ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА
на диссертацию Швеца Александра Валерьевича
«Взаимодействие информационных и лингвистических методов
в задачах анализа качества научных текстов»,
представленной на соискание ученой степени кандидата технических наук
по специальности 05.13.17 – «Теоретические основы информатики»

В настоящее время в связи с увеличением количества публикаций все больший интерес вызывает проблема оценки качества авторских текстов. Важным показателем качества является наличие в тексте различных нарушений устоявшихся научных норм. Это могут быть грамматические ошибки, нарушения требований к лексике, отсутствие четкой структуры, лексическая избыточность и т. п. Как правило, нарушения приводят к тому, что текст становится более сложным для восприятия и может вводить читателя в заблуждение. В результате содержание научного текста не позволяет оценить проведенное исследование, применить описанные в нем методы, воспроизвести экспериментальные исследования. К сожалению, подобные нарушения встречаются и в рецензируемых источниках, поэтому нельзя строить оценку качества научного текста исключительно на репутации издателя. Исследование А.В. Швеца посвящено решению задачи разработки методов автоматического выявления нарушений и определения качества текста, таким образом тема диссертации является крайне **актуальной**.

Диссертация состоит из введения, трех глав, заключения, списка использованных источников и приложения. Полный объем диссертации составляет 120 страниц; список литературы содержит 94 источника.

Во введении обоснована актуальность темы, определен предмет исследования, сформулированы цель и задачи исследования, научная новизна, а также охарактеризована теоретическая и практическая значимость полученных результатов.

В первой главе приведен анализ наиболее частых нарушений, допускаемых в научных текстах. Сделан вывод о том, что их обнаружение может быть выполнено автоматически с использованием лингвистических свойств текста. Затем приведен обзор методов, позволяющих на основе автоматического анализа текста определить его качественные характеристики. Отмечены недостатки и ограничения существующих методов. Сформулированы основные задачи исследования, предложен общий алгоритм выявления признаков, характеризующих качество текстов научной сферы. Вторая глава посвящена описанию и исследованию разработанных в рамках диссертации методов обнаружения различных нарушений. Представлены соответствующие методам алгоритмы, определены входные данные, необходимые для выполнения алгоритмов, предложены методы формирования этих данных. Проведены эксперименты, подтверждающие эффективность разработанных методов. Для извлекаемых признаков, характеризующих качество научного текста, установлены интервальные значения, показывающие степень отступления текста от устоявшихся требований к научным публикациям. В третьей главе для исследования применимости предложенных методов решается задача автоматического обнаружения псевдонаучных текстов. Предложен метод классификации коротких фрагментов текста для определения дополнительного признака, значение которого обозначает относительное количество фрагментов в тексте, классифицированных как псевдонаучные. Далее приведено сформированное пространство признаков, которое затем использовано при построении множества правил, устанавливающих принадлежность текста множеству псевдонаучных текстов. Правила формируются с помощью индуктивного метода порождения гипотез. В конце главы приведены результаты экспериментов, свидетельствующие об эффективности полученных правил. Заключение содержит основные результаты диссертационной работы. В

приложении приведено описание разработанных программных средств для определения качества текстов научной сферы.

Научную новизну диссертации составляют разработанные автором метод автоматического формирования общенаучного словаря словосочетаний, метод автоматического определения структуры научной публикации, метод обнаружения нарушений правил согласования, нарушений синтаксической и семантической связности, лексической избыточности, нарушений последовательности изложения, метод автоматического выявления псевдонаучных фрагментов текстов научной сферы, а также сформированное множество признаков, характеризующих качество текстов научной сферы, и построенное множество правил для обнаружения псевдонаучных текстов.

Практическая значимость диссертации определяется широкой областью возможного применения результатов работы: они могут быть использованы в системах поиска и анализа научной информации, в системах поддержки принятия решений при отборе заявок, проектов, приеме статей для публикации в научных журналах и в трудах конференций, а также для решения иных задач интеллектуального анализа текстов. Разработанные автором программные средства, которые реализуют основные методы и алгоритмы, предложенные в диссертационной работе, внедрены в различные информационно-аналитические системы четырех ведущих российских организаций в области электронно-библиотечных систем.

Диссертационная работа выполнена автором самостоятельно на высоком научном уровне. **Достоверность результатов** подтверждена вычислительными экспериментальными исследованиями.

Научные положения, выводы и рекомендации, полученные в диссертации, достаточно полно **обоснованы** и **аргументированы**. Автореферат диссертации отражает основные положения, выводы и результаты диссертационных исследований.

Результаты работы обоснованы и достаточно полно отражены в публикациях автора по теме исследования. По теме диссертации А.В. Швецом опубликовано 9 работ: из них 4 – в рецензируемых изданиях из Перечня ВАК РФ и приравненных к ним, 3 – в трудах российских и международных конференций, 2 – зарегистрированные программы для ЭВМ.

Отмечая высокое качество проработки решаемых задач, в то же время необходимо обратить внимание автора на **ряд критических замечаний** по существу отдельных положений диссертационной работы:

1. Во введении отсутствуют определения некоторых понятий, которые раскрываются лишь по мере появления в основной части работы (например, «псевдонаучный текст», «структура текста»).
2. В работе отсутствует пояснение, почему значения извлекаемых признаков, характеризующих качество научных текстов, являются дискретными. Лишь в третьей главе показано, что они используются для получения интерпретируемых экспертом правил. Стоило уточнить причину дискретности раньше, в первой или второй главе, где устанавливаются значения признаков.
3. Наряду с термином «снижение размерности признакового пространства» равнозначно используется термин «сокращение пространства признаков», который не является общепринятым. Следовало избежать употребления одного из терминов или, по крайней мере, пояснить, что они несут одинаковое значение.
4. В работе отсутствует анализ применимости предлагаемых инструментов в зависимости от области науки, которой принадлежат рассматриваемые публикации. В частности, неясно, подходят ли предложенные методы для гуманитарных наук.

5. В проведенном исследовании используется допущение, согласно которому научные публикации однородны по своему составу (научные состоят только из научных фрагментов, а псевдонаучные из псевдонаучных). На практике такие условия вряд ли выполняются для всех научных публикаций.

Отмеченные недостатки не снижают общую высокую оценку диссертационной работы.

В целом диссертационная работа является законченным научным исследованием, удовлетворяет всем требованиям Положения, предъявляемым к диссертациям на соискание ученой степени кандидата технических наук, а ее автор А.В. Швец заслуживает присуждения искомой ученой степени кандидата технических наук по специальности 05.13.17 – Теоретические основы информатики.

Младший научный сотрудник
Института вычислительных технологий
Сибирского отделения Российской академии наук,
кандидат технических наук _____ А.А. Князева
«10» сентября 2015 г.

Князева Анна Анатольевна, Институт вычислительных технологий СО РАН, 645055, г. Томск, Академический просп., 10/4, +7 (382) 249-17-74, aknjazeva@ict.nsc.ru

Подпись м.н.с. Князевой Анны Анатольевны удостоверяю
ученый секретарь ИВТ СО РАН
к.ф.-м.н. /Есипов Д.В./