# ОТЗЫВ

официального оппонента о диссертации
Мироновой Дины Марковны
«Автоматизированная классификация древних рукописей
(на материале 525 списков славянского Евангелия от Матфея XI-XVI веков)»
представленный на соискание степени кандидата филологических наук
по специальности 10.02.21 – Прикладная и математическая лингвистика

Диссертация Дины Марковны Мироновой посвящена кластерному анализу 525 списков славянского Евангелия от Матфея XI-XVI веков.
Актуальность работы состоит в том, что в том, что в рамках бесспорно значимой темы исследования библейской традиции на языках оригиналах и в переводах применяются автоматизированная методика классификации, дающая новые инструменты для сравнения рукописных текстов. Научные положения и выводы диссертации обоснованы как на теоретическом, так и на экспериментальном уровне, так же как достоверность и новизна полученных результатов. Автору удалось применить разные методы компьютерной классификации к обширному рукописному материалу, разработать и программно реализовать процедуру выделения текстологических примет групп рукописей, подробно описать алгоритм выделения узлов разночтения.

Представленное исследование является чрезвычайно важным не только для области текстологии древних рукописей, но и для гуманитарной науки в целом. С одной стороны, оставаясь абсолютно в рамках исходного понятийного поля, с другой стороны, предлагая современные компьютерные методы анализа близости рукописей, автор показывает нам образец работы, в котором компьютерные технологии используются для решения гуманитарных исследовательских задач. Более того, подходы к решению таких задач имеют весьма длительную историю и компьютерные технологии становятся очередным ее этапом, открывающим, однако, много

новых возможностей. Таким образом, результат диссертационного исследования органически встраивается в ряд текстологических исследований древних рукописей, но в тоже время вклад исследования в научный процесс состоит в технологическом прорыве, который, вне всякого сомнения, повлияет на будущие исследовательские практики.

Хорошая работа начинается с правильно поставленной цели. Целью исследования стала формализация критериев отбора текстовых фрагментов для текстологического анализа и последовательное практическое исследование влияния различных типов таких фрагментов на результат объединения рукописей в осмысленные группы посредством автоматического кластерного анализа. Для того чтобы реализовать эту цель автор предпринимает несколько промежуточных важных шагов:

- сравниваются разные подходы к кластерному анализу,
- готовится список узлов разночтений для 525 списков Евангелия от Матфея
- исследуется то, как разные классы разночтений влияют на конечный результат.

Представляется, что наиболее значимым и ярким результатом работы является создание комплексной системы описания узла разночтений. Система, созданная автором, базируется на текстологических принципах изучения рукописной традиции и формальных методах классификации рукописей. Еще одним важным достижением автора является разработка методики выделения текстологических примет для классификации рукописных источников текста.

Диссертация состоит из введения, трех глав, заключения, списка принятых сокращений, списка литературы, списка иллюстративного материала и четырех приложений.

Введение посвящено обоснованию темы, формулировки задач и целей исследования.

Первую главу автор начинает с исторического обзора подходов к формализации анализа отношений между текстами в разных рукописях. Особое внимание уделяется кластерному анализу, который и будет в дальнейшем применяться автором. Далее, во второй главе, автор сравнивает методы Ваттеля и Алексеева, и не ограничиваясь теоретическими рассуждениями, тестирует оба метода на хорошо изученном материале. Я бы хотела отдельно отметить выбранный автором подход, при котором экспериментально обосновывается выбор метода исследования. Мне кажется, что это очень важный шаг в направлении объективизации результатов исследований гуманитарных объектов. Но замечательно и то, что выбор делается не количественно (оба метода показывают примерно одинаковые результаты), но качественно – метод Алексеева позволяет визуально оценить близость каждой рукописи периферии к той или иной редакции путем сравнения процента сходства между рукописью и рукописями ядра различных редакций. Иными словами, метод Ваттеля предстает своего рода «черным ящиком», затрудняющим интерпретации того, какие факторы повлияли на результаты кластеризации, в то время как метод Алексеева позволяет провести качественный анализ того, какие различия между рукописями оказываются значимыми для кластеризации. Этому анализу и интерпретации наблюдаемых лингвистических расхождений посвящена третья глава диссертации. Автор опирается на понятие «узла разночтения» и далее проводит очень детальный лингвистический анализ, выделяя основные классы узлов и подробно останавливаясь на их заполнении. Исследование проводится на материале отрывка из Евангелия от Матфея, рассматриваются более 500 рукописей, оценивается влияние разных типов узлов на результаты кластеризации.

В целом хочется подчеркнуть, что перед нами очень глубокое, очень методологически продуманное исследование, выполненное с привлечением огромного материала. Значимость этого исследования, кроме достигнутых конкретных результатов, состоит в выработке методологии, чрезвычайно актуальной для сферы компьютерных методов в гуманитарных науках в целом. Очень важно, что автор находит идеальных баланс между

математикой – сама процедура кластеризации изложена максимально четко и доступно, - и лингвистическим исследованием. Автор убедительно показывает насколько ценнее и интереснее оказываются результаты применения квантитативного аппарата, когда они дополняются качественным лингвистическим разбором. Именно такого взаимопроникновения, интеграции гуманитарного и точного знания очень часто не хватает многим современным работам, связанным с цифровыми исследованиями текста. Также хочется отметить, что работа написана очень ясным, понятным языком, все рассуждения снабжены исчерпывающими примерами, работа замечательно структурирована и легко читается.

Небольшие замечания к работе носят дискуссионно-терминологический характер и никак не могут повлиять на общую оценку ее научной значимости и высокого уровня академического исполнения. Тем не менее, представляется, что заданный в работе вектор интеграции точных методов и лингвистического анализа мог бы быть усилен большей терминологической выверенностью терминов, связанных с анализом данных. Речь идет о некоторой культуре описания данных, которая, на мой взгляд, важна для того, чтобы результаты исследования были понятны не только специалистам по истории русского языка, но и специалистам в области data science. Приведу два примера. Так, в науках о данных принято разделять термины *классификация* и *кластеризация*. Классификацией называют автоматическое разделение массива текста по заданным классам, представленным, например, образцами текстов, или ключевыми словами, тогда как кластеризация – это деление корпуса текстов на определенное количество кластеров наиболее близких текстов. Безусловно, результаты кластеризации необходимо осмыслить, и возможность интерпретировать кластер с точки зрения нашего понимания о классах текстов говорит о том, что кластеризация проведена корректно. Однако при описании собственно процедуры анализа эти два подхода принято различать, что не было сделано в работе, ср. на стр. 53

«Классификация по всем узлам позволила выделить 13 кластеров, из которых в одном объединилось 308 рукописей, в одном – 48 рукописей, в двух – по 4 рукописи, в одном – 3 рукописи, а во всех остальных – по 2 рукописи».

Также представляется, что можно было эксплицитнее описать то, каким образом были получены данные для кластеризации, в частности узлы расхождений. Опять же в науках о данных чрезвычайно важной является информация о том, насколько мы можем доверять исходной разметке. Очень часто ошибки, допущенные при автоматической разметке больших корпусов текстов, существенно влияют на дальнейшие результаты анализа. В данном случае, вся разметка производилась в ручную (огромный труд!), то есть возможность ошибки была сведена к минимуму. Но это обстоятельство должно было быть прописано отдельно.

Тем не менее хочется отдельно подчеркнуть, что приведенные замечания ( а фактически это не замечания, а пожелания) никоим образом не влияют на содержательную значимость и академическую состоятельность рассматриваемой работы. Диссертация полностью соответствует требованиям, предъявляемым к кандидатским диссертациям и изложенным в пунктах 9-14 «Положения о порядке присуждения ученых степеней», утвержденного Постановлением Правительства РФ от 24.09.2013 № 842.

Дина Марковна Миронова безусловно заслуживает присуждения искомой ученой степени кандидата филологических наук по специальности 10.02.21 – прикладная и математическая лингвистика.

Кандидат филологических наук
Доцент школы лингвистики
факультета гуманитарных наук
Федерального государственного автономного образовательного
учреждения высшего профессионального образования
«Национальный исследовательский университет «Высшая школа экономики»
Адрес: 105066 г. Москва, ул. Старая Басманная д. 21/4, к. 519
Телефон: +7 (495) 772-9590, добавочный 22504
Электронный адрес: abonch@hse.ru
https://www.hse.ru/org/persons/32878143

Подпись заверяю
СПЕЦИАЛИСТ ПО КАДРОВОМУ ДЕЛОПРОИЗВОДСТВУ
Т.В.БАТОВА
30.03.18г.

Бонч-Осмоловская Анастасия Александровна
30 марта 2018 года.