Психология. Журнал Высшей школы экономики. 2021. Т. 18. № 4. С. 748–769.
Psychology. Journal of the Higher School of Economics. 2021. Vol. 18. N 4. P. 748–769.
DOI: 10.17323/1813-8918-2021-4-748-769

МЕТОД РЕЛЯЦИОННО-СИТУАЦИОННОГО АНАЛИЗА ТЕКСТА В ПСИХОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

С.Н. ЕНИКОЛОПОВ[a], Ю.М. КУЗНЕЦОВА[b], Г.С. ОСИПОВ[b], И.В. СМИРНОВ[b], Н.В. ЧУДОВА[b]

[a] *Федеральное государственное бюджетное научное учреждение «Научный центр психического здоровья», 115522, Россия, Москва, Каширское шоссе, д. 34*
[b] *Федеральное государственное бюджетное научное учреждение ФИЦ «Информатика и управление» РАН, 117312, Россия, Москва, пр-т 60-летия Октября, д. 9*

The Method of Relational-Situational Analysis of Text in Psychological Research

S.N. Enikolopov[a], Y.M. Kuznetsova[b], G. S. Osipov[b], I.V. Smirnov[b], N.V. Chudova[b]

[a] *Mental Health Research Center, 34 Kashirskoe shosse, Moscow, 115522, Russian Federation*
[b] *Federal Research Center "Computer Science and Control" Russian Academy of Sciences, 9 60-letiya Oktyabrya Ave, Moscow, 117312, Russian Federation*

Резюме

Разработка методов искусственного интеллекта, позволяющих получать достоверную информацию о психологических особенностях человека по его речи, является одним из направлений развития психодиагностического инструментария современного уровня. Создание подобных средств подразумевает взаимодействие психологов, лингвистов и представителей компьютерной науки, поэтому требует определенной методологической базы и организационных условий. В настоящей работе описан опыт разработки отечественного диагностического инструмента анализа письменных текстов, в том числе сетевых интеракций, осуществленной под руководством

Abstract

Some artificial intelligence methods for reliable identification of personality traits in speech can be useful for upgrade of psycho-diagnostics. Such tools should be developed by psychologists, linguists and computer scientists jointly, and therefore a preliminary special methodology and organization are required. The described in our paper original tool for analyzing written text was created at the Artificial Intelligence Research Institute of National Institute for Research in Computer Science and Control of RAS under the guidance of Dr. of Physical and Mathematical Sciences G.S. Osipov in collaboration with the members of Mental

Работа поддержана Министерством науки и высшего образования Российской Федерации, проект № 075-15-2020-799.

The study was funded be the Ministry of Science and Higher Education of the Russian Federation, grant No. 075-15-2020-799.

Г.С. Осипова в возглавляемом им Институте проблем искусственного интеллекта ФИЦ «Информатика и управление» РАН и при сотрудничестве с Научным центром психического здоровья, Институтом русского языка РАН и Пермским государственным университетом. В основе инструмента лежит предложенный Г.С. Осиповым метод реляционно-ситуационного анализа текста (РСА), учитывающий особенности коммуникативной грамматики русского языка Г.А. Золотовой. Извлекаемые с помощью инструмента «Машина РСА» данные позволяют представлять текст в виде совокупности предикатно-аргументных ролевых структур, отражающих особенности картины мира автора текста. Машина выделяет в тексте более ста семантических признаков и семантических ролей; привлечение ряда психолингвистических показателей и специально сформированных тематических групп слов расширяет перечень признаков, выявляемых «Машиной РСА», до 197. Проведенные эмпирические исследования позволили выявить специфику текста, связанную с психологическими особенностями (устанавливаемыми с помощью психодиагностических методик и экспертно), а также психиатрическим статусом (шизофрения и клиническая депрессия) испытуемых — авторов текстов, что открывает перспективы для диагностического и мониторингового применения полученных результатов. Возможность использовать данные, получаемые с помощью «Машины РСА», для машинного обучения делает инструмент адаптивным и развивающимся в соответствии с конкретными исследовательскими задачами.

Ключевые слова: искусственный интеллект, анализ текста, реляционно-ситуационный анализ, психологические особенности, психодиагностика, сетевые интеракции.

Ениколопов Сергей Николаевич — заведующий отделом, отдел медицинской психологии, ФГБНУ «Научный центр психического здоровья», кандидат психологических наук, доцент.
Сфера научных интересов: психология агрессии, психология юмора, клиническая психология.
Контакты: enikolopov@mail.ru

Health Research Center, the Institute of the Russian Language of the Russian Academy of Sciences and the Perm State University. G.S. Osipov proposed a method of relational-situational analysis of text (RSA) based on the ideas of the communicative grammar of the Russian language by G.A. Zolotova, and this approach is implemented in the tool. Analysis of the text using the tool "RSA Machine" turns it into a set of predicate-role structures that are associated with the specifics of the author's worldview. The RSA machine identifies 127 semantic features and semantic roles, and includes a set of psycholinguistic indicators and data of specially created groups of thematic lexis, extracts a total of 197 indicators from texts. The empirical studies have been conducted and have shown some singularities of texts written by people with certain psychological characteristics (according to psychodiagnostic methods and expert opinion), as well as with a psychiatric status (schizophrenia and clinical depression), and these results can be utilized for diagnosis and monitoring. By using the data that the RSA machine receives for further machine learning, it can adapt and develop according with specific research tasks.

Keywords: artificial intelligence, text mining, relational-situational analysis, psychological characteristics, psychodiagnostics.

Sergey N. Enikolopov — Associate Professor, Head of Department of Medical Psychology, Mental Health Research Centre, PhD in psychology.
Research area: psychology of aggression, medical psychology.
E-mail: enikolopov@mail.ru

Осипов Геннадий Семенович — директор института, Институт проблем искусственного интеллекта Федерального исследовательского центра «Информатика и управление» РАН, доктор физико-математических наук, профессор.
Сфера научных интересов: представление знаний, приобретение знаний интеллектуальными системами, динамические интеллектуальные системы, семантический поиск.
Контакты: gos@isa.ru

Кузнецова Юлия Михайловна — старший научный сотрудник, Федеральный исследовательский центр «Информатика и управление» РАН, кандидат психологических наук.
Сфера научных интересов: картина мира, психосемантика, психолингвистика, психодиагностика.
Контакты: kuzjum@yandex.ru

Смирнов Иван Валентинович — заведующий отделом, отдел «Интеллектуальный анализ информации» Федерального исследовательского центра «Информатика и управление» РАН, кандидат физико-математических наук, доцент.
Сфера научных интересов: обработка естественного языка, интеллектуальный анализ информации.
Контакты: ivs@isa.ru

Чудова Наталья Владимировна — старший научный сотрудник, Федеральный исследовательский центр «Информатика и управление» РАН, кандидат психологических наук.
Сфера научных интересов: психология агрессии, психология Интернета, картина мира.
Контакты: nchudova@gmail.com

Gennady S. Osipov — Head of Russian Artificial Intelligence Research Institute, Federal Research Center "Computer Science and Control", Russian Academy of Sciences, PhD and DSc in Mathematics, Full Professor.
Research Area: knowledge representation and reasoning, knowledge acquisition, dynamic intelligent systems, semantic search.
E-mail: gos@isa.ru

Yulia M. Kuznetsova — Senior Research Fellow, Federal Research Center "Computer Science and Control", Russian Academy of Sciences, PhD in psychology.
Research Area: worldview, psychosemantics, psycholinguistics, psychodiagnostics.
E-mail: kuzjum@yandex.ru

Ivan V. Smirnov — Head of department "Intelligent Processing of Information", Federal Research Center "Computer Science and Control", Russian Academy of Sciences, PhD in Mathematics, assistant professor.
Research Area: natural language processing, text and data mining.
E-mail: ivs@isa.ru

Natalia V. Chudova — Senior Research Fellow, Federal Research Center "Computer Science and Control", Russian Academy of Sciences, PhD in psychology.
Research Area: worldview, psychology of Internet, psychology of aggression.
E-mail: nchudova@gmail.com

Посвящаем эту работу памяти нашего коллеги и руководителя Геннадия Семеновича Осипова[1]

Введение

Методы искусственного интеллекта в последние полтора десятилетия все более активно используются для создания компьютерных систем в гуманитарных науках. Применение методов математического моделирования, обращение к цифровым базам данных и применение новых способов извлечения информации из текстов определяют интерес к специализированным интеллектуальным системам отечественных исследователей в области социологии (Михеенкова, 2012), политологии (Бурковская и др., 2004; Ясницкий, 2008), науковедении (Тихомиров и др., 2016), лингвистики (Роганов и др., 2007; Шелманов и др., 2016) и некоторых других гуманитарных наук.

Для психологии особый интерес представляют разрабатываемые в искусственном интеллекте методы обработки естественного языка и интеллектуального анализа текстов. К настоящему времени сформировалось направление междисциплинарных исследований, направленных на обнаружение связей между психологическими особенностями человека и характеристиками порождаемой им текстовой продукции. Особый интерес современных исследователей в качестве источника психодиагностической информации индивидуального, группового и популяционного уровней вызывают тексты сетевого общения (сетевых интеракций), которые становятся предметом международных проектов: SEMIOTIKS (Semantically-Enhanced Information Extraction for Improved Knowledge Superiority), MIMEX (Multivariant Information Management and Exploitation), ITA (International Technology Alliance in Network and Information Sciences), AKT (Advanced Knowledge Technologies), IEXTREME (Extremist Ideological Influences on Group Decision Making). Автоматическому выявлению маркеров личностных качеств и эмоционального состояния пользователей соцсетей посвящены систематически проводимые международные соревнования: INTERSPEECH Emotional Challenge, Speaker State Challenge, Speaker Trait Challenge, Computional Paralinguistic Challenge и др. Методам идентификации текстовых признаков психического неблагополучия посвящены соревнования CLEF и CLPsych.

В качестве базы подобных исследований выступают лингвистические ресурсы, позволяющие выявлять в текстах эмоционально маркированную

[1] Геннадий Семенович Осипов – крупнейший специалист в области искусственного интеллекта, президент Российской ассоциации Искуственного интеллекта (РАИИ) и постоянный член Европейского координационного комитета ИИ (ECCAI), бессменный руководитель Национальных конференций по ИИ в последние 20 лет. Долгие годы Г.С. Осипов поддерживал плодотворные научные контакты с психологами и ориентировал сообщество специалистов по ИИ на междисциплинарные фундаментальные исследования в области когнитивных наук. Он начинал с нами готовить эту работу, под его идейным руководством создан представленный в ней инструмент интеллектуального анализа текста, ориентированный на поддержку психодиагностических и психолингвистических исследований.

лексику: WordNet, WordNet-Affect, SentiWordNet. Наиболее популярным в мировом масштабе средством, применяемым в целях текстовой психодиагностики, является компьютерный инструмент LIWC (Linguistic Inquiry and Word Count), направленный на определение частотности слов различной часторечной принадлежности, аффективной лексики и лексических маркеров тем, имеющих эмпирически подтверждаемую значимость для выявления психологических особенностей автора текста (Pennebaker et al., 2007). Таким образом, решение диагностической задачи основывается преимущественно на традиционных лексических (морфологических) и стилистических характеристиках текста, при этом более глубокие уровни языка не учитываются. К тому же эти подходы и лингвистические ресурсы не могут быть адаптированы прямым переводом к анализу русскоязычных текстов.

Перспективы развития автоматического текстового анализа связываются с созданием методов, ориентированных на семантический уровень текста. Однако автоматический семантический анализ как цель, декларируемая разработчиками имеющихся информационно-аналитических, поисковых и психодиагностических систем, реализуется, как правило, в редуцированном виде, поскольку ограничивается учетом таких, по сути, косвенных данных, как семантические классы или статистические характеристики слов и их сочетаемости, что находится в противоречии с лингвистическими представлениями современности о неразрывной связи между семантикой высказывания и его синтаксисом. Авторский подход к изучению семантики языка образован на понятии значения в смысловом контексте высказывания. Реляционно-ситуационный анализ (РСА) текста (Осипов и др., 2008) представлен положениями коммуникативной грамматики русского языка (Золотова и др., 2004) и оперирует значениями синтаксем — минимальных синтаксических единиц, одновременно являющихся носителями элементарных смыслов.

Реализация идей коммуникативной грамматики русского языка с помощью математического аппарата неоднородных семантических сетей в виде инструмента интеллектуального анализа текста позволяет расширить возможности психологического исследования текстовой продукции русскоязычных авторов. Описанию этого инструмента «Машина РСА» и первых результатов его использования психологами посвящена данная работа.

Положение дел в области автоматического анализа текста

Традиционно автоматический анализ текстов применяется для поиска документов, их классификации, поиска близких документов, извлечения из документов фактов. Цифровое представление текста и его обработка осуществляются в рамках статистического или лингвистического подходов. В рамках первого подхода текст представляется в виде упорядоченного множества последовательностей символов (слов), обработка которого заключается в выявлении статистики встречаемости слов. Второй подход представляется отображением текста в виде набора языковых структур морфологического, синтаксического и семантического уровня. К сожалению, большинство современных

исследований речевой продукции ограничивается морфологическим анализом, при котором устанавливается словарная форма слова (леммы) и его грамматических значений.

В общем виде синтаксический анализ имеет своей целью экспликацию синтаксической структуры текста (Смирнов, Шелманов, 2013) в зависимости от применяемой модели — грамматики непосредственно составляющих (Chomsky, 1957) или грамматики зависимостей (Tesnière, 1959). Грамматика непосредственно составляющих высказывание разделяется на непересекающиеся проективные группы (в пределе — на отдельные слова), где по итогу высказывание представляется в виде иерархии составляющих его синтаксических групп. Описанный формализм отвечает структурной организации высказываний в языках с фиксированным порядком слов. В контексте грамматики зависимостей совершенно иным образом представлено предложение — в виде дерева зависимостей, в котором слова связаны ориентированными дугами, обозначающими синтаксическое подчинение (Апресян, 1995; Hudson, 1984; Melcuk, 1988). Для языков с произвольным порядком слов подходит именно этот формализм. Таким образом, для представления высказываний на английском и на русском языках требуются разные формализмы. Как нетрудно догадаться, эффективное использование в работах с русскоязычными текстами лингвистических анализаторов, построенных для работы с англоязычными текстами, возможно лишь для некоторых типов задач; для получения же данных о психологических особенностях русскоязычных авторов необходимо пользоваться анализаторами, созданными с учетом специфики русского языка.

Технически синтаксический анализ может быть поверхностным (shallow parsing) и глубоким, полным (deep parsing) (Abney, 1991). При поверхностном синтаксическом анализе предложения разделяются на рекурсивно невложенные синтаксические группы (chunking), сегментируются (выделяются отдельные речевые обороты и простые предложения в составе сложного), отражаются в виде поверхностного синтаксического дерева. Т.е. процедура построения полного синтаксического дерева предложения с максимальной связанностью, выявлением дальних связей и определением функций отдельных слов представлена глубоким синтаксическим анализом.

Большинство ранее созданных анализаторов, имеющих функцию глубокого синтаксического анализа (к примеру, коммерческие системы ABBYY (Anisimovich et al., 2012), Xerox XLE (Kaplan et al., 2004) и исследовательский проект ЭТАП-3 (Iomdin et al., 2012)) основано на системе правил. Для более современных разработок характерно привлечение методов машинного обучения с использованием синтаксически размеченных корпусов. При этом эффективность машинного обучения продемонстрирована как для поверхностного, так и для глубокого синтаксического анализа, вытесняя или дополняя подходы, основанные на ручном построении правил и грамматик. Постепенно формируются специальные цифровые ресурсы, обеспечивающие исследователей размеченными под задачи синтаксического анализа корпусами текстов. На основе этих корпусов разрабатываются обучаемые синтаксические

анализаторы, которые достигают высоких показателей качества синтаксического разбора. В качестве примера можно указать на ресурс Universal Dependencies, содержащий размеченные тексты на многих языках, в том числе русском.

Семантический анализ по необходимости опирается на одну из теорий семантики естественного языка (см., например: Апресян, 1974). Наиболее интересной и понятной для психологов представляется предикатно-аргументная (ролевая) семантика, системно представленная в работах Ч. Филлмора (1981). В ее основе лежит понятие падежа, выражающего семантическое содержание аргумента при предикате, или роль. Разработанные в рамках данного подхода приемы установления семантических ролей (semantic role labeling) слов и словосочетаний определили развитие целого направления зарубежных исследований в области «понимания текста» (text understanding). В значительной теоретической работе (Gildea, Jurafsky, 2002) описана схема определения семантических ролей посредством семантического фрейма. В данном процессе устанавливаются либо абстрактные семантические роли «Агенс» и «Пациенс», либо роли, более специфичные для некоторой предметной области. Тем самым фрейм, описывающий коммуникацию, включает роли «Говорящий» (Speaker), «Тема» (Topic), «Средство» (Medium), приписываемые как отдельным словам, так и словосочетаниям. Подобному фрейму, к примеру, соответствует предложение: [$_{\text{Speaker}}$ **We**] **talked** [$_{\text{Topic}}$ **about the proposal**] [$_{\text{Medium}}$ **over the phone**] — в квадратных скобках указаны аргумент и роль. Данный формализм позволяет рассматривать фрейм в качестве средства представления ситуации, включающей участников, их свойства и взаимосвязи, а роль оказывается частью (слотом) такого фрейма. На основе обучения статистических классификаторов с учетом лексических и синтаксических характеристик предложений можно получить вероятностные модели замещения определенным речевым аргументом той или иной позиции в фрейме (Ibid.).

В зарубежных исследованиях семантический анализ чаще всего проводится с применением методов автоматического обучения грамматик и различных классификаторов. Данный подход предназначен для выявления семантического значения синтаксической единицы; определения синтаксических и морфологических характеристик анализируемого предложения; проведения анализа данных о взаимном расположении элементов предложения, об их подчинении, о пути от одной лексической единицы до другой в синтаксическом дереве и типах выявляемых синтаксических групп. Для машинного обучения здесь чаще всего используется статистический, а не индуктивный (логический) метод анализа данных. Выявляемые в результате условные вероятности и их веса в линейной комбинации не подразумевают возможности содержательной интерпретации. Другой недостаток рассмотренных подходов заключается в том, что они задействуют преимущественно лексический уровень языка, что связывает результаты конкретного исследования со строго определенной предметной областью и принадлежащие к ней обучающие корпусы. В анализе русскоязычных текстов необходимо учитывать, что многие используемые в

зарубежных системах признаки (такие как позиция слова в предложении относительно предиката или залог глагола) применительно к русскому языку относительно менее информативны.

Реляционно-ситуационный анализ и «Машина РСА»

В настоящей работе мы представляем для психологов метод искусственного интеллекта, получивший название реляционно-ситуационного анализа (РСА). Он основан на теории коммуникативной грамматики русского языка, созданной в ИРЯ РАН, и на теории неоднородных семантических сетей, созданной в ФИЦ ИУ РАН. Метод широко применяется в задачах поиска релевантной информации, сравнения документов и классификации коллекций текстов (Шелманов и др., 2016; Осипов, Смирнов, 2016). В последние два года метод РСА, реализованный в программно-аппаратном комплексе «Машина РСА», стал применяться для выявления текстовых признаков психологических особенностей авторов (Ениколопов и др., 2019б).

Метод РСА описан (Осипов и др., 2008). Не излагая математические и программистские идеи, реализованные в нашем инструменте, более подробно остановимся на лингвистических идеях, определивших особенности РСА.

Основной идеей коммуникативной грамматики русского языка является утверждение о связи синтаксиса и семантики. Предметом синтаксиса как науки выступают средства и способы построения связной речи и ее коммуникативных единиц — предложений. При этом важно учитывать, где в речи проходят границы синтаксиса и семантики: слово-лексема как единица словарного состава языка, не являясь структурным компонентом высказывания, не передает его смысл. К примеру, в локативных формах *в лесу*, *за лесом*, *над лесом*, *из леса*, *из-за леса*, *возле леса* актуализируется значение места, в творительном пути движения *лесом* — значение протяженности пространства, в объектных формах *губить лес*, *рубить лес*, *сажать лес*, *любоваться лесом* актуализируется предметное значение. Данные значения могут иметь различные синтаксические единицы от слова *лес*, создавая обобщенный тип, могут объединяться с синтаксическими единицами от других слов (*в саду*, *в комнате*). Рассмотренные примеры представлены минимальными синтаксическими единицами с обобщенными значениями (в первом случае локативным) — «в+предл. падеж» и «за+твор. падеж», во втором случае — со значением пути движения — «твор. падеж» и др. Данные синтаксические единицы получили название «синтаксема» — минимальная семантико-синтаксическая единица русского языка, выступающая одновременно как носитель элементарного смысла и как конструктивная компонента более сложных синтаксических построений (Золотова, 1988).

Помимо морфологической формы слова, ключевой ролью в формировании высказывания выступает категориально-семантический класс, к которому принадлежит конкретная лексема. Определяемое этой принадлежностью категориальное значение лексемы задает ее синтаксические возможности и способы функционирования. Поэтому в предложении *«Мама мыла раму»* имя

мама невозможно заменить на любое другое (*белка*, *тарелка*...), поскольку при такой замене предложение перестанет иметь смысл.

Итак, единицей смысла в конкретном предложении выступает слово в составе синтаксемы. Смысл всего высказывания передается с помощью сложных конструкций, образуемых отдельными синтаксемами. Смысл всего текста с точки зрения описываемого подхода восстанавливается путем выделения в нем отдельных синтаксем, установления значений выделенных синтаксем и определения отношения на множестве установленных значений синтаксем. Можно прийти к заключению, что одно из центральных пониманий языкового опосредствования высших психических функций человека в рамках культурно-исторического подхода связано с синтаксемным подходом к анализу речи. Синтаксема, в отличие от значения отдельного слова, лишь указывающего на предмет, отражает ситуацию бытования предмета в культуре, представляет обстоятельства его использования в социуме, характеризует тот класс предметов, которые являются эквивалентными по значению для исполняемого действия. В качестве единицы речевого мышления выступает скорее значение, представленное в синтаксеме, нежели словарное значение слова. На синтаксемном уровне подвергается обобщению сам момент действования с предметом, и предмет предстает в типовой ситуации действования с предметом.

Основанный на коммуникативной грамматике метод РСА позволяет выявлять предикатно-аргументативную структуру высказываний на естественном языке, т.е. представить содержание высказывания в форме действия, события или ситуации, которые выражаются предикатами. Данная модель, не зависящая от предметной области, способна передать семантику почти любого высказывания на естественном языке. В лингвистическом анализаторе «Машины РСА» категориальная семантика глаголов и других предикатных слов отражается в словаре предикатных слов. Семантико-грамматическая классификация глаголов Г.А. Золотовой (1982) выступила теоретическим фундаментом построения такого словаря. Отметим важную для психологов особенность этого подхода. Лингвистический анализатор, проводящий реляционно-ситуационный анализ, опирается на психологически и лингвистически оправданное представление об особой роли глаголов в разворачивании мысли говорящего. О предикативности внутренней речи писал, как известно, Л.С. Выготский, роль предикативной организации высказывания анализировалась в работах А.Р. Лурии и его школы. Лингвисты также подчеркивают особую роль предикатных слов для русскоязычных текстов. Так, например, Г.П. Мельников считает, что «канон внутренней формы языков флективного типа, наиболее последовательно выдерживаемый славянскими языками, а среди славянских — русским, заключается в наличии тенденции по возможности любое сообщение представлять через внешнюю форму канонического предложения, номинативный смысл которого вписывается в такой канон внутренний формы, как образ развивающегося события. Поэтому смысловые поля русских лексем, вещественные и грамматические значения морфем, синтаксические связи между словоформами предложения и т.д. — все это

обусловлено потребностью дать слушающему возможность без труда догадаться, образ какого развивающегося события имел в своем замысле говорящий» (Мельников, 2000, с. 263). Было обнаружено, что доля глаголов от общего числа слов в русском тексте значительно выше, а также «лидерство русского языка по уровню динамичности семантики за счет насыщенности текста глаголами предстает как функционально оправданное своеобразием внутренней формы».

Этап перехода от синтаксем к их значениям и от значений синтаксем к значению предложения предстает в качестве ключевой задачи РСА. С целью реализации данной задачи применяется неоднородная семантическая сеть с расширенным семейством отношений и строится отображение из множества синтаксем в вершине семантической сети. Снятие многозначности отображения нескольких синтаксем во множество их значений осуществляется за счет множества построенных контекстных правил. Ребра полученной сети — элементы отношений на множествах значений. В результате сопоставляются множество значений синтаксем и фрагмент семантической сети на множестве таких значений. Данная конструкция может быть названа семантическим образом предложения, благодаря которому выполняются различные формальные операции: обобщения, конкретизации, сравнения с образами других предложений, исследуются свойства отношений семантической сети и выполняются операции поиска релевантных структур с точки зрения семантического образа предложения и т.д.

Итак, работа нашего лингвистического анализатора состоит из следующих шагов:

Вход: текст на естественном языке.

Выход: неоднородная семантическая сеть.

Шаг 1. Морфологический анализ. Выделение в тексте предложений и слов. Установление для слов нормальных форм и морфологических признаков. Снятие омонимии.

Шаг 2. Синтаксический анализ. Выделение клауз. Установление синтаксических зависимостей между лексемами и выделение синтаксем.

Шаг 3. Реляционно-ситуационный анализ. Выявление значений синтаксем и семантических связей между ними. В результате предложение текста представляется в виде неоднородной семантической сети как совокупность именных групп, ролей и бинарных связей между ними в окрестности одного предикатного слова.

На рисунке 1 приведен пример результата работы анализатора для предложения «Доходы направлены на повышение производства». В узлах полученной сети находятся слова, в качестве ребер выступают синтаксические связи, ролевые связи между глаголом «направлены» и аргументами «доходы» и «повышение», а также семантическое отношение DES (дестинативная связь, один компонент которой обозначает назначение для другого компонента) между указанными аргументами.

В последние годы совершенствование метода происходило за счет разработок в области автоматического установления семантических ролей как наиболее

Рисунок 1

Пример неоднородной семантической сети

сложной и значимой задачи. Впервые для русского языка (Shelmanov, Smirnov, 2014) были разработаны методы установления семантических ролей на основе машинного обучения по размеченным данным с применением подходов самообучения (self-learning). При этом выполняется семантико-синтаксический анализ, когда установление ролей и синтаксический анализ выполняются в единой процедуре.

Далее, с целью установления семантических ролей для русского языка стали использоваться современные нейросетевые подходы (Shelmanov, Devyatkin, 2017), которые обучаются на различных наборах лингвистических признаков с помощью различных архитектур нейронных сетей. Особое внимание уделяется проблеме «неизвестных» предикатов, которые не представлены в обучающей выборке. Установление ролей для таких предикатов происходит с помощью векторных представлений слов (эмбеддингов).

При установлении семантических ролей применяются подходы машинного обучения без учителя (Larionov et al., 2019). Разработан конвейер (pipeline), решающий все подзадачи установления семантических ролей (выявление предикатов, аргументов и их ролей, оптимизация распределения ролей в предложении) с помощью машинного обучения. Для решения проблемы «неизвестных» предикатов используются предобученные языковые модели и эмбеддинги (word2vec, FastText, ELMo, BERT), при этом показано, что эмбеддинги, полученные из предобученных языковых моделей, позволяют получить наилучшее качество установления ролей как для «известных», так и «неизвестных» предикатов.

Указанные выше методы основаны на машинном обучении и требуют размеченных текстов, создание которых может быть трудоемко. В связи с этим нами разработан облегченный метод установления семантических ролей, основанный на правилах, генерируемых из словаря предикатных слов. Принцип работы алгоритма можно разделить на 5 этапов:

1) фильтрация: отбрасываются все тексты, которые не содержат ни одного предиката из имеющегося словаря;

2) анализируемый текст разделяется на предложения, а сами предложения разделяются на клаузы;

3) выделение предикатно-аргументных структур из полученных клауз;

4) обогащение текста морфологической информацией; установление семантических ролей путем применения правил;

5) применение набора ограничений для разрешения возникшей неоднозначности в установленных ролях.

Правила, используемые в алгоритме, представляют собой набор признаков, которые должны быть у анализируемых предиката и аргумента для того, чтобы установить некоторую семантическую роль. Для предиката такими признаками являются: Возвратность, Девербативность, Статус категории состояния. Также есть возможность задать конкретную словоформу для предиката. Для аргумента признаками являются: Одушевленность/Неодушевленность; Падеж аргумента; Наличие какого-либо конкретного предлога перед аргументом. Правила могут действовать как для всех предикатов в словаре, так и для отдельных настраиваемых подмножеств.

Машина РСА как исследовательский инструмент анализа текста, в том числе текстов сетевых интеракций, наряду с показателями частоты встречаемости семантических ролей и семантических связей содержит показатели частотности лексики определенного типа. Лексико-частотный анализ активно используется в системах обработки текста, в зарубежных и отечественных работах. Однако словари «Машины РСА» отличаются двумя особенностями.

Во-первых, лексемы в лексико-частотном модуле «Машины РСА» сгруппированы по принципу тематической принадлежности, т.е. образуют тематические группы слов (ТГС). В отличие от строго лингвистического подхода к созданию словарей метод ТГС предполагает достаточно свободную тактику отбора входящих в них единиц, в основе которой лежит группирование предметов и явлений в соответствии с экстралингвистическим принципом сопряженности с определенной темой. Т.е. при формировании ТГС ведущим критерием является содержательно-тематическая эквивалентность лексем, позволяющая игнорировать такие существенные формальные признаки, как, например, принадлежность лексемы к определенной части речи.

Вторая особенность лексического ресурса «Машины РСА» связана с основной функцией инструмента как средства социогуманитарных исследований. В соответствии с этой направленностью в лексике русского языка искались единицы с семантикой напряжения и психологического неблагополучия. В соответствии с данными, представленными в лингвистических исследованиях и посвященными проблеме языковых средств выражения фрустрации и эмоциональных состояний, были определены темы для ТГС. ТГС, используемые в «Машине РСА», сформированы методом сплошной выборки из материалов следующих словарей: Русский орфографический словарь Российской академии наук, Словарь русской брани (Мокиенко, Никитина), Большой словарь мата (Плуцер-Сарно), Юрислингвистический словарь инвективной лексики русского языка (Голев, Головачева), Словарь современного русского города (Гайдамак и др.), Большой словарь молодежного сленга (Левикова), Словарь молодежного

сленга (Никитина). На данный момент комплекс созданных ТГС представляет собой относительно полный перечень существующих в русском языке лексем — в ТГС психического напряжения входит приблизительно 47 тыс. лексических единиц. Кроме того, разработаны и используются ТГС (ок. 3000 ед.), тематика которых соответствует выделяемым социологами социально-экономическим причинам социального стресса. Таким образом, в лексико-частотном модуле «Машины РСА» задействовано более 50 тыс. лексических единиц.

Лингвистический анализатор «Машины РСА» определяет правильные изменяемые формы лексических единиц, внесенных в ТГС. Вероятностная идентификация применяется в случае форм, образованных с нарушениями правил русского языка или содержащих орфографические ошибки. Результаты анализа представляются количественными данными о частотности лексем из каждой ТГС. Психолингвистический модуль нашего инструмента позволяет вычислять так называемые психолингвистические показатели — преимущественно показатели, описанные в психолингвистических и лингвопсихиатрических исследованиях докомпьютерной эпохи: коэффициент опредмеченности действия (соотношение количества глаголов к количеству существительных), коэффициент Трейгера (отношение количества глаголов к количеству прилагательных), количество существительных и глаголов по сравнению с прилагательными и наречиями и др. Данные показатели в свое время не проверялись на больших корпусах текстов и возможности их использования при автоматическом анализе текстов в качестве маркеров эмоциональности нуждаются в изучении.

В настоящий момент данные «Машины РСА» представляют собой набор из 197 признаков: семантические роли — 92 признака, семантические связи — 35 признаков, словари психического напряжения — 20 признаков, словари социального стресса — 9 признаков, психолингвистические показатели — 27 признаков, части речи — 14 признаков.

Направления использования «Машины РСА» в психологии: примеры проведенных исследований и полученные результаты

Инструмент автоматического анализа текстов в социогуманитарных интересах «Машина РСА» был создан в 2018 г. и продолжает совершенствоваться. В его отладке и в получении с его помощью первых данных наряду с математиками и программистами ФИЦ ИУ РАН принимают участие психологи и лингвисты ФИЦ ИУ РАН, НЦПЗ РАН, ИРЯ РАН, ПермГУ. К настоящему моменту уже накоплен опыт использования в психологических исследованиях данных автоматического анализа текстов и появилась возможность выделить направления работы в данной области.

Поиск текстовых маркеров психологических особенностей

К настоящему моменту с помощью «Машины РСА» выделены текстовые признаки, коррелирующие с результатами психодиагностического исследования и экспертной оценки:

– Агрессивность как личностная черта (повышение частотности в текстах лексики социального разобщения (*вымогатель, бесправный, дразнить* и т.п.), прилагательных и др.); склонность к физической агрессии (высокая частота встречаемости семантической роли «деструктив», глаголов первого лица, лексики страдания — *беда, чахлый, стонать* и т.п.); склонность к гневу (высокая частотность лексики с семантикой отрицательной рациональной оценки — *путаный, абсурд, потерять* и т.п., большое количество знаков препинания, причастий, деепричастий, частиц и др.); высокая враждебность (лексика аффектации и напряжения — *хотеть, волевой, добиваться* и т.п.) (Ковалёв и др., 2019).

– Отсутствие эмпатии как компонент нарциссизма: высокая частотность безысключительной и усилительной лексики, лексики отрицательной рациональной оценки, лексики социальной разобщенности; коэффициент опредмеченности действия; коэффициент Трейгера; высокая частотность всех местоимений и личных местоимений первого и третьего лица; частотность глаголов первого лица единственного числа в прошедшем времени; средняя длина слов (Девяткин, Кузнецова, 2018).

– Актуальное на момент написания текста состояние фрустрации: повышение количества местоимений первого лица, знаков препинания, словоформ, имеющих отрицательные приставки (*ненадежный, бестолково, нигде* и т.п.), слов с семантикой аффектации (*чудовищный, прекрасно, счастье, офигеть* и т.п.), инвектив (*подлецы, фрик, мерзкий, свинство* и т.п.), склонность строить высказывания, содержание которых связано с выявлением причин, указанием на объект реального или потенциального разрушения или ликвидации (синтаксемы: каузатив, ликвидатив, деструктив) (Ениколопов и др., 2019а).

– С помощью применения к данным «Машины РСА» алгоритма установления каузальных связей AQ-JSM выделено 16 лингвистических маркеров клинической депрессии у пациентов НЦПЗ и 27 маркеров депрессивности как личностной особенности, выявленной у здоровых людей с помощью шкалы Бека (Smirnov et al., 2020).

Статистика текстовых параметров как средство изучения психических процессов и состояний, содержание сознания и картины мира

Выявлены следующие факты:

– Предикатно-ролевые структуры отражают на вербальном уровне устройство компонентов или «участков» картины мира субъекта, обычно не становящиеся предметом наблюдения. Так, с помощью «Машины РСА» в текстах эссе «Я. Другие. Мир» выявлены тенденции, характеризующие влияние психического отклонения: частота употребления объекта Я в семантической роли «авторизатор» (субъект оценки, восприятия, речи-мысли) в текстах испытуемых из группы психически больных встречается вдвое реже, чем в текстах здоровых, а в роли «адресат» (лицо, к которому обращено информативное, донативное или эмотивное действие) — в два раза чаще (Кузнецова и др., 2019).

– Выделены два типа текстовой объективации психологического барьера, характеризующего состояние фрустрации в сетевой интеракции: автор, относящийся к типу «деятеля», описывает преимущественно свои собственные действия, а автор, относящийся к типу «наблюдателя», склонен к описанию обстоятельств (Кузнецова и др., 2020).

– Социальный компонент регуляции, отражаемый в комплексе текстовых признаков, значим для пациентов с клинической депрессией и включен в систему их самооценок. К примеру, при шизофрении происходит распространение частно-индивидуального опыта на всех людей, при этом техника включения себя в общий опыт не свойственна; испытуемые данной группы не видят изменений на уровне внутреннего индивидуального опыта даже после видимых проявлений болезни (Ениколопов и др., 2019в).

– Параметром, различающим группы здоровых, больных шизофрений и больных депрессией, выступает морфологическая база обобщенно личного значения: для группы здоровых характерно использование всех трех местоимений «ты», «вы», «мы» и соответствующих форм глагола; для группы с диагнозом «депрессия» более характерны использование местоимения «мы» и возможность присоединить свой опыт к общему; в группе с диагнозом «шизофрения» обобщенное личное значение востребовано достоверно ниже, при этом обобщение если и происходит, то посредством подведения всех под свой собственный опыт (Никитина, Онипенко, 2019).

Проведение психолингвистических исследований и развитие методов сетевой психодиагностики

Помимо определения описанных выше признаков психологического (не)благополучия, отражающихся в сетевом контенте, данное направление работы «Машины РСА» может включать и специфические диагностические приемы. Так, с точки зрения организации широких мониторинговых мероприятий в соцсетях интерес может представлять поиск таких отдельных текстовых показателей, которые в силу количественных и качественных характеристик своих связей с психодиагностическими данными обладают повышенной информативностью. Например, было выявлено, что показатель «отношение числа инфинитивов к общему числу глаголов на текст» взаимосвязан со свойствами психологически благополучной личности: экстраверсия (общительность, раскованность, поиск впечатлений, привлечение внимания), эмоциональная устойчивость (беззаботность, расслабленность, эмоциональный комфорт), добросовестность, склонность к ощущению собственной уникальности, отсутствие социального негативизма и зависти (Кузнецова, 2019).

Другим примером может служить работа по созданию алгоритма распознавания реакций на фрустрацию, встречающихся в тестах пользователей социальных сетей. На первом этапе нами была решена задача создания алгоритма автоматической классификации ответов испытуемых при прохождении Теста фрустрационного реагирования Розенцвейга. В этой работе для каждой категории высказываний лингвистом-русистом были созданы лингвистические

описания, включающие от 3 до 12 правил. Эти описания были формализованы и в ходе машинного обучения использованы для построения выделяемого нашим лингвистическим анализатором признакового описания текстовых фрагментов. Полученные таким способом лингвистические правила формируют высокоуровневое признаковое описание фрагментов текста, позволяющее с высокой полнотой ($F > 0.8$) выявлять высказывания, относящиеся к различным типам реакций. Эти результаты дали основание полагать, что построенные лингвистические правила являются универсальными в отношении самих фрустрирующих ситуаций и что типологически схожие речевые реакции будут наблюдаться в любой ситуации фрустрации, в том числе и в тех ситуациях, которые встречаются в сетевой коммуникации. Однако в отличие от ответов в тесте Розенцвейга сообщения в социальных сетях часто содержат большое количество грамматических ошибок, эллипсов, ненормативной лексики и пр. Поэтому на втором шаге для анализа таких текстов было скомбинировано несколько типов лингвистических признаков, включая лексические, морфологические и высокоуровневые, полученные путем сопоставления текстов с лингвистическими шаблонами, векторные представления предложений, полученные с помощью языковых моделей типа BERT. Далее все эти признаки использовались для обучения простых, хорошо интерпретируемых моделей (ансамблей деревьев решений, ядерных классификаторов). Такой подход позволил создать инструмент для выявления реакций на фрустрацию в сетевых интеракциях, результаты обучения которого интерпретируемы специалистами в области психодиагностики.

Таким образом, результаты эмпирических исследований позволяют выделить в качестве основных перспективных направлений применения «Машины РСА» в области психологических исследований индивидуальную и групповую диагностику психологических особенностей и психического неблагополучия, а также мониторинговые мероприятия, направленные на определение уровня представленности признаков неблагополучия в сетевом контенте и локализации «горячих точек» коммуникативного пространства. В силу статистической природы выявляемых с помощью «Машины РСА» признаков очевидно, что адекватность выводов обеспечивается объемом обрабатываемой текстовой информации, поэтому наилучшие результаты будут демонстрировать мониторинговые исследования. Однако эта связь не является строго линейной — как показывают наши данные, в экстремальных случаях (например, при клинически выраженных отклонениях) своеобразие отдельных текстов выражено настолько ярко, что может проявляться на диагностически значимом уровне.

Дальнейшее развитие «Машины РСА» в интересах психодиагностики предполагает три взаимосвязанных процесса. Собственно психологическая линия развития инструмента включает расширение перечня диагностируемых по текстам особенностей их авторов. В лингвистическом отношении требуется разработка адекватных психологическим категориям тематических групп слов (например, для диагностики клинических нарушений — групп морбуальной лексики, метафоры замкнутого пространства), создание моделей

разорванности и синтаксического однообразия текста, синтаксиса повествования и рассуждения, способов выделения критериев энтропии сирконстантовых и детерминантов, формирование групп предикатов в соответствии с их семантическим содержанием и др. (Ениколопов, Мишланов, 2019; Кузнецова и др., 2019). Совершенствование программной составляющей «Машины РСА» осуществляется путем реализации ее модульной концепции, обеспечивающей адаптивность инструмента при решении различающихся по предмету исследовательских задач.

Результаты представленных в настоящей статье исследований и разработок применяются в TITANIS — новом инструменте для анализа текста из социальных сетей, предназначенном для изучения реакций общества на различные значимые события с точки зрения психоэмоционального анализа (Smirnov et al., 2021). Инструмент предлагает набор текстовых параметров и методов для обработки естественного языка, которые позволяют оценить психологические состояния авторов текстов в социальных сетях. Помимо широко используемых сегодня подходов к обработке текстов, таких как tf-idf и анализ тональности, TITANIS соедержит психолингвистический, семантический, дискурсивный и другие виды анализа, позволяющие выявлять различия в текстах пользователей с разными психоэмоциональными состояниями. К настоящему моменту словарь предикатных слов в TITANIS значительно расширен за счет класса так называемых эмотивов; на основе установления семантических ролей при эмотивных предикатах TITANIS позволяет определить по тексту, ***кто*** [субъект] и ***от чего*** [причина] испытывает определенную эмоцию, например, *Мы* [субъект] *обрадовались подарку* [причина]. TITANIS представляет собой библиотеку на языке программирования Python и является промышленным решением с прикладным программным интерфейсом (API), обеспечивающим встраивание инструмента в сторонние системы или использование инструмента в научных исследованиях с минимальными затратами на развертывание. Пробная версия TITANIS с открытом кодом и ограниченной функциональностью доступна в Интернете (TITANIS, 2021).

Заключение

В рамках работ, ориентированных на создание методов автоматизации работы с текстом для специалистов социогуманитарного профиля, лингвистика играет роль медиатора между психологией (или другой дисциплиной социогуманитарной области) и математикой. Роль лингвистов состоит, прежде всего, в моделировании признаков речевой системности того или иного вида, возникающей при объективации в речи различных психических процессов и состояний. В свою очередь, специалисты по ИИ могут не только предоставить психологам готовые средства анализа текста, построенные на достижениях мировой и отечественной лингвистики, но и совместно с психологами решать задачи развития исследовательских методов, которые опираются на математические модели и программные средства, разработанные в области анализа неструктурированной информации.

Литература

Апресян, Ю. Д. (1974). *Лексическая семантика: Синонимические средства языка*. М.: Наука.

Апресян, Ю. Д. (1995). *Избранные труды: Т. 2. Интегральное описание языка и системная лексикография*. М.: Школа «Языки русской культуры».

Бурковская, Ж. И., Михеенкова, М. А., Финн, В. К. (2004). Об интеллектуальной системе для анализа электорального поведения. В кн. *IX Национальная конференция с международным участием «Искусственный интеллект-2004». Тверь, сентябрь 8–11, 2004 г. Труды конференции* (т. 1, с. 120–128). М.: Физматлит.

Девяткин, Д. А., Кузнецова, Ю. М. (2018). Психолингвистические и лексические маркеры нарциссизма. В кн. *Восьмая международная конференция по когнитивной науке: Тезисы докладов. Светлогорск, 18–21 октября 2018 г.* (с. 324–326). М.: Изд-во «Институт психологии РАН».

Ениколопов, С. Н., Ковалёв, А. К., Кузнецова, Ю. М., Чудова, Н. В., Старостина, Е. В. (2019а). Признаки, характерные для текстов, написанных в состоянии фрустрации. *Вестник Московского университета. Серия 14. Психология, 3*, 66–85.

Ениколопов С. Н., Кузнецова Ю. М., Смирнов И. В., Станкевич М. А., Чудова Н. В. (2019б). Создание инструмента автоматического анализа текста в интересах социо-гуманитарных исследований. Ч. 1. Методические и методологические аспекты. *Искусственный интеллект и принятие решений, 2*, 28–38.

Ениколопов, С. Н., Медведева, Т. И., Воронцова, О. Ю. (2019в). Лингвистические особенности текстов людей с разным психическим статусом. *Вестник Московского государственного областного университета (Электронный журнал), 3*, 119–128. https://evestnik-mgou.ru/ru/Articles/Doc/965

Ениколопов, С. Н., Мишланов, В. А. (2019). Особенности речевой организации текстов, порождаемых людьми с психическими отклонениями (к проблеме автоматического выявления лингвистических маркеров психического неблагополучия). *Филология в XXI веке, 1*(3), 22–30.

Золотова, Г. А. (1982). *Коммуникативные аспекты русского синтаксиса*. М.: Наука.

Золотова, Г. А. (1988). *Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса*. М.: Наука.

Золотова Г. А., Онипенко Н. К., Сидорова М. Ю. (2004). *Коммуникативная грамматика русского языка*. М.: Институт русского языка РАН им. В.В. Виноградова.

Ковалёв, А. К., Кузнецова, Ю. М., Минин, А. Н., Пенкина, М. Ю., Смирнов, И. В., Станкевич, М. А., Чудова, Н. В. (2019). Методы выявления по тексту психологических характеристик автора (на примере агрессивности). *Вопросы кибербезопасности, 4*(32), 72–80.

Кузнецова, Ю. М. (2019). Относительное количество инфинитивов как показатель текстовой диагностики личности. В кн. *Теория речевой деятельности: вызовы современности. Материалы XIX международного симпозиума по психолингвистике и теории коммуникации. Москва, 06–08 июня 2019 г.* (с. 117–118). М.: Изд-во «Канцлер».

Кузнецова, Ю. М., Курузов, И. А., Смирнов, И. В., Станкевич, М. А., Старостина, Е. В., Чудова, Н. В. (2020). Текстовые проявления фрустрированности пользователя социальных сетей. *Медиалингвистика, 1*, 4–16.

Кузнецова, Ю. М., Смирнов, И. В., Станкевич, М. А., Чудова, Н. В. (2019). Создание инструмента автоматического анализа текста в интересах социо-гуманитарных исследований. Ч. 2. Машина РСА и опыт ее использования. *Искусственный интеллект и принятие решений, 3*, 40–51.

Мельников, Г. П. (2000). *Системная типология языков: синтез морфологической классификации языков со стадиальной*. М.: Изд-во РУДН.

Михеенкова, М. А. (2012). *Принципы и логические средства интеллектуального анализа социологических данных* [Докторская диссертация, Всероссийский институт научной и технической информации РАН (ВИНИТИ РАН)].

Никитина, Е. Н., Онипенко, Н. К. (2019). Когнитивно-лингвистическая интерпретация результатов автоматического анализа текстов психически больных. *Искусственный интеллект и принятие решений, 3*, 60–69.

Осипов, Г. С., Смирнов, И. В. (2016). Семантический анализ научных текстов и их больших массивов. *Системы высокой доступности*, *1*, 41–44.

Осипов, Г. С., Смирнов, И. В., Тихомиров, И. А. (2008). Реляционно-ситуационный метод поиска и анализа текстов и его приложения. *Искусственный интеллект и принятие решений*, *2*, 3–10.

Роганов, В. Р., Роганова, С. М., Новосельцева, М. Е. (2007). *Методы искусственного интеллекта для машинного перевода текстов*. Пенза: Изд-во Пензенского государственного университета.

Смирнов, И. В., Шелманов, А. О. (2013). Семантико-синтаксический анализ естественных языков. Часть I. Обзор методов синтаксического и семантического анализа текстов. *Искусственный интеллект и принятие решений, 1,* 41–54.

Тихомиров, И. А., Соченков, И. В., Швец, А. В. (2016). Наукометрия и полнотекстовая аналитика в российских реалиях. В кн. *Науковедческие исследования, 2016: Сборник научных трудов* (с. 197–212). М.: ИНИОН РАН.

Филлмор, Ч. (1981). Дело о падеже (В. А. Звегинцев, пер. с англ.). В кн. *Новое в зарубежной лингвистике: Вып. X. Лингвистическая семантика* (с. 369–495). М.: Прогресс.

Шелманов, А. О., Каменская, М. А., Ананьева, М. И., Смирнов, И. В. (2016). Семантико-синтаксический анализ текстов в задачах вопросно-ответного поиска и извлечения определений. *Искусственный интеллект и принятие решений, 4,* 47–61.

Ясницкий, Л. Н. (2008). О возможностях применения методов искусственного интеллекта в политологии. *Вестник Пермского университета. Серия: Политология, 2,* 147–155.

Ссылки на зарубежные источники см. в разделе References.

References

Abney, S. P. (1991). Parsing by chunks. In R. C. Berwick, S. P. Abney, & C. Tenny (Eds.), *Studies in linguistics and philosophy: Vol. 44. Principle-based parsing: Computation and psycholinguistics* (pp. 257–278). Dordrecht: Kluwer Academic.

Anisimovich, K. V., Druzhkin, K. Ju., Minlos F. R., Petrova M. A., Selegey V. P., & Zuev, K. A. (2012). Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2012): Vol. 2. Papers from special sessions* (Iss. 11(18), pp. 91–104). Moscow: RGGU.

Apresyan, Yu. D. (1974). *Leksicheskaya semantika: sinonimicheskie sredstva yazyka* [Lexical semantics: synonymous means of language]. Moscow: Nauka.

Apresyan, Yu. D. (1995). *Izbrannye trudy: T. 2. Integral'noe opisanie yazyka i sistemnaya leksikografiya* [Selected Works: Vol. 2. The integrated description of language and system lexicography]. Moscow: Shkola "Yazyki russkoy kul'tury".

Burkovskaya, Zh. I., Miheenkova, M.A., & Finn, V. K. (2004). Ob intellektual'noi sisteme dlya analiza elektoral'nogo povedeniya [Intellectual system for electoral behavior analysis]. In *IX Natsional'naya konferentsiya s mezhdunarodnym uchastiem "Iskusstvennyj intellekt-2004". Tver', sentyabr' 8–10, 2004 g. Trudy konferentsii* [IX National Conference with International Participation "Artificial Intelligence-2004". Tver', 2004, September 8–11. Proceedings of the Conference] (Vol. 1, pp. 120–128). Moscow: Fizmatlit.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Devyatkin, D. A., & Kuznetsova, Yu. M. (2018). Psiholingvisticheskie i leksicheskie markery narcissizma [Psycholinguistic and lexical markers of narcissism]. In *Vos'maya mezhdunarodnaya konferentsiya po kognitivnoj nauke: Tezisy dokladov. Svetlogorsk, 18–21 oktyabrya 2018 g.* [The Eighth International Conference on Cognitive Science. 2018, October 18–21, Svetlogorsk, Russia. Abstracts] (pp. 324–326). Moscow: Institute of Psychology of the RAS.

Enikolopov, S. N., & Mishlanov, V. A. (2019). Peculiarities of speech organization of the texts produced by people with mental deviations (to the problem of automatic detection of linguistic markers of mental distress). *Filologiya v XXI Veke, 1*(3), 22–30. (in Russian)

Enikolopov, S. N., Kovalev, A. K., Kuznetsova, Yu. M., Chudova, N. V. & Starostina, E. V. (2019a). Features of texts written by a frustrated person. *Moscow University Psychology Bulletin, 3,* 66–85. (in Russian)

Enikolopov, S. N., Kuznetsova, Yu. M., Smirnov, I. V., Stankevich, M. A., & Chudova, N. V. (2019b). Creating a tool for automatic text analysis in the interests of socio-humanitarian research. Part 1: Methodical and Methodological Aspects. *Iskusstvennyi Intellekt i Prinyatie Reshenii, 7,* 28–38. (in Russian)

Enikolopov, S., Medvedeva, T., & Vorontsova, O. (2019c). Linguistic characteristics of texts of people with different mental status. *Vestnik Moskovskogo gosudarstvennogo oblastnogo universiteta (Elektronnyj zhurnal) [Bulletin of Moscow State Regional University (electronic journal)], 3,* 119–128. https://evestnik-mgou.ru/ru/Articles/Doc/965 (in Russian)

Fillmore, Ch. J. (1981). Delo o padezhe [The case of case]. In Novoe v zarubezhnoi lingvistike: Iss. X. Lingvisticheskaya semantika [New in foreign linguistics: Iss. X. Linguistic semantics]. Moscow: Progress.

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics, 28*(3), 245–288.

Hudson, R. (1984). *Word grammar*. Oxford: Basil Blackwell.

Iomdin, L., Petrochenkov, V., Sizov V., & Tsinman, L. (2012). ETAP parser: state of the art. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2012): Vol. 2. Papers from special sessions* (Iss. 11(18), pp. 119–131). Moscow: RGGU.

Kaplan, R. M., Riezler, S., King, T. H., Maxwell, J. T., & Vasserman, A. (2004). Speed and accuracy in shallow and deep stochastic parsing. In *Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT/NAACL'04)* (pp. 97–104). https://aclanthology.org/N04-1013.pdf

Kovalev, A. K., Kuznetsova Y. M., Minin, A. N., Penkina, M. Y., Smirnov, I. V., Stankevich, M. A., & Chudova, N. V. (2019). Text analysis approach for identifying psychological characteristics (with aggressiveness as an example). *Voprosy Kiberbezopasnosti, 4*(32), 72–80. (in Russian)

Kuznetsova, Yu. M. (2019). Otnositel'noe kolichestvo infinitivov kak pokazatel' tekstovoi diagnostiki lichnosti [Relative number of infinitives as an indicator of textual personality diagnostics]. In *Teoriya rechevoi deyatel'nosti: vyzovy sovremennosti. Materialy XIX mezhdunarodnogo simpoziuma po psikholingvistike i teorii kommunikatsii. Moscow, 06–08 iyunya 2019 g.* [Theory of speech activi-

ty: challenges of modernity. Proceedings of the XIX International Symposium on Psycholinguistics and Communication Theory. Moscow, 2019, June 06–08] (pp. 117–118). Moscow: Kantsler.

Kuznetsova, Yu. M., Kuruzov, I. A., Smirnov, I. V., Stankevich, M. A., Starostina, E. V., & Chudova, N. V. (2020). Textual manifestations of frustration of a user of social networks. *Medialingvistika [Media Linguistics Journal], 7*(1), 4–16. (in Russian)

Kuznetsova, Yu. M., Smirnov, I. V., Stankevich, M. A., & Chudova, N. V. (2019). Creating a text analysis tool for socio-humanitarian research. Part 2. RSA machine and the experience of using it. *Iskusstvennyi Intellekt i Prinyatie Reshenii, 3,* 40–51. (in Russian)

Larionov, D., Shelmanov, A., Chistova, E., & Smirnov, I. (2019). Semantic role labeling with pretrained language models for known and unknown predicates. In *Proceedings of International Conference on Recent Advances of Natural Language Processing. Varna, Bulgaria, 2019, Sep 2–4* (pp. 619–628). Shoumen, Bulgaria: INCOMA Ltd. https://doi.org/10.26615/978-954-452-056-4_073

Melcuk, I. (1988). *Dependency syntax: theory and practice*. Albany, NY: The SUNY Press.

Melnikov, G. P. (2000). *Sistemnaya tipologiya yazykov: sintez morfologicheskoi klassifikatsii yazykov so stadial'noi* [System typology of languages: Synthesis of morphological classification of languages with stadial one]. Moscow: RUDN University.

Mikheenkova, M. A. (2012). *Printsipy i logicheskie sredstva intellektual'nogo analiza sotsiologicheskih dannyh* [Principles and logical means of intellectual analysis of sociological data] [Doctoral dissertation, Russian Institute of Scientific and Technical Information of the Russian Academy of Sciences].

Nikitina, E. N., & Onipenko, N. K. (2019). A cognitive linguistic interpretation of statistical analysis results based on texts by persons with mental disorder. *Iskusstvennyi Intellekt i Prinyatie Reshenii, 3,* 60–69. (in Russian)

Osipov, G. S., & Smirnov, I. V. (2016). Semantic analysis of large-scale collections of scientific texts. *Sistemy Vysokoi Dostupnosti, 1,* 41–44. (in Russian)

Osipov, G. S., Smirnov, I. V., & Tikhomirov, I. A. (2008). Relyatsionno-situatsionnyi metod poiska i analiza tekstov i ego prilozheniya [Relational-situational method of search and analysis of texts and its applications]. *Iskusstvennyi Intellekt i Prinyatie Reshenii, 2,* 3–10.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. Austin, TX: LIWC.net. https://www.researchgate.net/publication/228650445_The_Development_and_Psychometric_Properties_of_LIWC2007

Roganov, V. R., Roganova, S. M., & Novosel'tseva, M. E. (2007). *Metody iskusstvennogo intellekta dlya mashinnogo perevoda tekstov* [Methods of artificial intelligence for machine translation of texts]. Penza: Izdatel'stvo Penzenskogo gosudarstvennogo universiteta.

Shelmanov, A. O., & Devyatkin, D. A. (2017). Semantic role labeling with neural networks for texts in Russian. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2017)* (Iss. 16, pp. 245–256). Moscow: RGGU.

Shelmanov, A. O., Kamenskaya, M. A., Ananieva, M. I., & Smirnov, I. V. (2016). Cemantiko-sintaksicheskij analiz tekstov v zadachah voprosno-otvetnogo poiska i izvlecheniya opredelenij [Semantic-syntactic analysis for question answering and definition extraction]. *Iskusstvennyi Intellekt i Prinyatie Reshenii, 4,* 47–61.

Shelmanov, A. O., & Smirnov, I. V. (2014). Methods for semantic role labeling of russian texts. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2014)* (Iss. 13(20), pp. 580–592). Moscow: RGGU.

Smirnov, I. V., & Shelmanov, A. O. (2013). Semantic and syntactic analysis of natural languages. Part I. Review of methods for syntactic and semantic analysis of text. *Iskusstvennyi Intellekt i Prinyatie Reshenii, 1,* 41–54. (in Russian)

Smirnov, I., Stankevich, M., Kuznetsova, Y., Suvorova, M., Larionov, D., Nikitina, E., Savelov, M., & Grigoriev, O. (2021). TITANIS: A tool for intelligent text analysis in social media. In *Russian Conference on Artificial Intelligence. The Nineteenth Russian Conference on Artificial Intelligence RCAI-2021. Taganrog, Russia, 2021, October 11–16*. Cham: Springer.

Smirnov, I. V., Ushakova, A. V., & Chudova, N. V. (2020). Method for detecting text markers of depression and depressiveness. In *Russian Conference on Artificial Intelligence* (pp. 325–337). Cham: Springer. https://doi.org/10.1007/978-3- 030-59535-7_24

Tesnière, L. (1959). *Elements de syntaxe structurale*. Paris: Klincksieck.

Tikhomirov, I. A., Sochenkov, I. B., & Shvets, A.V. (2016). Naukometriya i polnotekstovaya analitika v rossijskih realiyah [Scientometrics and full-text analytics in the Russian realias]. In *Naukovedcheskie issledovaniya, 2016: Sbornik nauchnyh trudov* [Science research, 2016: Scientific papers collection] (pp. 197–212). Moscow: INION RAN.

TITANIS: A Tool for Intelligent Text Analysis in Social Media. (2021, September 15). URL: https://github.com/tchewik/titanis-open

Yasnitsky, L. N. O vozmozhnostyah primeneniya metodov iskusstvennogo intellekta v politologii [The opportunities of application of artificial intelligence methods in political science]. *Vestnik Permskogo universiteta. Seriya: Politologiya [Bulletin of Perm University. Political Science], 2,* 147–155.

Zolotova, G. A. (1982). *Kommunikativnye aspekty russkogo sintaksisa* [The communicative aspects of the Russian syntax]. Moscow: Nauka.

Zolotova, G. A. (1988). *Sintaksicheskii slovar': Repertuar elementarnykh edinits russkogo sintaksisa* [Syntax dictionary: The repertoire of elementary units of Russian syntax]. Moscow: Nauka.

Zolotova, G. A., Onipenko, N. K., & Sidorova, M. Yu. (2004). *Kommunikativnaya grammatika russkogo yazyka* [Communicative grammar of the Russian language]. Moscow: Institut russkogo yazyka RAN im. V.V. Vinogradova.