Conference on Knowledge Based and Intelligent

Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France. Procedia Computer Science. 2017. 112. P. 1964–1973.
14. *Caplan, J., Adams, K., & Boyd, R.* Language and personality. The Wiley-Blackwell Encyclopedia of Personality and Individual Differences. 2017. URL: https://www.researchgate.net/publication/315671233.
15. *Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H., & Eichstaedt J.C.* Detecting depression and mental illness on social media: an integrative review // Current Opinion in Behavioral Sciences 2017. 18. P. 43–49.
16. *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. URL: http://dx.doi.org/10.1037/met0000091.
17. *Osipov G., Smirnov I., Tikhomirov I., Shelmanov A.* Relational-Situational Method for Intelligent Search and Analysis of Scientific Publications // Proceedings of the Integrating IR technologies for Professional Search Workshop. Moscow, 2013. P. 57-64
18. *Pennebaker J, Boyd R, Jordan K, Blackburn K.* The development and psychometric properties of LIWC2015. 2015. URL: https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf.
19. *Pennebaker, J.W.* Mind mapping: Using everyday language to explore social & psychological processes // Procedia Computer Science, 2017, 118: 100–107. URL: https://utexas.influuent.utsystem.edu/en/publications/mind-mapping-using-everyday-language-to-explore-social-amp-psycho.
20. *Tandera, T., Hendro, Suhartono, D., Wongso, R., & Prasetio, Y.L.* Personality Prediction System from Facebook Users // Procedia Computer Science. 2017. 116. P. 604–611.
21. *Ziemer, K.S., & Korkmaz, G.* Using text to predict psychological and physical health: A comparison of human raters and computerized text analysis // Computers in Human Behavior. 2017. 76. P. 122-127.

Enikolopov Sergey N., PhD in psychology, head of department, Mental Health Research Center (MHRC), Moscow, Russian Federation, enikolopov@mail.ru

Chudova Natalia V., PhD in psychology, Senior Researcher, Institute for Systems Analysis, Federal Research Center "Computer Science and Control" of RAS, Moscow, Russian Federation, nchudova@gmail.com

Kuznetsova Yulia M., PhD in psychology, Senior Researcher, Institute for Systems Analysis, Federal Research Center "Computer Science and Control" of RAS, Moscow, Russian Federation, kuzjum@yandex.ru

Minin Alexej N., Senior Lecturer, Kurgan State University, Kurgan, Russian Federation, aminlex@yandex.ru

Penkina Marina Y., Senior lecturer, Moscow State University of Psychology and Education (MSUPE), Moscow, Russian Federation, mpenkina@mail.ru

Smirnov Ivan V., PhD, head of department, Institute for Systems Analysis, Federal Research Center "Computer Science and Control" of RAS, ivs@isa.ru

Stankevich Maxim A., Institute for Systems Analysis, Federal Research Center "Computer Science and Control" of RAS, Moscow, Russian Federation, stankevich@isa.ru