С. А. Хозяинов

НЕКОТОРЫЕ ПРОБЛЕМЫ И МЕТОДЫ КВАНТИТАТИВНО-СТРУКТУРНОГО ИЗУЧЕНИЯ АВТОРСКИХ СТИЛЕЙ

Работа представлена кафедрой математической лингвистики Санкт-Петербургского государственного университета. Научныйруководитель - доктор филологических наук, профессор М. А. Марусенко

В статье представлен анализ некоторых современных исследований в области атрибуции письменных текстов. Основное внимание уделено особенностям постановки и решения агрнбуционных задач (в аспекте квантитативно-структурного изучения авторских стилей).

ТЬе article presents the analysis of some modern researches in the area of written texts' authorship afiribvtkin. Basic attention is given to the features of statement and decision of attribution tasks (in the aspect of quantitative and structural research of authors' styles).

Современные математические методы аярибугми отталкиваются от количественной **концепции** стиля. Все многообразие проблем изучения авторских стилей (АС) кванлпативными методами можно свести к двум главным: 1) определения специфики **АС;** 2) точности атрибуции[1].

Первая проблема осложняется тем, что **АС** не однообразны и зависят от воздействий хронологического, жанрового и тематичеаои© факторов. Так, известно, что средний размер длины предложения примерно одинаков для текстов, написанных разными писателями в одном жанре, но существенно различен для произведений одного и того же писателя, выполненных в разных жанрах[2]. В связи с этим в рамках квантитативного изучения стилей выдвигается категорическое требование однородаости исследуемого материала по каждому из названных факторов.

Конкретное решение проблемы определения специфики АС зависит также оттого,

какие структурные уровни языка изучаемого текста оказываются предметом исследования. Наиболее распространенными являются морфология, лексика, синтаксис. В последнее время появились методики атрибуции, в которых предлагается обследовать явления текста, напрямую не соотносимые с традиционно выделяемыми уровнями языка. В связи с этим встает два вопроса: 1) каким образом эти явления текста отражают специфику АС; 2) насколько адекватны задачам атрибуции результаты, получаемые на основе изучения этих явлений языка текста.

Способы решения проблемы точности атрибуции определяются спецификой используемого метода. В любом случае в силу природы квантитативных методов мера этой точности поддается вычислению или оценке. Так, точность результатов разработанного И. П. Севбо метода определения автора путем сравнения графов зависимостей, характеризующих синтаксическую структуру авторской фразы, зависит от количества сравниваемых авторов. Для двух авторов вероятность правильного решения составляет 75%, для трех - 50%, для четырех - 35% и т. д.[3]

В зависимости от способов постановки и решения первой проблемы в современной науке различается несколько подходов к решению задач атрибуции. Рассмотрим некоторые из них на примере ряда работ. При этом обратим внимание на то, как в них формулируется атрибуционная задача и определяется объект анализа и в какой мере то и другое отвечает филологической и квантитативно-структурной трактовкам проблемы авторства (ПА) - трактовкам, подразумевающим изучение АС.

Под руководством профессора В. Н. Захарова коллектив исследователей ПетрГУ с 1993 г. занимается созданием баз данных дтя многоаспектного филологического анализа литературных текстов. Одной из целей этой работы является решение атрибуционных задач[4]. Долгое время в центре внимания исследователей была ПА статей, приписываемых Ф. М. Достоевскому. Последним результатом работы коллектива стало создание информационной системы «Статистические методы анализа литературных текстов», или ИС СМАЛТ\ В ней реализована возможность морфолого-синтаксического анализа текстов и их последующей атрибуции, предусмотрена возможность подключения шобых методик атрибуции - в виде динамических библиотек.

Процедуры обработки лингвистических данных и самой атрибуции в системе четко формализованы. Не так обстоят дела с постановкой задачи. Вместо понятия АС (имеющего много определений) исследователи используют понятие «стилистический инвариант», предложенное Гейром Хетсо[6]. Дифференциальные признаки этого понятия не учитывают важных аспектов изучения АС, описанных в классической филологии и ее прикладных дисциплинах[7]. Постановка атрибуционной задачи основана главным образом на предположении о том, что стилистический инвариант представлен распределением частей речи на первых трех и последних трех позициях предложений". (Хетсо исследовал дистрибуцию частей речи первых двух и последней позиций предложения в качестве детализации сравнительной частотности частей речи в тексте)[9]. Выбор исследователей не прокомментирован в аспекте задач изучения АС. Неисследованным оказывается материал, равный объему остальных позиций предложений, в то время как он не менее информативен в отношении определения АС. Необследованным остается синтаксис текстов. Однако известно, что именно на синтаксическом уровне свобода авторского выбора менее всего связана факторами, влияющими на АС[10].

Ю. В. Сидоров констатирует невозможность решения ПА притшеьшаемой Ф. М. Достоевскому публицистики рассмотренными

им методами (корреляционных плеяд, иерархического кластерного анализа, оценки парной связи грамматических классов). Причину этого исследователь видит в том, что обследованные тексты слишком малы"[11]. В отношении текстов малого объема эффективнее методика атрибуции М. А. Марусенко, основанная на применении методов теории распознавания образов. В недавнее время с ее помощью была предпринята более успешная попытка решения названной ПА[12]. Одним из достоинств этой методики является се тесная связь с филологическим аспектом изучения ПА[13]. Постановка атрибуционной задачи в работах исследователей Петр ГУ, напротив, во многом лишена этого взаимодействия. В то же время ими были учтены жанровый и хронологический критерии определения объекта исследования, поэтому полученные результаты соотносимы с поставленной задачей.

Нельзя сказать этого о результатах исследования Д. В. Хмелева[14]. Как полагает исследователь, общим недостатком современных методик является то, что они не были апробированы на материале произведений большого количества писателей. Опираясь на этот тезис, Д. Хмелев ставит задачу создания методики, позволяющей проверить ее на большом материале и, в сущности, *в краткие сроки.* Предлагаемый им метод предполагает автоматический математический анализ матриц, отражающих содержание изучае.мых текстов в виде так называемых цепей Маркова (ЦМ), с последующим определением ранга принадлежности текста тому или иному автору. В качестве реализации ЦМ рассматривается последовательность букв текста. В основном метод Д. Хмелева дал положительные результаты, но анализ условий эксперимента показывает, что в определенной мере успешность этих результатов объясняется несоблюдением названного выше требования однородности исстедуемого материала.

В исследовании даны два примера применения методики. В первом анализируются тексты трех авторов, во втором - 82. В каждом из примеров не соблюдены требования хронологического и жанрового соответствия. Исследователь сопоставляет тексты авторов XIX-XX вв. - например, произведения А. П. Чехова и В. Шукшина; тексты, написанные в разное время в разных жанрах, повести и рассказы Н. В. Гоголя, повести и рассказы В. Набокова, фантастические повести К. Булычева, сказочные повести А. Волкова. Более того, в одном ряду оказываются изученными произведения разных *родов* литературного творчества это лирика (А. С. Пушкин), эпос, драма («Ревизор» Н. В. Гоголя), а вместе с ними и образцы переводной литературы (произведения С. Лема). Объемы исследованных совокупностей авторских текстов существенно рознятся - от 70 181 до 6 571 689 символов ASCII (отношение между ними составляет » 1:93), но вопросы репрезентативности выборок в этом отношении не оговорены. Между тем при сплошном обследовании совокупностей объемом в миллионы символов неизбежно возникновение шума в области исследуемых характеристик совокупности.

Таким образом, *конкретная реализация* методики некорректна в двух аспектах: 1) постановки задачи; 2) статистическом. Это делает затруднительным вывод о том, насколько она эффективна для решения конкретных ПА. Решение последних подразумевает сопоставление текстов, написанных в одном жанре и в ограниченный несколькими годами период. В этих текстах, как правило, поднимаются одни и те же темы, излагаются похожие мысли в сходных стилистических формах и лексике, что и затрудняет однозначное решение проблемы[15]. Важно также, что объемы контрольных текстов в эксперименте Д. Хмелева составили не менее 70 тыс. символов. Но нередко конкретные ПА касаются текстов в десятки раз

меньших. Существенно и то, что правильно было определено авторство 69 текстов, т. с. доля ошибочных ответов составляет почти 16%, а ранг правильного варианта авторства для некоторых текстов равен 29, 46 и даже 63 - из 82 возможных.

Важен и вопрос о непосредственном объекте анализа. Как полагает Д. В. Хмелев, последовательность пар букв отражает морфемную структуру словоформ текста. Действительно, регулярно в тексте должны появляться определенные ;твухбуквенные флексии, суффиксы, префиксы, а также константные элементы этих и корневых морфем. Но верно и то, что огромное количество парных сочетаний букв неадекватно морфемному членению слов и случайно совпадает с теми или иными морфемами. Впоследствии в рамках этой методики объект анализа был расширен главным образом за счет грамматических классов слов и их семантико-грамматических разрядов, но эксперимент был повторен при тех же исходных условиях[16].

Методологически последовательный способ атрибуции произведений, основанный на изучении частот парной встречаемости грамматических классов слов, был описан Л. И. Бородкиным в 70-е гг. XX в.[17] И сследовато!гь предлагает в целях атрибуции использовать механизмы анализа, выработанные в рамках теории графов. В своих рассуждениях Бородкин отталкивается от индекса различия Р, введенного В. Фуксом. Индекс вычисляется для любых двух текстов, каждый из которых представлен матрицей частот переходов, и представляет собой «сумму мер различия по всем соответствующих клеткам» этих двух матриц. Однако из-за большого количества ячеек сравниваемых матриц при выгчислении индекса *Р* важные различия могут смешиваться с несущественными. Так, Л. И. Бородкин приходит к выводу о необходимости вычленения из общего фона грамматических связей только самых сильных. При этом «сила» связи понимается статистически: редко встречающиеся связи случайны, несущественны для АС. Для вычленения сильных грамматических связей вводится некоторое пороговое значение, с которым сравниваются полученные частоты. Опыт исследований Л. И. Бородкина не был учтен в методике атрибуции Д. В. Хмелева. Вероятно, поэтому в ней не предусмотрены процедуры оценки информативности (значимости) тех или иных частот парной встречаемости изучаемьгх явлений текста.

Интересен подход к атрибуции, предложенный И. О. Тарнопольской[18]. В качестве реализации ЦМ здесь также рассматривается последовательность букв в тексте, но основанием для сравнения текстов является коэффициент диаграммной энтропии (КДЭ), или количественная мера (информационной) неопределенности текста. Значение КДЭ растет вместе с увеличением длины выборки лишь до так называемой точки стабилизации, а затем фиксируется. Для всех текстов в значении КДЭ неизменна цифра до запятой (7), а различия между текстами выражаются в двух цифрах после запятой. КДЭ одинаков для разных текстов одного автора и различен для текстов разных авторов.

Исследователь ставит задачу определения длины выгборки, достаточной для достижения «точки стабилизаггии». Для разных текстов эта длина колеблется в интервале от 10 до 17 тысяч знаков, что почти на порядок меньше длин выгборок, исследуемых Д. В. Хмелевым. Начальная точка поиска оптимальной длины 1200 знаков, длина шага - 1000 знаков. Определение чувствительности метода в каждом конкретном случае также выгодно отличает методику от предложенной Д. Хмелевым. Специально оговорена зависимость инвариантного значения КДЭ от вставок чужого текста в авторский. В описанном И. О. Тарнопольской эксперименте по атрибуции соблюдены требования жанровой и хронологиче-

ской однородности исследуемых текстов - были изучены две редакции «Синопсиса»: основной текст, написанный в Кисво-Печерской лавре в 1674 г., и дополнение к нему, сделанное в 1680 г. Метод позволил выявить в рамках редакции 1674 г. три отрывка, которые характеризуются различными значениями КДЭ, отличающимися и от его значения для редакции 1680 г. Для всех отрывков «точка стабилизации» была достигнута при объемах менее 12 тысяч знаков.

Методики атрибуции, рассматривающие в качестве реализации ЦМ парную встречаемость букв в тексте, направлены на изучение таких закономерностей естественных текстов, которые еще не имеют названия с точки зрения лингвистики и теории стилей. Остается неясным, что именно описывают эти закономерности и характеризуют ли они АС. Существенно ограничивает область применения этих методик то, чго они не позволяют атрибутировать тексты малых объемов.

Как отмечает Л. В. Милов, в русской науке наибольшее развитие получили два подхода к атрибуции, основанных на анализе синтаксиса текста[19]: 1) направленный на построение и анализ «графов синтаксических связей в рамках типических фраз и предложений» (в качестве примера приведена методика И. П. Севбо); 2) смыкающийся со стилемегрией[20] и направленный на изучение закономерностей «во взаимосвязях между различного рода синтаксическими структурами» (в качестве примера приведена методика М. А. Марусенко).

Графические изображения синтаксических структур типических авторских фраз наглядны, но верное определение автора при сравнении этих изображений затруднено ввиду действия двух факторов. Во-первых, один и тот же автор можег быть обладателем сразу нескольких типических фраз (отражая структуру предложений, графы зависимостей должны быть отзывчивы на «пульсацию» стиля писателя). При большом объеме выборок это становится серьезной проблемой, которую И. П. Севбо решает введением семи определенных статистических характеристик (параметров) графа. Здесь и вступает в силу действие второго фактора - нельзя знать заранее, какие именно параметры способны учесть специфику АС. Поэтому избранные Севбо параметры не всегда эффективны в описании стилистических различий.

Неверны предположения о том, что какой-либо один уровень языка текста позволяет полностью описать особенности индивидуального АС его создателя, что существуют параметры, оптимальные в этом отношении. В рамках задач классификации текстов исследователи приходят к выводу о необходимости перехода на системы многомерной классификации, а проблему выбора параметров решают введением особых этапов исследования определения параметрического пространства и снижения его размерности[21].

Выполнение обеих этих задач предусмотрено в рамках методики атрибуции, основанной на применении методов теории распознавания образов. Лингвистический анализ текста здесь направлен на описание его морфолого-синтаксической структуры. Синтаксический анализ текста, как и в методике И. П. Севбо, в первую очередь предполагает описание структурной организации синтаксиса текста. Информация о линейной организации синтаксиса не менее ценна для характеристики АС, так как отражает порядок следования синтаксических структурных элементов. Важной задачей является поиск путей совместного использования синтаксических параметров, отражающих как структурные, так и линейные порядки внутри синтаксиса текста.

ПРИМЕЧАНИЯ

[1] Опираемся на освещение этого вопроса в кн.: *Марусенко М. А.* Атрибуция анонимных и псевдонимных произведений методами распознавания образов. Л.: Изд-во ЛГУ, 1990. С. 19-23.

[2] Там же. С. 18-19.

[3] *Севбо И. П., Петунии Ю. И., Г'алюта Е. Д.* Эксперимент по распознаванию автора, основанный на предварительном статистическом исследовании синтаксических структур // Структурная и математическая лингвистика. Киев, 1977. Вып. 5. С. 103.

[4] *Захаров В. Н., Рогов А. А., Сидоров Ю. В.* Проблема грамматического инварианта Достоевского и атрибуция анонимных и псевдонимных статей в журналах «Время» и «Эпоха» (! 861 -1865) // русский язык: исторические судьбы и современность: Труды и матер. МеЖдунар. контр. (13-16 марта 2001 г.). М., 2001. С. 404-405.

[5] *Рогов А. А., Сидоров 10. В., Король А. В.* Автоматизированная система обработки и анализа литературных текстов «СМАЛТ» // Русский язык: исторические судьбы и современность: Труды и матер. И-го Междунар. контр, исслед. рус. яз. (18-21 марта 2004 г.). М., 2004. С. 485-486; *Сидоров Ю. В.* Математическая и информационная поддержка методов атрибуции литературных текстов на основе формально-грамматических параметров: Автореф. дис. на соис. учен, степени канд. тех. наук. Петрозаводск: Изд-во Петр ГУ, 2002.

[6] *Захаров В. К, Рогов А. А., Сидоров Ю. В.* Указ соч. С. 405.

[7] *Виноградов В. В.* Проблема авторства и теория стилей. М.: Гослитиздат, 1961; *Марусенко М. А.* Указ. соч. С. 4-23; *Слепак Б. Я.* Некоторые теоретико-методологические предпосылки качественно-количествешюй концепции стиля //Уч. зап. Тартуского гос. ун-та. Тарту, 1982. Вып. 619: Вопросы сопоставительной и прикладной лингвистики. С. 107-117.

[8] *Захаров В. П., Рогов А. А., Сидоров Ю. В.* Указ соч. С. 405.

[9] Кто написал «Тихий Дон»? (Проблема авторства «Тихого Дона») / Г. Хетсо, С. Густавссон, Б. Бекман, С. Гил. М.: Книга, 1989.

[10] *Марусенко М. А.* Указ. соч. С. 15-16, 18-19, 22.

[11] *СидоровЮ. В.* Указ. соч. С. 19.

[12] *Марусенко М. А., Мельникова Е. Е., Родионова Е. С.* Атрибуция анонимных и псевдонимных статей, опубликованных в журналах «Время» и «Эпоха» в 1861-1865 годах// Квантитативная лингвистика: исследования и модели (КЛИМ-2005): Материалы Всеросс. науч. конф. (6-10 июня 2005 г.). Новосибирск, 2005. С. 283-294.

[13] См., напр.: *Марусенко М. А.* Указ. соч. С. 25.

[14] *Хмелев Д. В.* Распознавание автора текста с использованием цепей А. А. Маркова // Вестник МГУ. Серия Филология. 2000. № 2. С. 115-126.

[15] О конкретных проблемах авторства см.: В поисках потерянного автора: Эподы атрибуции / М. А. Марусенко, Б. Л. Бессонов, Л. М. Богданова и др. СПб.: Филологический ф-т СПбГУ, 2001.

[16] *Кукушкина О. В., Поликарпов А. А., Хмелев Д. В.* Определение авторства текста с использованием буквенной и грамматической информации // Проблемы передачи информации. 2001. Т. 37. №2, С." 96-108.

[17] От Нестора до Фонвизина: Новые методы определения авторства/Л. В. Милов, Л. И. Бородкин, Т. В. Иванова и др. М.: Прогресс, 1994. С. 33-39.

[18] *Тарнопольская И. О.* Диаграммная энтропия текста и атрибуция анонимных текстов: результаты тестирования методики // Информационный бюллетень ассоциации «История и компьютер». М., 1998. Март. № 23. С. 66-68.

[19] От Нестора до Фонвизина: Новые методы определения авторства. С. 8-9.

[20] *Мартыненко Г. Я.* Основы стилеметрии. **Л.: Изд-во ЛГУ,** 1988.

[21] См. об этом: *Марусенко М. А.* **Указ.** соч. С. 82-91.