РЕЦЕНЗИЯ
НА СБОРНИК ТРУДОВ МЕЖДУНАРОДНОЙ
КОНФЕРЕНЦИИ «КОРПУСНАЯ ЛИНГВИСТИКА-2019»
РАЗДЕЛ «Семантика и извлечение информации из корпусов»

Сборник трудов международной научной конференции «КОРПУСНАЯ ЛИНГВИСТИКА-2019» освещает направления современной корпусной лингвистики. Одним из традиционных направлений являются извлечение информации из специализированных корпусов текстов, к нему присоединяют в настоящее время корпусное обеспечение семантических разработок. В рецензируемом разделе представлены статьи отечественных и зарубежных исследователей, посвященные актуальным вопросам создания и использования методологии корпусного анализа отдельных языковых единиц и текстов, а также математических и компьютерных методов в процессе создания и использования корпусов.

В статье ***О.В. Блиновой, С.А. Белова*** **«РУССКИЕ ОФИЦИАЛЬНЫЕ ТЕКСТЫ ДОМЕНА «ЗДРАВООХРАНЕНИЕ» И ОЦЕНКА ИХ ЛЕКСИЧЕСКОЙ СЛОЖНОСТИ С ИСПОЛЬЗОВАНИЕМ КЛЮЧЕВЫХ СЛОВ»** представлен классический подход к выделению ключевых слов в текстах, описывающих документооборот в сфере здравоохранения, при этом специфические слова выделяется на фоне частотности слов в современных частотных словарях. Данное исследование имеет большое значение в силу необходимости уточнения лексических параметров делового стиля при оформлении таких документов, как договор на оказание платных услуг, правил госпитализации и проч. с целью обеспечить однозначное и непротиворечивое понимание договорных условий между человеком-пациентом и медицинским учреждением.

В статье ***I. Kapič*** **«AUTOMATIC TERM EXTRACTION – EFFICIENCY OF SELECTION AND RELEVANCE OF EXTRACTED TERMS AS APPLIED TO THE SPECIALIZED CORPUS OF LIBRARY AND INFORMATION SCIENCE IN SLOVENE LANGUAGE»** представлен способ автоматического выделения терминов из корпуса по тематике информационного поиска и библиотечного дела размером в 4 млн токенов при помощи системы Sketch Engine. Описываются проблемы морфо-синтаксической сложности словенского языка, Выделенная совокупность на 60% состоит из терминов, представленных в терминологических словарях для данной области, что показывает перспективность предлагаемого подхода.

В статье ***А.А. Новиковой*** **«ИСПОЛЬЗОВАНИЕ ИНСТРУМЕНТА SKETCH ENGINE ДЛЯ ИЗВЛЕЧЕНИЯ ТЕРМИНОЛОГИИ»** используется тот же инструментарий для анализа текстов предметной области «Водоснабжение и водоотведение» для текстов на трех языках: русском, английском и немецком. Как и в предыдущем случае, Sketch Engine показал свою эффективность при выделении терминов в текстах на английском и русском языках: соответственно 66% и 77%

совпадения выделенных кандидатов с эталонными списками. Однако для немецкого языка этот показатель не превышает 16%, что требует отдельно объяснения.

В статье ***V. Bobicev, Y. Hlavcheva, O. Kanishcheva, V. Lazu*** **«AUTHORSHIP ATTRIBUTION IN SCIENTIFIC PUBLICATIONS»** в качестве анализа авторства рассматривается атрибуция языка научных текстов: украинского или русского. Используется несколько алгоритмов на платформе Weka: байесовский классификатор (простой и по линомиальный), метод опорных в екторов (SVM) и авторский метод символьной компрессии PPM. Анализируются тексты разных предметных областей: от информационного поиска до экономики. Результаты показывают, что авторский метод дает наилучшие результаты комплексной F-меры выполнения поставленной задачи даже для небольших текстов объемом 150 слов.

В статье ***М.Н. Михайлова, Ю.В. Соума*** **«СОЗДАНИЕ ПАРАЛЛЕЛЬНОГО КОРПУСА МЕЖГОСУДАРСТВЕННЫХ ДОГОВОРОВ: ТЕХНИЧЕСКИЕ И МЕТОДОЛОГИЧЕСКИЕ ПРОБЛЕМЫ»** обсуждаются вопросы баланса корпуса финско-русских параллельных текстов. В задачу исследователей входит анализ лексического состава в текстах международных договоров трех хронологических разделов: 1917-1944, 1945-1991, 1992-2016 гг. Наблюдается существенный дисбаланс в количестве договоров в эти периоды: количество документов во втором разделе практически суммарно равно объему первого и третьего разделов вместе взятых. Авторы видят способ достижения баланса в пополнении третьей группы договорами более «низкого уровня»: между компаниями, городами и проч.

В статье ***N.B. Krizhanovskaya, A.A. Krizhanovsky*** **«SEMI-AUTOMATIC METHODS FOR ADDING WORDS TO THE DICTIONARY OF VEPKAR CORPUS BASED ON INFLECTIONAL RULES EXTRACTED FROM WIKTIONARY»** описывается процедура пополнения словаря морфологической разметки в Открытом корпусе вепсских и карельских текстов VepKar. Для этой цели используется система динамических шаблонов Английского Викисловаря, используя которую порождается система примерно из 40 форм для вепсских глаголов. Затем список полученных форм просматривается экспертами, носителями языка для внесения корректировки.

В статье ***В.А. Шульгинова, В.А. Шульгинова*** **«КОРПУСНОЕ ИССЛЕДОВАНИЕ АВТОРСКОЙ РЕЦЕПЦИИ В СТРУКТУРЕ ЭЛЕКТРОННОГО ГИПЕРТЕКСТА»** показано использование нескольких стратегий, реализуемых в текстовой части гиперссылок, которые собраны автоматически по электронным новостным ресурсам в корпус гипертекстем — совокупностей целевого и исходного текстов с гиперссылкой плюс номинация ссылок. Первая стратегия гиперссылки предполагает сильную семантическую связь, при которой целевой текст выступает в качестве пресуппозиции для исходного. Вторая стратегия обеспечивает доказательность текста, чаще всего реализуется в виде глагола, для которого целевой текст заполняет предикатно-аргументные позиции.

В статье ***VB. Broz, Vlatko*** **«A CORPUS-BASED CRITICAL DISCOURSE ANALYSIS OF BREXIT IN THE ENGLISH LANGUAGE PRESS»** рассматриваются коллокации нового понятия «Брекзит» в текстах английской прессы. Отмечается, что в течение короткого периода времени возникла система метафорических определений явления (*мягкий* и *жесткий*), которые описывают 2 варианта реализации выхода Великобритании из Евросоюза. Показано, что число определений с отрицательной тональностью значительно превосходит положительные оценки, чт о показывает существующее в стране настроение общества по отношению к Евросоюзу.

В статье ***R.R. Rebechi*** **«'GOD', 'NATION' AND 'FAMILY' IN THE IMPEACHMENT OF A BRAZIL'S PRESIDENT: A CORPUS-BASED APPROACH TO DISCOURSE»** рассматривается дискурсивный отклик на политическое событие — импичмент Дилмы Русефф — в выступлениях представителей бразильского парламента, которые голосовали «за» или «против» отставки президента. Испо льзуется мера log-likelyhood для выделения слов, коррелирующих с той или иной позицией голосующих. Полученные списки из 101 и 65 слов отчасти отражают дисбаланс текстов «за» и «против» отставки. Дальнейший дискурсивный анализ позволил показать, что слова, связанные с концептами «бог» и «семья», использовались примерно в равной степени в текстах той или иной группы, вероятно, в разных значениях. А вот слова, связанные с концептом «нация», чаще использовались в выступлениях «за» отставку.

В статье ***С.Ю. Семеновой*** и ***А.С. Паниной*** **«ОПЫТ ИСПОЛЬЗОВАНИЯ ДАННЫХ НКРЯ ПРИ ОПИСАНИИ ПОЛИСЕМИИ В ПРИКЛАДНОМ СЕМАНТИЧЕСКОМ СЛОВАРЕ»** описывается использование статистики НКРЯ при создании специального семантического словаря РУСЛАН, ориентированного на использование в системах автоматического анализа, что накладывает ограничения на количество выделяемых значений по лисемантичных слов — не более 5. На начальных этапах разработки словарь опирался на интроспекцию своего автора Н.Н. Леонтьевой. Данные словаря корректируются как в плане формулировки значений, так и в плане структуры валентностей, наличия устойчивых сочетаний и проч. на базе корпусной статистики, при этом сравниваются статистические данные в подкорпусах НКРЯ.

В статье ***V.P. Zakharov, I.V. Azarova*** **«TOWARDS A COMPUTATIONAL ONTOLOGY OF RUSSIAN PREPOSITIONS»** рассматривается построение семантико-грамматического описания предложных конструкций в русском языке, которое создается на основании данных больших современных корпусов русских текстов. В семантическом плане используется модифицированная система синтаксем Г.Золотовой, которые объединяются к группы сходных, квазисинонимичных конструкций, образуя так называемые «рубрики». В статье представлено описание медиативной и транзитивной рубрик.

В статье ***Линь Цзиньфэн, Д.М. Семёновой, С.Л. Пущина, Т.Г. Петрова, М.Н. Бабарико, С.В. Чебанова*** **«РУЧНАЯ РАЗМЕТКА КОРПУСА ДЛЯ**

ИЗУЧЕНИЯ СТАТИСТИКИ КОНЦЕПТОВ» представлено оригинальное исследование ручной разметки корпусов пословиц для русского и китайского языков. Базовая трихотомия «тело-дух-душа» находит свое отражение в текстах в виде лексических единиц (слов и словосочетаний), а также в виде косвенных упоминаний (описаний и аллюзий). В отношении первого компонента триады «тело» вводится 5 уровней на основании отношения «часть-целое». Описывая соотношения рангов частот упоминания концептов в полученных файлах, авторы приходят к выводам о закономерностях в выделении ядерных концептов в корпусах китайских и русских пословиц.

Материалы раздела представляют собой оригинальные научные исследования по актуальным вопросам извлечения информации из текстов и его семантическому анализу с использованием статистических корпусных данных и технологий для эффективного корпусного анализа. Результаты исследований имеют важную теоретическую и практическую значимость.

Целесообразность публикации сборника и соответствие его материалов высокому научному стандарту не вызывают сомнения. Сборник трудов Международной конференции «КОРПУСНАЯ ЛИНГВИСТИКА - 2019», может быть РЕКОМЕНДОВАН К ПЕЧАТИ.

РЕЦЕНЗЕНТЫ

канд.филол.наук, доцент кафедры математической лингвистики СПбГУ
Азарова И.В.
ivazarova@gmail.com

канд.филол.наук, доцент кафедры математической лингвистики СПбГУ
Митренина О.В.
mitrenina@gmail.com

ЛИЧНУЮ ПОДПИСЬ ЗАВЕРЯЮ Азаровой И. В., Митрениной О. В.
ДОКУМЕНТ ПОДГОТОВЛЕН ПО ЛИЧНОЙ ИНИЦИАТИВЕ

17 ИЮН 2019

ТЕКСТ ДОКУМЕНТА РАЗМЕЩЕН В ОТКРЫТОМ ДОСТУПЕ НА САЙТЕ СПбГУ ПО АДРЕСУ HTTP://SPBU.RU/SCIENCE/EXPERT.HTML

ведущий специалист по кадрам
Селиверстова В. В.

ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
Санкт-Петербургский государственный университет (СПбГУ)
ИНН 7801002274