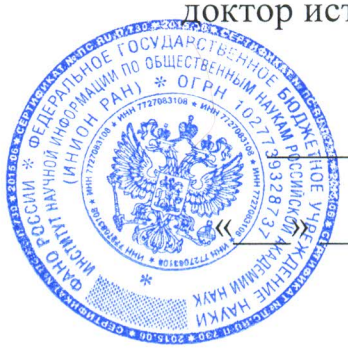
УТВЕРЖДАЮ

Вр. и. о. директора ФГБУН ИНИОН РАН

доктор исторических наук, доцент

Зайцев И.В.



«____» _____ 2016 г.

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

о диссертации Куликова Сергея Юрьевича

«Автоматическое извлечение мнений: лингвистический аспект»,
представленной на соискание
ученой степени кандидата филологических наук

по специальности 10.02.21 – Прикладная и математическая лингвистика

Актуальность исследования определяется необходимостью повышения эффективности функционирования автоматизированных систем идентификации оценочных высказываний по письменному Интернет-дискурсу, а также потребностью привлечения лингвистического обеспечения с целью модернизации существующих систем автоматического извлечения мнений.

Научная новизна исследования заключается в разработке лингвистических принципов автоматизированного составления ресурсов для систем автоматического извлечения мнений общего назначения применительно к русскоязычному текстовому интернет-контенту, а также в дополнении существующих классификаций оценочной лексики.

Теоретическая значимость работы состоит в комплексном описании лингвистического компонента в составе систем автоматического извлечения мнений общего назначения, специализированных систем, а также систем идентификации противозаконного контента.

Практическая значимость исследования реализуется в разработке рекомендаций по повышению эффективности (повышение точности анализа, полноты выдачи материала, увеличение скорости обработки данных) существующих систем автоматического извлечения мнений за счет совершенствования лингвистического обеспечения в их составе.

Разработанная в диссертационном исследовании методика может быть использована при создании словарей оценочной лексики, баз данных, размеченных корпусов текстов (на русскоязычном материале).

Материалом экспериментальной части исследования послужил корпус сообщений в блогах и пользовательских отзывов «ruTenTen» (15,8 млрд. словоупотреблений), корпус русскоязычных СМИ и блогов (25,5 млрд. словоупотреблений).

Объектом исследования являются языковые и внеязыковые способы реализации оценочных суждений (мнений) в текстах интернет-источников.

Предметом исследования выступают способы формализации оценочных суждений (на русскоязычном материале).

Структура и содержание диссертации. Диссертация состоит из введения, трех глав, которые сопровождаются выводами, заключения, списка литературы и списка иллюстративного материала, использованного в диссертации, а также четырех приложений. Общий объем диссертации составляет 200 страниц.

Во введении описывается актуальность исследования, научная новизна, теоретическая и практическая значимость работы, экспериментальный языковой материал, четко сформулированы объект и предмет исследования, методология и методы исследования, цель работы, которая состоит в разработке «принципов создания лингвистического обеспечения системы автоматического извлечения мнений для анализа текстов на русском языке» (с. 6). Подробно описаны задачи, решение которых сделает возможным достижение поставленной цели. На защиту выносятся пять положений. Представлены данные об апробации работы.

Глава первая описывает исторические этапы и современное состояние разработки исследуемой тематической области. Оформление автоматического извлечения мнений как отдельного исследовательского направления обусловливается запросом из практики и интересом ученых к проблеме субъективности в целом. Исторический анализ развития исследований в этой области свидетельствует о возрастающем интересе к лингвистическому компоненту систем автоматического извлечения мнений и стремлении к сближению прикладных разработок и теоретических исследований, что, в свою очередь, ведет к расширению сферы применения подобных технологий.

Современный этап характеризуется следующими основными тенденциями: а) значительным увеличением числа ресурсов (размеченные текстовые корпусы, демонстрационные версии программ); б) ростом числа языков, для которых разработаны системы автоматического извлечения мнений; в) продолжением совершенствования методов машинного обучения; г) переосмыслением задачи автоматического извлечения мнений (с. 25).

Отечественные разработки в этой области отличаются использованием преимущественно методов на основе словарей оценочной лексики.

Наряду с преимуществами разрабатываемых систем автор отмечает существенные сложности, связанные с отсутствием унифицированной терминологии для сферы автоматического извлечения мнений; наличием множества уникальных программных продуктов, которые привязаны к определенному типу текстового контента и с трудом адаптируются к новым предметным областям.

На основе компонентного анализа терминов, описывающих исследуемую область (*анализ субъективности – subjectivity analysis*, *автоматическое извлечение мнений – opinion mining*, *анализ чувств – sentiment analysis*, *анализ / определение тональности*, *с(а/е)нтимент-анализ*, *извлечение оценочных слов*, *анализ мнений*), автор приходит к заключению об использовании в качестве наиболее общего термина «автоматическое извлечение мнений».

Глава вторая раскрывает структуру лингвистического обеспечения в системах автоматического извлечения мнений. Следует положительно оценить тщательность и детализированность описания каждого этапа автоматического определения оценочного суждения, в котором особое внимание автор уделяет лингвистическому анализу на различных этапах работы системы. Под лингвистическим обеспечением понимается «совокупность средств и правил для формализации естественного языка» (с. 39). В составе модуля автоматического извлечения мнений наряду с подмодулем определения объекта мнения, субъекта мнения и подмодуля оценки включен подмодуль определения свойств объекта. Автор акцентирует внимание на необходимости разграничения субъекта мнения от источника и канал связи с оговоркой, что «данные явления не характерны для блогов и иных социальных медиа и встречаются в основном в текстах СМИ» (с. 81).

В списке трудностей, с которыми сталкивается разработчик систем автоматического извлечения мнений применительно к русскоязычному материалу, автор указывает проблему одушевленности собирательных существительных, проблему формального представления словоизменительных парадигм существительных-заимствований.

Центральный компонент системы автоматического извлечения мнений – словарь оценочной лексики – автор предлагает формировать на основе расширенной классификации оценочной лексики (см., например, работы M. Taboada, M. Klenner), дополнив ее такими лексическими классами, как: положительные усилители, отрицательные усилители, субъективно-нейтральные слова и словосочетания, субъект-позитивные глаголы, субъект-негативные глаголы, однореферентные слова (с. 107). При формировании словаря предлагается учитывать также частеречную информации, особенности фразеологических (фразеологизмы с вариативной частью и нерасчленимые фразеологизмы) и терминологических сочетаний. В частности, нерасчленимые фразеологизмы предлагается заносить в отдельный словарь.

В Главе третьей представлены результаты собственно экспериментального исследования на основе авторской методики. Применение разработанной автором двухступенчатой методики позволило сформировать первичный словарь оценочных прилагательных с частотой словоупотреблений в корпусе выше десяти, встречающихся на расстоянии 1 слева от слова в запросе. Как справедливо отмечается, «выбор имен прилагательных связан с тем, что именно они наиболее сильно модифицируют оценку слов в текстах Сети Интернет, где значительный процент предложений не имеет глаголов или предикативных слов» (с. 148). Тестирование разработанной методики проводилось с использованием технология «Sketch Engine» на корпусе «ruTenTen» (объем корпуса на русском языке составляет 15,8 млрд. словоупотреблений), для морфологической разметки корпуса применялся анализатор «TreeTagger».

Ценность представляет и разработанная автором база данных этнофобонимов (85 вхождений – лексических единиц, имеющих своим референтом страну / регион проживания и ее жителей), которые относятся к типу однореферентных оценочных слов. По мнению автора, «наиболее адекватным способом формального описания однореферентных слов является тезаурус» (с. 172), т.к. он позволяет описывать разнообразные лингвистические свойства (принадлежность к части речи, референциальные и прагматические свойства слов). В базе данных этнофобонимов единицы тезауруса сгруппированы по словообразовательному принципу, порядок расположения словообразовательных моделей зависит от их оценочной силы. Валидность применяемой в диссертационном исследовании технологии морфемного синтеза в целях автоматического пополнения словаря этнофобонимов была подтверждена на основе построения слообразовательных моделей для ксенофобонимов.

Следует отметить, что описание разработанных автором методов сопровождается уточнением специфики работы в ручном и автоматическом режимах, а также ссылками на недостатки применяемого подхода, что указывает на профессионализм автора в отношении полноты предоставления данных об интеллектуальном продукте.

Выбор материала для проведения экспериментального исследования дает основания для доказательного анализа в соответствии с поставленными в диссертации целями. Большой объем анализируемых текстов делает выводы исследования убедительными. Соображения и выводы автора сопровождаются многочисленными примерами.

Заключение обобщает результаты изложенной в диссертации работы и содержит основные выводы по обозначенной тематике.

Список литературы, цитируемой в диссертации, составляет 178 научных источников, в том числе 63 источника на иностранных языках. Обзорно-аналитическая часть работы охватывает труды основателей научного направления автоматического извлечения мнений и современные работы, отражающие актуальное состояние разработок и исследований.

Замечания. Отмечая высокий уровень эрудиции автора диссертации, строгий научный стиль изложения материала, четкость выводов и заключений, наличие значительного числа инкорпорированных в текст диссертации рекомендаций по повышению эффективности существующих систем автоматического извлечения мнений с применением лингвистического обеспечения, хотелось бы уточнить некоторые отдельные аспекты экспериментальной части исследования.

1. Желательно более акцентировано описать основания выделения дополнительных классов оценочной лексики (с. 106-107), предложенных автором, а также описать процедуру эмпирического исследования (если таковое имело место), в процессе которого автору удалось формализовать данные классы.

2. Было бы полезно описать исследование (с. 120-124), которое проводилось на материале лингвистического обеспечения системы «Аналитический курьер», позволившее автору утверждать, что «представление уменьшителей и переключателей оценки в виде правил, а не словарей, направлено на более компактное описание оценочной лексики за счет сокращения количества словарей, а также позволяет более оперативно исправлять возникающие ошибки при помощи дополнительных ограничений на оценочную сочетаемость» (с. 124).

3. В связи с тем, что автор среди прочих недостатков существующих методов, построенных на базе обращения к словарным ресурсам, указывает на отражение в последних устаревшей лексики (см. с. 145), необходимо уточнение временно́го интервала текстового контента, составившего сверхбольшой морфологически размеченный корпус текстов (свыше 1 млрд. словоупотреблений), который используется на первом этапе авторской двухступенчатой методики формирования первичного оценочного словаря (с. 147), а также указание на соотнесенность «возраста» полученного первичного оценочного словаря с корпусом веб-текстов (15,8 млрд. словоупотреблений) за 2011 год (с. 149).

4. Вопрос о фиксации силы оценки, затронутый в п. 2.4.1.5.2 (с. 105-106), не отражен при описании практической реализации разработанных автором принципов и методов автоматического извлечения мнений. При это представляется интересным дальнейшее развитие вывода автора о том, что «деривационные компоненты (*этно- и религиофобонимов – пояснение наше, Л.К.*) позволяют провести разграничение не только по частям речи, но и по силе оценки» (с. 168) в части диагностического потенциала словообразовательных моделей указанных явлений при автоматическом извлечении мнений.

Замечания, высказанные в настоящем отзыве, не влияют на высокую положительную оценку проделанной автором работы, результатом которой является настоящая диссертация, и носят характер рекомендаций, которые ориентированы на углубление исследуемой проблематики в будущем. Диссертация Куликова С.Ю. является самостоятельным исследованием,

отличается детальным анализом теоретического материала, обладает теоретической и практической значимостью в сфере прикладной и математической лингвистики.

Результаты диссертационного исследования могут найти широкое применение в практике разработки систем автоматического извлечения мнений. Материалы диссертации могут быть полезны при составлении курсов по прикладному языкознанию, формальным моделям в языкознании и спецкурса по новым информационным технологиям в лингвистике.

Работа прошла апробацию на заседаниях сектора прикладного языкознания Института языкознания РАН и различных научных конференциях. Основное содержание диссертации отражено в автореферате и 24 научных публикациях автора, в том числе в трех публикациях в журналах, рекомендованных ВАК Министерства образования и науки РФ.

Диссертация «Автоматическое извлечение мнений: лингвистический аспект», представленная на соискание ученой степени кандидата филологических наук, является научно-квалификационной работой, которая соответствует критериям, установленным пп. 9-14 «Положения о порядке присуждения ученых степеней» (утвержденного постановлением Правительства Российской Федерации от 24 сентября 2013г. № 842), а ее автор, Куликов Сергей Юрьевич, заслуживает присуждения ученой степени кандидата филологических наук по специальности 10.02.21 – Прикладная и математическая лингвистика.

Отзыв составлен и. о. старшего научного сотрудника отдела языкознания ФГБУН ИНИОН РАН кандидатом филологических наук Комаловой Лилией Ряшитовной.

Отзыв обсужден и утвержден на заседании отдела языкознания ФГБУН ИНИОН РАН (Протокол № 2 от 30.08.2016г.).

Заведующая
отделом языкознания
Центра гуманитарных
научно-информационных исследований
ФГБУН ИНИОН РАН,
д-р филол. наук, профессор Яковлева Э.Б.

Руководитель Центра гуманитарных
научно-информационных исследований
ФГБУН ИНИОН РАН,
д-р филос. наук, профессор Скворцов Л.В.

Подпись Яковлевой Э.Б. УДОСТОВЕРЯЮ
Зав.отделом кадров