

ЗАДАЧИ АВТОРСКОЙ АТРИБУЦИИ ТЕКСТА В АСПЕКТЕ ГЕНДЕРНОЙ ПРИНАДЛЕЖНОСТИ (к проблеме междисциплинарного взаимодействия лингвистики и информатики)

Обсуждаются вопросы междисциплинарного взаимодействия лингвистики и информатики в связи с решением задач авторской атрибуции текста в аспекте гендерной принадлежности. Выявляются проблемные зоны теоретических и прикладных аспектов такого взаимодействия (соотношение качественных и количественных методов, ограничения материала исследования и параметров анализа текстов в гендерной лингвистике и автороведческой экспертизе) и определяются направления их решения в рамках совместного междисциплинарного проекта лингвистов и информатиков.

Ключевые слова: автороведческая экспертиза; гендерная лингвистика; междисциплинарные методы исследования; информатика; лингвистика.

Постановка задачи взаимодействия лингвистики и информатики

Тенденция к междисциплинарной интеграции как одна из наиболее значительных в развитии современной науки стимулируется необходимостью решения проблем, возникающих в пределах проблемных зон развития теории и методологии конкретных наук и в связи с разрешением прикладных задач, имеющих междисциплинарный характер.

В статье обсуждаются проблемные зоны, актуализирующиеся в настоящее время в гендерной лингвистике и в автороведческих экспертных исследованиях (в исследовательском сегменте определения пола автора текста), намечаются пути разрешения ряда вопросов в проекте, совмещающем методы лингвистического анализа с использованием приемов информатизационной обработки текстовых массивов.

Стилеметрия как междисциплинарное направление прикладных исследований, решающее задачи определения авторства «спорных» текстов, сформировала методическую основу своих исследований на совмещении качественных и количественных методов [1]. Мощный стимул к ее развитию был дан на этапе привлечения методов информатики, компьютерной обработки текстов, усиливших результативность «ручных» количественных способов анализа. Уже на первом этапе решения задач определения авторства спорных художественных текстов выявились проблемные зоны:

1) определение *типа языковых единиц*, статистический анализ которых достаточен для идентификации текста;

2) объем текстов, на основе анализа которых выявляется состав релевантных признаков (так называемые «обучающие тексты»);

3) объем исследуемых текстов, способных обнаружить значимый достоверный результат;

4) соотнесенность качественных и количественных методов анализа и др.

При разрешении первой проблемы исследователи двигались в двух противоположных направлениях, стремясь либо к минимизации, либо к максимальному расширению списка релевантных для определения автора текста языковых единиц. В первом случае в подсчетах, как правило, используются асемантические элементы или опосредованно семантические элементы.

Ярким примером проявления такой тенденции являются труды А.А. Маркова, разработавшего метод идентификации автора текста на основе учета сугубо формальных признаков. Метод апробировался на примере статистического анализа «распределения доли гласных и согласных среди первых 20 000 букв “Евгения Онегина”» (цит. по: [2]). В аналитических работах второго направления в состав элементов, проверяемых на релевантность при определении авторства, включаются как асемантические, так и семантические единицы разных языковых уровней: длины слов, длины предложений, служебные слова, частоты встречаемости отдельных букв, буквосочетаний, знаков препинания, закономерности в повторении этих единиц, сложность и частоты появления отдельных грамматических конструкций и т.п. [3, 4 и др.].

Решение первой проблемы непосредственно коррелирует с другими: так, введение в стилеметрический анализ семантических элементов (морфологических, синтаксических, лексических) требует предварительного лингвистического анализа (приводимого традиционно с привлечением так называемых качественных методов), как следствие – уменьшение объема привлекаемых для анализа текстов. «При таком подходе, – отмечает Д.В. Хмелев, – большой вычислительный эксперимент с целью проверки методики на большом числе авторов практически неосуществим» [2. С. 115].

В настоящее время методы, выработанные в стилеметрии, активно применяются в системе юридической экспертизы текста, и, таким образом, формируются новые условия для развития этого направления, что, во-первых, стимулируется наличием активного *социального запроса* на определение автора текстов в криминалистической экспертизе и, во-вторых, обеспечивается накоплением знаний в теоретической лингвистике, в различных направлениях антропоцентрической, функционалистской парадигмы современного *языкознания*, более широким использованием в лингвистических исследованиях методологического аппарата активно развивающейся *информатики*.

Данное направление прикладных исследований существенно расширило круг вопросов, решаемых в стилеметрии, по-новому высветило проблемные зоны разработанной методологии, а также актуализировало ряд новых.

1. Остается актуальным вопрос о типе единиц, релевантных при автороведческой экспертизе. Этот вопрос по-новому разворачивается в связи с разнообразием конкретных задач, решаемых в криминалистических экспертизах текстов. Для разрешения определенных экспертных задач может оказаться достаточным определение одного из параметров языковой личности автора текста: гендерного, возрастного, образовательного, культурного и т.д.

2. Как правило, проблемы авторской атрибуции текста решались и решаются в настоящее время преимущественно на материале художественных произведений. Такое положение дел, во-первых, имеет конкретное историческое обоснование – проблема определения авторства впервые была поставлена применительно к так называемым спорным художественным текстам. Во-вторых, обращение к материалу художественных текстов мотивируется и тем, что в настоящее время это направление переживает этап становления, в связи с чем доступность материала исследования, его разноаспектность, значительные объемы текстового материала, принадлежащего одному автору, создают безусловные преимущества художественных текстов для отработки методик, проверки релевантности признаков идентификации по сравнению, например, с устной речью, объемы текстовых массивов которой, принадлежащие одному автору и доступные для исследования, весьма невелики в настоящее время.

Однако при этом остается открытым, требующим неперменного решения вопрос о том, как отработанные на материале объемных художественных текстов методики будут проявляться в условиях ограниченного текстового материала. Вопрос о достаточной для проявления статистических закономерностей длине текста в автороведческой экспертизе является принципиальным.

Существенным является также вопрос о том, как данные методики будут работать при их переносе на текстовый материал других стилей и дискурсов, в том числе дискурсов, обусловленных в своем появлении новейшими информационными технологиями: Интернет-, смс-коммуникация и т.п.

Как представляется, все эти проблемные вопросы автороведческой экспертизы могут и должны решаться на принципах междисциплинарного взаимодействия лингвистики и информатики с привлечением данных, выработанных в новейших направлениях современной лингвистики, и новейших методов количественного анализа, разрабатываемых в информатике: статистических методов, методов искусственного интеллекта и т.д.

Так, разрешение проблемы выбора языковых единиц, по отношению к которым применяются методы количественного анализа, во-первых, должно опираться не только на классические знания, выработанные в традиционной лингвистике, – знания о единицах уровня членения языка, представления о ядерных и периферийных единицах, их уровневой принадлежности, семантической и формальной специфике и т.д., но и на данные новейших направлений языкознания: лингвоперсонологии (теории языковой личности), структуры языкового сознания, данные психолингвистики и нейролингвистики, когнитивной лингвистики, гендерной и онтолингвистики и т.д.

В названных отраслях научного знания с применением соответствующих либо внутринаучных, либо междисциплинарных методов выработаны знания о проявлении в текстовых структурах разных аспектов когнитивной деятельности человека в аспекте и индивидуального варьирования, и типологических признаков.

Методология гендерных исследований

Проблема определения *пола автора* текста в автороведческой экспертизе в настоящее время не может решаться вне опоры на знания о системе языковых маркеров гендера – языковых единицах, моделях, конструкциях, противопоставляющих речь в гендерном аспекте.

Уже в ранних исследованиях особенностей речевого поведения мужчин и женщин на материалах различных лингвокультур были выявлены дискурсивные маркеры доминирования мужчин в речевых практиках, обозначены лексические показатели гендерной специфики, проявляющиеся в аспекте количественного преобладания использования одних и тех же языковых единиц (цветообозначений, экспрессивной, эмоциональной лексики), стратегий и тактик ведения коммуникации и т.д. Были определены различия в характере общения однородных и неоднородных в гендерном аспекте групп коммуникантов: мужчины с мужчинами и женщины с женщинами осуществляют стратегии развертывания текста иначе, нежели в ситуациях межполовой коммуникации (см. обзоры результатов гендерных исследований в [5, 6] и др.).

В настоящее время на материале русского языка охарактеризованы коммуникативные модели, общий эмоциональный и волитивный фон, специфика выражения дескриптивных и эмоциональных смыслов и другие признаки типично мужской и типично женской речи ([6–9] и др.).

Вместе с тем обнаружилась и относительность выводов, представленных в такого рода исследованиях, так как в них либо не учитывалось, либо учитывалось в слабой степени влияние других личностных качеств на реализацию гендерных характеристик: «Ученые все больше и больше настаивают на том, что расово-этническую, социально-классовую позицию и гендерный статус нельзя рассматривать как аналитически различные, скорее, к ним необходимо подходить как к трехчастной стратификационной системе, в которой каждый компонент переплетен с другими и каждая система зависит от двух других» [10. С. 190].

В связи с этим встает вопрос о разработке процедур текстологических исследований, позволяющих дифференцировать влияние различных аспектов языковой личности на характер ее речевой реализации.

Анализ широкого спектра гендерных исследований позволяет сделать также вывод о том, что обычно исследователи анализируют один тип дискурсивной реализации коммуникантов: изучается либо разговорная, либо письменная деловая, художественная и т.д. речь, но, как правило, не ставится вопрос об устойчивости / вариативности признаков гендерного противопоставления в разных типах речевых ситуа-

ций. При этом данная проблема представляется нам весьма существенной, и ее решение также требует коррекции методик, конкретных процедур гендерных исследований.

Таким образом, можно сделать вывод о том, что накопление научных данных активно развивающегося направления гендерных исследований в современной русистике актуализировало проблему взаимной верификации результатов исследований, ведущихся на разном материале, решающих различные частные проблемы и применяющих разные методы.

Одно из направлений в продвижении по этому пути – решение вопроса о соотношении качественных и количественных методов анализа.

Размышляя по этому поводу, Е.И. Горошко, одна из лидеров русскоязычных гендерных лингвистических исследований, отмечает явное преобладание в них качественной методологии, позволяющей преодолевать «слабые места позитивизма в изучении социума», и в то же время наметившуюся устойчивую тенденцию применения количественных приемов анализа и их соотношения с качественными методами, так как, говоря о противопоставленности мужской и женской моделей коммуникации, исследователи прежде всего ссылаются на количественное преобладание определенных языковых единиц. Автор отмечает также, что наиболее явно тенденция активного применения количественных методов проявляется в исследованиях прикладной направленности [11. С. 35–36].

Вопрос о соотношенности результатов применения качественных и количественных методов в прикладных лингвистических исследованиях до сих пор остается одним из нерешенных, однако продвижение в этом направлении представляется важным не только для прикладной, но и для теоретической лингвистики.

Проблемы автороведческого гендерного анализа текста

Таким образом, как свидетельствует обзор путей развития и проблемных зон гендерных и автороведческих экспертных работ, в данных направлениях междисциплинарных исследований, развивающихся относительно самостоятельно, обнаруживается пересечение, во-первых, предметных сфер, во-вторых, применяемых методов. Вместе с тем наблюдается и определенная общность проблемных зон, требующих своего разрешения, к числу которых относим в первую очередь характер взаимного вовлечения результатов исследований в смежных областях знания, гендерологии и автороведческой экспертизы, и решение вопроса о соотношении и взаимной верификации применения качественных и количественных методов анализа, как следствие – характера междисциплинарной интеграции.

В настоящее время лишь отдельные аспекты гендерных исследований вовлечены в практику автороведческих экспертиз, когда гендерологи работают в тесной связи с экспертами-практиками. Такой весьма положительный пример демонстрирует харьковская школа гендерной лингвистики, вовлекающая результаты исследований в решение проблем автороведческой экспертизы текста. В серии статей и диссертационном

исследовании Е.С. Горошко был проведен в сотрудничестве со специалистами в области криминалистической экспертизы текста анализ русской письменной речи, относящейся к бытовому стилю письма, и составлены основной и дополнительный списки диагностических характеристик мужской и женской письменной речи, на основе которых возможно создание методики установления пола автора письменного текста. Как отмечает автор, проведенное исследование показало, что «все различия между мужской и женской письменной речью носят не инвентарный, а вероятностный характер, проявляясь в определенных особенностях употребления речевых средств как на качественном, так и на количественном уровне анализа этих различий» [12].

Работы подобной направленности [13–14] ведутся и коллективом авторов Томского университета систем управления и радиоэлектроники. В работах А.С. Романова и Р.В. Мещерякова проблемы определения автора текста короткого электронного сообщения и пола автора решались с использованием методов машинного обучения, таких как искусственные нейронные сети и машины опорных векторов.

В качестве параметров, отражающих авторский стиль или пол, использовались признаки текста уровня символов и слов: частоты встречаемости в тексте униграмм, биграмм и триграмм символов, знаков пунктуации, служебных слов, а также частотный словарь, распределение слов по частям речи и др. [15, 16].

Стоит отметить, что данные признаки не могут контролироваться человеком сознательно в отличие от синтаксических и семантических характеристик текста, поэтому их использование в автороведческой экспертизе является предпочтительным. Вместе с тем встает вопрос о том, как данные параметры соотносятся с семантически релевантными признаками.

Однако наблюдения и выводы, сделанные в работах по лингвистической гендерологии, либо не проверяются, либо проверяются в минимальной степени на способность выступать в качестве признака, который может работать как диагностирующий при гендерной идентификации автора текста. Об этом пишет, в частности, Л.В. Кузниченко, отмечая отрицательные следствия подобного положения дел для формирования методик криминалистической экспертизы: «Методики комплексного решения вопросов установления пола, возраста и профессиональной принадлежности лица, составившего текст, являются неразработанными. Поэтому не всегда экспертам удается решить поставленные перед ними вопросы» [17].

Автор отмечает необходимость привлечения «криминалистически интерпретированных данных и методов широкого круга научных дисциплин: лингвистики, психологии, психолингвистики, социолингвистики, коммуникативной лингвистики, социологии и т.д.» [Там же].

Перечисляя ряд аспектов автороведческой экспертизы, нуждающихся в дальнейшей серьезной разработке, Л.В. Кузниченко в качестве важнейших выделяет определение диагностических признаков языковых и интеллектуальных навыков и решение проблемы количественного анализа текстов с помощью

современных компьютерных технологий: «Важно отметить, что и разработка каждой конкретной методики установления определенного классификационного параметра личности является сложной самостоятельной проблемой» [Там же]. Автор также пишет о негативном влиянии подобной неразработанности и на формирование теоретических основ классификационной экспертизы.

Нам бы хотелось подчеркнуть существенное значение вовлечения в парадигму собственно гендерных исследований выводов, полученных в практике автороведческой экспертизы, прежде всего о нерефлексируемых собственно формальных признаках речи, но в то же время являющихся глубинной основой противопоставления мужского и женского речевого поведения. Вместе с тем нуждаются в количественной проверке выделенные качественными методами признаки «типично мужской» и «типично женской» речи.

В целом последовательное применение методов количественного анализа с привлечением новейших приемов информатики и вовлечение полученных с их помощью новых гендерно-релевантных признаков речевого поведения в научный лингвистический контекст создаст основу дальнейшего развития гендерной теории.

В свою очередь, информатики решают задачи проверки эффективности применяемых методов в условиях смены параметров измерения текстовых структур: типов единиц, длины и дискурсивной принадлежности текста и других признаков. Так, в исследованиях А.С. Романова и Р.В. Мещерякова при определении авторства удалось достичь точности 70%, при определении пола – 74%, в настоящее время применение в методике усовершенствованных метаклассификаторов позволяет достичь точности более 80% [18, 19].

Заключение

Решение комплекса обозначенных в статье взаимосвязанных теоретических и прагматических задач предполагается в междисциплинарном проекте, выполняющемся в настоящее время объединенным коллективом авторов кафедры комплексной информации

и безопасности электронно-вычислительных систем Томского государственного университета систем управления и радиоэлектроники и коллективом кафедры общего, славяно-русского языкознания и классической филологии Томского государственного университета.

Сформирована совместная программа исследований, направленных на решение проблем разноаспектной текстологической экспертизы в сочетании качественных и количественных методов анализа, на выявление влияния на результативность применяемых методов авторской атрибуции параметров:

- 1) сфер коммуникации;
- 2) длины коммуникативных фрагментов;
- 3) соотношения выявляемых асемантических и семантических единиц текстовых структур.

Решение проблем автороведческой экспертизы, намеченных в проекте, предполагает формирование особой эмпирической основы исследований.

Объединенный коллектив авторов в настоящее время ставит перед собой следующие задачи:

1. Создание корпуса текстов различной дискурсивной принадлежности и его разметка в соответствии с задачами автороведческой экспертизы. На первом этапе предполагаются формирование корпуса текстов коротких электронных сообщений и его лингвистическая разметка. Такой корпус можно будет использовать как для непосредственно исследований вопросов определения авторства и гендера, так и для создания автоматизированных средств обработки текста с применением машинных методов обучения.

2. Расширение перечня используемых для идентификации гендера признаков и проведение исследований по определению их информативности для решаемой задачи, что приведет к расширению функциональности программного комплекса «Авторовед» [18].

3. Исследование метаклассификаторов с целью повышения точности идентификации автора текста.

4. Обоснование полученных результатов как математическими методами, так и с точки зрения современных направлений языкознания. Обобщение ранее полученных моделей и методик идентификации автора текста [19] (на первом этапе – применительно к анализу коротких интернет-текстов).

ЛИТЕРАТУРА

1. Марков А.А. Об одном применении статистического метода // Известия Императорской Академии Наук. Сер. VI. 1916. Т. X, № 4.
2. Хмелев Д.В. Распознавание автора текста с использованием цепей А.А. Маркова // Вестник МГУ. Сер. 9: Филология. 2000. № 2. С. 115–126. URL: <http://www.philol.msu.ru/~lex/khmelev/published/vestnik/vestnik2000.html>
3. Шевелев О.Г. Разработка и исследование алгоритмов сравнения стилей текстовых произведений : автореф. дис. ... канд. техн. наук. Томск, 2006.
4. Романов А.С., Мещеряков Р.В. Идентификация автора текста с помощью аппарата опорных векторов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Междунар. конф. «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). М. : РГГУ, 2009. Вып. 8 (15). С. 432–437.
5. Кирилина А.В. Гендер: лингвистические аспекты. М., 1999.
6. Горошко Е.И., Кирилина А.В. Гендерные исследования в лингвистике сегодня // Гендерные исследования. Харьков, 1999. № 2. С. 234–241.
7. Земская Е.А., Китайгородская М.А., Розанова Н.Н. Особенности мужской и женской речи // Русский язык и его функционирование. М. : Наука, 1993. С. 90–136.
8. Анищенко А.В. О гендерных особенностях реализации эмоциональных реакций // Гендер: Язык, Культура, Коммуникация : материалы Третьей Междунар. конф. 27–28 ноября 2003 г. М., 2003. С. 18–19.
9. Борисова Е.Г. Употребление междометий в речи женщин и мужчин // Гендер: Язык, Культура, Коммуникация : материалы Третьей Междунар. конф. 27–28 ноября 2003 г. М., 2003. С. 18–19.
10. Лорбер Дж., Фаррелл С. Принципы гендерного конструирования // Хрестоматия феминистских текстов. Переводы / под ред. Е. Здравомысловой, А. Темкиной. СПб. : Дмитрий Буланин, 2000. С. 187–192.

11. *Горошко Е.И.* К вопросу о соотношении количественных и качественных методов анализа данных в лингвистической гендерологии // Гендер: Язык, Культура, Коммуникация : материалы Третьей Междунар. конф. 27–28 ноября 2003 г. М., 2003. С. 35–36.
12. *Горошко Е.И.* Судебно-автороведческая классификационная экспертиза: проблема установления пола автора документа // 36. науково-практичних матеріалів конференції «Теорія та практика судової експертизи і криміналістики». Харків : Право, 2003. Вип. 3. URL: ТЕКСТОЛОГИЯ.RU
13. *Вул С.М., Горошко Е.И.* Судебно-автороведческая классификационная диагностика: установление половой принадлежности автора документа // Современные достижения науки и техники в борьбе с преступностью : материалы науч.-практ. конф. Минск, 1992. С. 139–141.
14. *Вул С.М., Мартынюк А.П.* Теоретические предпосылки диагностирования половой принадлежности автора документа // Современное состояние и перспективы развития традиционных видов криминалистической экспертизы. М., 1987. С. 105–112.
15. *Романов А.С., Мецерыков Р.В.* Определение пола автора короткого электронного сообщения // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Междунар. конф. «Диалог» (Бекасово, 25–29 мая 2011 г.). М. : Изд-во РГГУ, 2011. Вып. 10 (17). С. 620–626.
16. *Романов А.С., Мецерыков Р.В.* Идентификация авторства коротких текстов методами машинного обучения // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Междунар. конф. «Диалог» (Бекасово, 26–30 мая 2010 г.). М. : Изд-во РГГУ, 2010. Вып. 9 (16). С. 407–413.
17. *Кузниченко Л.В.* Проблемы теории и практики судебно-автороведческой классификационной экспертизы. URL: <http://www.law.edu.ru/doc/document.asp?docID=1308749>. 2008 (Харьковский НИИСЭ).
18. *Романов А.С.* Структура программного комплекса для исследования подходов к идентификации авторства текстов // Доклады Томского государственного университета систем управления и радиоэлектроники. 2008. Ч. 1, № 2(18). С. 106–109.
19. *Романов А.С., Шелупанов А.А., Мецерыков Р.В.* Разработка и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста. Томск : В-Спектр, 2011.

Статья представлена научной редакцией «Филология» 9 марта 2013 г.