УДК 81'33
ББК 81.1

Степан Сергеевич Верхозин,
аспирант,
Иркутский государственный лингвистический университет
(Иркутск, Россия), e-mail: stepaverkhozin@rambler.ru

К вопросу о лингвотеоретических основах методик авторизации текста

В данной статье рассматривается такое актуальное направление прикладной лингвистики, как авторизация текста. Авторизация текста – соотнесение тексту соответствующих ему атрибутов, к которым причисляется имя создателя, но также жанр, время и место создания текста. Типовые случаи авторизации текста описываются тремя ситуациями: множественной неопределённостью, сравнением по образцу и конкуренцией образцов. Данная проблема относится к числу классических и часто рассматривается в рамках «количественной стилистики» – стилеметрии. История применения количественных методов в зарубежной лингвистике берёт своё начало в XVIII веке в работах Э. Мэлоуна, а в отечественной в XX веке в статье Н. А. Морозова. Автороведческая экспертиза имеет три основные практические области применения: судебная экспертиза, исследование языка и литературы, а также образование и психология. Методы авторизации текста достаточно разнообразны и позволяют взглянуть на данную проблему со стороны разных уровней языка.

Ключевые слова: квантитативные методы, авторизация текста, атрибуция текста, доля служебных слов, среднее значение, коэффициент синтаксического разнообразия, коэффициент лексического разнообразия речи.

Stepan Sergeevich Verkhozin,
Postgraduate Student,
Irkutsk State Linguistic University
(Irkutsk, Russia), e-mail: stepaverkhozin@rambler.ru

On Issue of Linguistic and Theoretical Foundations of Authorship Attribution Methods

This article focuses on such an urgent sphere of applied linguistics as authorization, also known as authorship attribution. Authorship attribution is identification of correlation between a text and its attributes, such as an author's name, genre, etc. The basic types of authorization can be described by the following three situations: multiple ambiguity, pattern comparison and pattern competition. This problem is a classical one and is often studied in quantitative stylistics. The history of using quantitative methods in foreign linguistics originates from E. Malone's works and in Russian one – from an article by N. Morozov. Authorship examination has three main spheres of practical use: forensic inquiry, language and literature researches, education and psychology. Authorization approaches are rather various and allow one to study texts according to different language levels.

Keywords: quantitative methods, text authorship, authorship attribution, percent of function words, average value, syntactical diversity coefficient, lexical diversity coefficient.

Количественные, или квантитативные, методы в последнее время прочно закрепились в качестве весьма полезного и информативного инструмента лингвистических исследований. Они могут применяться как отдельно для решения определённой языковедческой проблемы, так и в комплексе с другими, качественными подходами.

Существует множество областей применения квантитативных методов с лингвистической точки зрения: к ним можно отнести как самостоятельные дисциплины, такие как математическая или статистическая лингвистика, так и некоторые другие направления, где количественные методики выступают в качестве средств разрешения тех или иных вопросов, связанных с функционированием языка. К последним можно отнести психолингвистику, компьютерную лингвистику, судебную лингвистику и т. д.

Среди этих областей для нас наибольший интерес представляет судебная лингвистика в общем и определение авторства текста в частности.

© С. С. Верхозин, 2013

В широком смысле, авторизация текста – соотнесение тексту соответствующих ему атрибутов, к которым причисляется не только имя создателя, но также жанр, время и место создания текста [7].

Типовые случаи авторизации текста описываются следующими ситуациями [1, с. 43]:

- Множественная неопределённость;
- Сравнение по образцу;
- Конкуренция образцов.

Современные формальные методики установления авторства текста опираются на стохастическую модель порождения речи. В основе этой модели лежит представление о том, что с возрастанием объёма текста авторские особенности языкового оформления содержания становятся устойчивыми с вероятностной точки зрения, что позволяет устанавливать авторство по стабильно повторяющимся формальным характеристикам текста [5].

Проблема авторизации текста относится к числу классических проблем филологического исследования. Часто она рассматривается в рамках «количественной стилистики» – стилеметрии [1, с. 41]. Стилеметрия – прикладная филологическая дисциплина, занимающаяся измерением стилевых характеристик текстов с целью их упорядочивания и систематизации (типологии, диагностики, атрибуции, классификации, датировки и тому подобное). Задачей этой дисциплины является поиск каких-то формальных количественных параметров текста, которые позволяли проводить строгие различия между двумя, максимум тремя авторами [4].

Важно отметить, что использование методов установления авторства, основанных на анализе формальных характеристик, или, наоборот, на чисто стилистическом и литературном анализе текста, не представляется возможным. С другой стороны, грамматические особенности авторского стиля – частота употребления неполнозначных, синсемантических слов и дискурсивных маркеров (частиц, союзов, предлогов, некоторых модальных слов, вводных выражений) – для текстов порядка 1000–2000 словоупотреблений сохраняются. Это и даёт возможность подтверждать или опровергать имеющиеся гипотезы относительно авторства текста. Необходимо подчеркнуть, что гипотеза об авторстве в любом случае должна присутствовать, причём она должна формулироваться на экстралингвистических основаниях [5].

Перспектива объективизации экспертного знания была обнаружена в использовании количественных, статистических методов анализа текста. В зарубежной лингвистике квантитативные методы впервые стали применяться ещё в XVIII веке: в 1787 году специалист в хронологии и происхождении пьес Уильяма Шекспира Эдмонд Мэлоун (Edmond Malone) оспорил авторство трёх частей «Henry IV», написанного английским драматургом. Мэлоун основывал свои предположения на некоторых, нехарактерных для Шекспира, метрических показателях: например, на редком использовании двойных окончаний и рифмованных строк [9, с. 8–12].

В нашей стране основателем квантитативного направления можно считать Н. А. Морозова, опубликовавшего в 1915 г. статью «Лингвистические спектры. Средство для отличия плагиатов от истинных произведений того или другого известного автора. Стилеметрический этюд». В этой работе Н. А. Морозов предположил, что для индивидуального стиля писателя показательными являются именно служебные слова, поскольку они никак не связаны с темой и содержанием книги [7].

С тех пор было выработано и предложено множество различных методов авторизации текста, но ни один из них не является общепризнанным. С точки зрения разных подходов, в качестве достоверных количественных критериев определения авторства текста признавались: доля служебных слов как авторский инвариант, средняя длина слова, средняя длина предложения, индекс разнообразия лексики и многие другие. У каждого из подходов есть свои преимущества и свои недостатки, свои сторонники и свои противники. Некоторые учёные стараются совмещать качественные методики и количественные, дополняя, таким образом, результаты анализа по каждой из них. Однако уместным будет сказать, что квантитативные, или формальные, методы определения авторства видятся более объективными, поскольку позволяют установить некоторые языковые явления, которые очень трудно, если вообще возможно подделать, как то процент использования автором служебным слов.

На сегодняшний день методы атрибуции текста отличаются большим разнообразием: одни направлены на изучение лексических показателей, другие на изучение синтаксических или грамматических характеристик. Существуют также некоторые другие подходы, авторы которых предлагают комплексный анализ текста на нескольких языковых уровнях. Для обоснования перспективности того или иного метода необходимо провести их обзор и анализ.

Первым из рассматриваемых методов является так называемая «методика Хетсо» – норвежского учёного, специалиста по использованию компьютерных технологий в атрибуции литературных произведений. В 1979 году он исследовал спорные тексты, авторство которых приписывается Ф. М. Достоевскому, применив при этом несколько лингвостатистических критериев [там же]. К ним относятся:

• средняя длина слова в буквах (вычисляемая на основании выборок размером 500 текстовых слов);

• общее распределение длины слова, средняя длина предложения в словах (вычисляемая на основании выборок размером в 30 предложений);

• общее распределение длины предложения;

• лексический спектр текста на уровне словаря;

• лексический спектр текста на уровне текста;

• индекс разнообразия лексики.

К недостаткам этого подхода следует отнести то, что не все статистические критерии, использованные Г. Хетсо, могут дать достаточно надёжные результаты. Например, такая характеристика, как распределение длины слова, по словам Т. С. Менденхола, может быть полезна в тех случаях, когда надо исключить авторство того или иного человека, а распределение длины предложения может дать достоверные результаты только тогда, когда выборка исследуемого текстового материала достаточно велика. Иными словами, сомнительным является то, что, используя этот критерий, можно определить авторство, например, небольшого письма или записки.

А. Н. Баранов в книге «Введение в прикладную лингвистику» предложил метод, суть которого заключается в проведении количественного анализа квазисинонимичных лексем – близких по значению слов или устойчивых словосочетаний (фразеологизмов) [1, с. 43]. Автор основывается на том, что интересную информацию об идиолекте писателя даёт изучение частотных характеристик служебных и модальных слов. Так, частицы «разве» и «неужели» по-разному распределены в романах М. Булгакова «Мастер и Маргарита» и «Белая гвардия». «Разве» значительно чаще встречается в «Мастере и Маргарите», а «неужели» – наоборот. Значение частицы «разве» предполагает более активную, действенную позицию говорящего, подвергающего сомнению некоторое положение дел. В противоположность «разве», частица «неужели» скорее указывает на то, что некоторое положение дел практически принимается говорящим, и он лишь недоумевает, удивляется, почему оно имеет место. Учитывая значение этих частиц, указанные факты распределения «разве» и «неужели» можно интерпретировать как лингвистический коррелят авторской позиции в изображении событий и действий героев.

В основе следующего метода авторизации лежит одно из главных коммуникативных качеств речи – её богатство. В основе содержания понятия «богатая речь» лежит мысль о том, что речь тем богаче, чем реже повторяются в ней одни и те же знаки и цепочки знаков, способы внутренней организации этих последовательностей. Это означает, что речь тем богаче, чем она разнообразнее на всех своих языковых уровнях. Такое обстоятельство позволяет отождествлять понятия «богатство» и «разнообразие» речи [3, с. 25].

Существующие методы оценки богатства (разнообразия) речи в пределах того или иного её фрагмента, целого текста или их совокупности опираются, как правило, на интуицию исследователя, что может привести к субъективизму этой оценки. Крайне редки попытки применения объективных, опирающихся на количественные меры способов характеризации этой стороны речи, таких как строгие и статистически внятные коэффициенты речевого разнообразия. Расчёт коэффициента речевого разнообразия должен предполагать взаимосвязь, как

минимум, двух параметров: лексического разнообразия и степени синтаксической сложности. На этих уровнях легко достигнуть формальной соотносимости коэффициентов [там же].

Объективным критерием определения авторства может являться доля всех служебных слов в последовательном фрагменте [6, с. 768]. Простота характеристики позволяет провести объёмный статистический эксперимент. Как показали результаты предыдущих исследований, базирующихся на данном методе, на протяжении всего периода творчества писателя доля служебных слов (два знака после запятой) остаётся постоянной. В то же время эта доля значительно варьируется от писателя к писателю, принимая значения от 0,20 до 0,30. Это позволило авторам метода серьёзно обосновать плагиат М. А. Шолохова, поскольку его авторский инвариант значительно отличается от значения найденного параметра на первых частях «Тихого Дона» [там же].

Очень интересным и одним из самых новых является метод, предложенный шведским учёным Себастьяном Бернгардссоном (Sebastian Bernhardsson) из университета города Умео в докторской диссертации «The meta book and size dependent properties of written language» (2009). Сущность подхода заключается в изучении так называемого концепта «мета-книги» (meta book), как способа описания зависимости от размера некоторых статистических свойств, наблюдаемых в книгах [8, с. 43–54].

Идея автора состоит в том, что само создание текста может быть описано как процесс, в ходе которого автор берёт часть текста из большой материнской книги (метакниги) и переносит её на бумагу. Эта «метакнига» – воображаемая бесконечная книга, которая даёт представление о характеристиках частотности слова, описывающих всё, что автор мог когда-либо написать. Это не имеет никакого отношения к семантике и к фактическому значению того, что написано, а скорее говорит о лексическом словаре, уровне и типе образования писателя, о его личных предпочтениях [там же].

В рамках исследования С. Бернгардссона было проверено, как в текстах трёх писателей: Томаса Харди, Генри Мелвилла и Дэвида Лоуренса – реализуется так называемый закон Ципфа. Этот закон, открытый в 1935 году лингвистом Джорджем Ципфом (George Kingsley Zipf), гласит, что частота какого-либо слова в тексте обратно пропорциональна его рангу – месту в списке слов текста, отсортированных по частоте.

Шведские физики в своей статье показали, что этот закон не так универсален, как считал Джордж Ципф. Они обнаружили, что частота появления новых слов по мере роста объёма текста меняется у разных авторов по-разному, причём эта закономерность не зависит от конкретного текста, а только от автора.

Последний метод, который будет описан в статье – метод распознавания образов, в котором текст рассматривается как сложный лингвистический объект, характеризующийся обширным инвентарём элементов и многоуровневостью анализа. В основу нового метода атрибуции анонимных и псевдонимных произведений был положен многомерный статистический анализ, представленный в его наиболее развитой форме – теории распознавания образов [2].

В терминах распознавания образов стиль определяется как набор свойств (параметров), характеризующих состав, способы объединения и статистико-вероятностные закономерности употребления речевых средств, образующих данную разновидность языка. Набором свойств, характеризующих структуру текста в синтаксическом аспекте, становится в данном случае совокупность информативных параметров, чей состав определяется путём выполнения специальной процедуры отбора информативных параметров для каждого конкретного случая.

Важное теоретическое положение рассматриваемой работы заключается в том, что процедура атрибуции расчленяется на три относительно самостоятельных этапа [2]:

• Формирование литературно-критической атрибуционной гипотезы, которое выполняется методами традиционного филологического анализа с применением всех доступных субъективных и объективных методов и приёмов атрибуции.

• Поверка литературно-критической гипотезы, для выполнения которой используются средства теории распознавания образов.

• Интерпретация результатов проверки атрибуционной гипотезы.

Гипотеза считается статистически подтверждённой, если результаты распознавания согласуются с исходной литературно-критической атрибуционной гипотезой (при установленном уровне значимости). В противном случае гипотеза считается опровергнутой и проверяется либо альтернативная гипотеза, либо переформулированная исходная гипотеза. При реализации такой схемы атрибуции статистико-вероятностные методы анализа языка и стиля используются лишь в качестве вспомогательных средств проверки исходной атрибуционнной гипотезы, сформированной с помощью филологических методов атрибуции [там же].

Разумеется, рассмотреть все существующие методы атрибуции текста в рамках одной статьи, не представляется возможным. У каждого лингвостатистического подхода есть как свои достоинства, так и свои недостатки. Порой результаты анализа одного текста или набора текстов, обработанных с помощью статистических критериев на предмет установления авторства разными учёными с помощью разных методов, получаются диаметрально противоположными: один лингвист атрибутирует материал конкретному автору, а второй это авторство опровергает. Поэтому не случайно, что ни один из методов не имеет особого преимущества перед другими, с точки зрения своей распространённости и популярности среди учёных. Лингвист, предложивший сравнительно новый подход или использующий определённый уже существующий, зачастую рассматривает его как единственно верный и отвергает другие. По этой причине перспективным видится использование целого комплекса различных статистических методов, рассматривающих различные аспекты атрибутируемого текста. Результаты, полученные на основе такого многомерного подхода, позволят сделать более объективные выводы об авторстве, увеличить надёжность общего результата, а также исключить возможную ошибку при авторизации. Конечно, такая методика может вызвать определённые проблемы при обработке текстового материала, вследствие его потенциально большого объёма или сложности математического либо статистического аппарата, применяемого в некоторых подходах, но современное состояние информационных технологий, компьютерной техники, а также программного обеспечения позволяют значительно снизить трудоёмкость такой работы.

В заключение, следует отметить, что авторизация текста как область применения статистических знаний – сравнительно новое, быстро развивающееся и востребованное направление в науке в целом и в лингвистике в частности. Практическое же применение методы атрибуции текста находят в области судебной лингвистики (например, судебные дела о плагиате, определение авторства анонимного письма и т. д.). Конечно, в сферу этого направления прикладного языкознания входит не только собственно установление авторства, предмет этой науки гораздо шире и охватывает многие аспекты исследования языковых явлений в деле судопроизводства (в судебной лингвистике выделяют два больших раздела: судебная фонетика и судебная лексикология). В данной статье внимание уделено именно методам атрибуции текстов, коих, как было показано, насчитывается довольно много. Слабым местом можно назвать то, что отсутствует один общепринятый подход, который применяло бы большинство экспертов. Конечно же, каждый учёный, имеющий соответствующую учёную степень и авторитет в учёном сообществе, имеет право на выработку своего собственного метода. В конце концов, все статистические гипотезы носят вероятностный характер. Однако работы над разработкой новых количественных инструментов исследования авторства и усовершенствования старых продолжаются и видятся крайне перспективными.

Список литературы

1. Баранов А. Н. Введение в прикладную лингвистику: учеб. пособие. М.: Эдиториал УРСС, 2001. 347 с.

2. В поисках потерянного автора. Этюды атрибуции: [сайт]. URL: http://attribution.corneille-moliere.com/ (дата обращения: 15.04.2011).

3. Варфоломеев А. П. Психосемантика слова и лингвостатистика текста: метод. реком. к спецкурсу / Калинингр. ун-т. Калининград, 2000. 37 с.

4. Мартыненко Г. Я. Основы стилеметрии. Л.: Изд-во Ленингр. ун-та, 1988. 173 с.

5. Рахимова А. А. Лингвистическая экспертиза // Вестник КАСУ № 3. 2005. С. 63–67. URL: http://www.vestnik-kafu.info/journal/3/83/ (дата обращения: 8.03.2011).

6. Фоменко В. П., Фоменко Т. Г. Авторский инвариант русских литературных текстов / предисл. А. Т. Фоменко // Фоменко А. Т. Новая хронология Греции: Античность в средневековье. М.: Изд-во МГУ, 1996. Т. 2. 886 с.

7. Хмелёв Д. Краткая история разработки методик определения авторского стиля. URL: http://rusf.ru/books/analysis/history.htm/ (дата обращения: 15.01.2010).

8. Bernhardsson Sebastian. Structures in Complex Systems Playing dice with Networks and Books. – Print & Media, Umea 2009. 77 c.

9. Grieve Jack William. Quantitative Authorship Attribution: A history and an evaluation of techniques. – Simon Fraser University. 2005. 282 c.

References

1. Baranov A. N. Vvedenie v prikladnuju lingvistiku: ucheb. posobie. M.: Jeditorial URSS, 2001. 347 s.

2. V poiskah poterjannogo avtora. Jetjudy atribucii: [sajt]. URL: http://attribution.corneille-moliere.com/ (data obrashhenija: 15.04.2011).

3. Varfolomeev A. P. Psihosemantika slova i lingvostatistika teksta: metod. rekom. k speckursu / Ka-liningr. un-t. Kaliningrad, 2000. 37 s.

4. Martynenko G.Ja. Osnovy stilemetrii. L.: Izd-vo Leningr. un-ta, 1988. 173 s.

5. Rahimova A. A. Lingvisticheskaja jekspertiza // Vestnik KASU № 3. 2005. C. 63–67. URL: http://www.vestnik-kafu.info/journal/3/83/ (data obrashhenija: 8.03.2011).

6. Fomenko V. P., Fomenko T. G. Avtorskij invariant russkih literaturnyh tekstov / predisl. A. T. Fomenko // Fomenko A. T. Novaja hronologija Grecii: Antichnost' v srednevekov'e. M.: Izd-vo MGU, 1996. T. 2. 886 c.

7. Hmeljov D. Kratkaja istorija razrabotki metodik opredelenija avtorskogo stilja. URL: http://rusf.ru/books/analysis/history.htm/ (data obrashhenija: 15.01.2010).

8. Bernhardsson Sebastian. Structures in Complex Systems Playing dice with Networks and Books. – Print & Media, Umea 2009. 77 c.

9. Grieve Jack William. Quantitative Authorship Attribution: A history and an evaluation of techniques. – Simon Fraser University. 2005. 282 c.

Статья поступила в редакцию 13.10.2012