УДК 519.25:004.8 +81.13

З.И. Резанова, А.С. Романов, Р.В. Мещеряков

О ВЫБОРЕ ПРИЗНАКОВ ТЕКСТА, РЕЛЕВАНТНЫХ В АВТОРОВЕДЧЕСКОЙ ЭКСПЕРТНОЙ ДЕЯТЕЛЬНОСТИ

В статье анализируется результативность вовлечения разнотипных признаков текста в авторoведческую экспертную деятельность, осуществляемую с привлечением количественных методов анализа, разрабатываемых в теории информации и интеллектуальном анализе данных. Характеризуется релевантность количественных методов анализа языковых элементов текстовой структуры (единицы означивания, в том числе фонетико-графические элементы, элементы грамматики текста), а также собственно текстовых признаков (структура и графическое оформление, метаданные документа). Противопоставляются параметры, количественная квалификация которых проявляет диагностическую силу относительно широкого спектра текстов, и параметры, действенность которых зависит от конкретных целей атрибуции, своеобразия анализируемых текстов и их авторов.
Ключевые слова: *авторoведческая экспертиза, стилометрия, междисциплинарные методы исследования, информатика, интеллектуальный анализ данных, авторская атрибуция текста.*

Вопросы авторской атрибуции текста относятся к числу проблем, которые могут быть поставлены и разрешены как междисциплинарные с опорой на теоретико-методологический аппарат юриспруденции, лингвистики, информатики. Данные науки занимают вполне определенные позиции относительно рассматриваемой проблемы: необходимость авторской идентификации текста актуализируется в области юридической (реже – филологической) практики, разрешается она с привлечением количественных методов анализа, разрабатываемых в теории информации и интеллектуальном анализе данных, параметризация текста осуществляется с опорой на результаты лингвистического анализа. Однако мы полагаем, что в настоящее время не достигнута искомая степень интеграции данных наук: лингвисты в анализе не используют потенциал количественных методов анализа, разработанных в современной теории информации, пользуясь элементарными количественными подсчетами, оперируя, как правило, понятиями относительного преобладания того или иного свойства/признака текста. Специалисты в области теории информации не учитывают результаты лингвистических исследований в области теории языковой личности, лингвоперсонологии, лингвистики текста, стилистики. Именно недостаточная интегрированность методов и приемов авторoведческих исследований, проводимых представителями разных наук, на наш взгляд, замедляет прогресс в данной научной сфере.

В статье обсуждается одна из проблем авторoведческой экспертизы – выбор единиц текста, количественный анализ которых может быть наиболее эффективным, с позиций задач, *разрабатываемых и решаемых в теории информации и интеллектуальном анализе данных* в общем проблемном поле авторской атрибуции текста. Так как данные задачи по своему целеполага-

нию непосредственно соотнесены с рядом смежных проблем, решаемых собственно лингвистическими методами, мы полагаем, что вовлечение в обсуждение данных проблем лингвистов было бы весьма полезным. Данная статья, написанная лингвистом и специалистами в области информатики, и инициирует такое обсуждение в научном лингвистическом издании.

Выбор языковых характеристик авторского стиля является важнейшим этапом при идентификации авторства текста. По подсчетам Дж. Рудмана [1], для идентификации автора разными исследователями используется порядка тысячи различных групп характеристик. Уже этот факт косвенно свидетельствует, во-первых, об объективных причинах такого исследовательского разнобоя – это показатель реальной сложности, многомерности текста, обусловливающего необходимость проверки релевантности разного рода единиц в данном аспекте, их способности быть устойчивым показателем своеобразия авторского стиля. Во-вторых, это показатель того, что данное направление науки находится в ситуации становления и далеко от разрешения поставленной проблемы. В-третьих, многообразие выделяемых единиц, лишь частично пересекающихся в практике авторoведческой экспертизы, свидетельствует о различии типов текстов, с которыми работают авторы, и о разнообразии конкретных задач, решаемых исследователями.

Актуализация вопроса о составе единиц, диагностирующих своеобразие авторского стиля, может быть связана с разрешением разнонаправленных задач: либо с поиском признаков текста, релевантных при решении конкретной исследовательской задачи, либо с выявлением относительно универсального набора признаков (при этом принципиален вопрос о степени универсальности такого корпуса). Решение второй задачи имеет и общетеоретическую, и практическую направленность и связано с частными языковедческими направлениями: теорией текста, лингвоперсонологией, лингвистикой универсалий и рядом других направлений. Выделение такого набора признаков опирается на знания об универсальных свойствах языков, о наличии в них типовых единиц, не зависящих от глобальных типологических языковых различий. Актуален этот вопрос вследствие того, что в поле зрения европейских исследователей находятся, как правило, языки флективного типа. В известных нам работах не приводятся данные о наборе признаков текста, написанных на языке, к примеру, изолирующего типа. Данное замечание также применимо и к признакам графического оформления текста: в имеющейся в нашем распоряжении литературе обсуждаются особенности графического облика текста, написанного в звукобуквенных системах письма (ср. глобальные отличия текста, написанного на основе иероглифики и звукобуквенной графики).

Значимы и менее «сильные» различия синтетичности и аналитичности в области флективных языков: в последнем случае исследователи могут работать кооперативно, так как накопленные данные о верификационной силе параметра, доказанной для одного типа языков, могут быть проверены относительно другого (см. об этом далее). В настоящее время в том случае, когда говорят об универсальном наборе признаков, как правило, имеют в виду их набор для текстов разных стилей и жанров в пределах *одного языка.* И это во многом обусловлено тем, что стилометрия сложилась как практически ориен-

тированная отрасль знания, область задач которой ограничивалась данными одного этнического языка. Однако и при работе с текстами одного языка проблему составляет соотношение универсальности параметров текста, способных устойчиво идентифицировать речевое произведение автора безотносительно к смене его коммуникативных позиций, речевых жанров, и свойств текста, наиболее результативно «работающих» при решении частных исследовательских задач. Эта проблема в наибольшей степени актуализировалась в настоящее время, когда стилометрия из сферы филологической квалификации текста (в начале решались задачи определения авторов художественных текстов) все более перемещается в область криминалистической экспертизы, что предопределяет вовлечение в сферу анализа текстов новых систем коммуникации – обыденных письменных, электронной коммуникации и т.д.

В данной статье, обсуждая вопросы выбора единиц текста, вовлекаемых в анализ при автороведческой экспертизе, которая осуществляется с привлечением количественных методов анализа, разрабатываемых в теории информации и интеллектуальном анализе данных, остановимся на следующих аспектах: проверенность результативности признака на материале конкретных языков; релевантность признака при анализе текстов разных стилей и жанров, при решении конкретных исследовательских задач; наличие или отсутствие осложняющих факторов при выборе конкретных единиц, выступающих в качестве параметра идентификации автора.

Признаки текста, используемые в настоящее время для идентификации автора, можно разделить на группы и подгруппы (рис. 1).

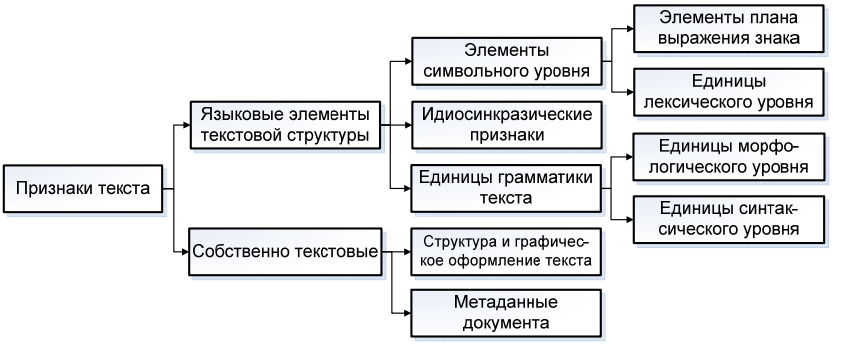

Рис. 1. Типы признаков текста, используемых в автороведческой экспертизе

Языковые и собственно текстовые признаки могут быть как формальными, так и формально-семантическими. Как показал анализ, признаки выделенных групп в разной степени обнаруживают контекстную и конситуативную зависимость.

1. Языковые элементы текстовой структуры

1.1. Единицы символьного уровня

На этом уровне выявляется специфика номинативного состава текста в широком смысле этого слова, анализ ведется по формальным и формально-

семантическим единицам означивания. В группу включаются собственно знаковые единицы и элементы плана выражения знаковых единиц.

1.1.1. Прежде всего в методиках этого типа выявляется специфика текста на уровне *единиц плана выражения*: буквенный состав, распределение слов по длине и др.; анализируются как частота появления отдельных символов, так и их последовательность с детализацией по цифрам, буквам, знакам пунктуации, а также наиболее частых *N*-грамм символов вместо использования полного набора [2–9]. Кроме того, каждое слово имеет такие свойства, как, например, количество слогов или определенных морфем и т.д. Простейшие свойства слов в той или иной мере использовались в автороведческой экспертизе в сочетании со статистическими мерами [10, 11], однако, как правило, проигрывали более простым в плане трудоемкости извлечения/анализа признакам.

Исследования отечественных и зарубежных авторов свидетельствуют, что данная группа признаков нередко является более информативной при идентификации автора, чем признаки высоких уровней, причем данный принцип работает для многих языков, в числе которых английский, датский, русский, итальянский, греческий и др.

Высокая релевантность этих параметров определяется и тем, что, будучи собственно формальными, эти признаки в обычных условиях речепорождения не рефлексируются коммуникантами, вследствие чего вероятность намеренного моделирования текста по данным параметрам практически равна нулю.

Вследствие их принадлежности к низшему уровню языковой системы высока вероятность их инвариантного проявления в текстах разной жанровой и стилевой принадлежности одного автора.

Информативность данной группы признаков определяется и тем, что для выявления статистических закономерностей в области графического, собственно символьного аспекта достаточными являются тексты меньшей длины, нежели при установлении статистики относительно единиц высших уровней языковой системы.

Преимуществом использования символов и их последовательностей является и простота анализа. Любой программист способен написать компьютерную программу, подсчитывающую частоту появления символов в тексте, немаловажно и то, что при членении текста на единицы данного текстового среза, как правило, программисту нет необходимости опираться на сугубо специфичные лингвистические знания.

1.1.2. Одним из самых простых способов, позволяющих подтвердить или опровергнуть авторство текста, считается использование характерных *особенностей словаря автора*. В данном случае слово интерпретируется не как последовательность графических символов, а как двусторонняя единица лексической системы, маркированная не только семантически, но и функционально-стилистически.

В автороведческой экспертизе словарный срез текста анализируется как минимум в двух аспектах.

Во-первых, выявляется использование *определенных слов автором*, которое может быть явным признаком его индивидуальности. Человек, обладаю-

щий богатым словарным запасом, выражает свои мысли, как правило, более емкими словами и фразами, наиболее близкими к описываемой ситуации, его речь последовательна и выразительна. Люди с небольшим словарным запасом вынуждены ограничиваться повторением одних и тех же слов, отчего их устная и письменная речь выглядит более примитивной.

По словарю текста можно судить об историческом времени, в которое был написан документ, так как все его слова на момент написания уже должны были существовать и, вероятнее всего, входили в активный словарь. Специфические слова могут означать принадлежность автора к определенной группе: по профессионализмам можно судить о профессии автора, уровне образования, по диалектизмам – о географическом месте, в котором вырос автор или жил во время создания текста, жаргонизмы позволяют судить об уровне культуры человека и т.д.

Данный прием автороведческой экспертизы более других соотнесен с практикой исследования идиостиля автора в филологических работах. Однако отметим проблемность использования описанного подхода в автороведческой экспертной деятельности, что, на наш взгляд, определяется рядом факторов. Во-первых, ярких характерных особенностей лексикона у текста, равно как и у автора, может и не быть. Во-вторых, если текст имеет выраженные особенности, то существует большая вероятность их намеренного моделирования. К недостаткам следует также отнести и тот факт, что выявление отличительных черт авторского лексикона во многом носит субъективный характер, зависит от личности исследователя. Кроме того, данный признак проявляет значительную зависимость от смены темы и жанра коммуникации (данное замечание справедливо прежде всего относительно знаменательных единиц, но в целом относится к единицам всех грамматических классов) и, вследствие этого, не может выступать в качестве надежного идентификатора при вовлечении в экспертизу разножанровых текстов одного автора. Существенным ограничением применения данного приема в автороведческой экспертизе, осуществляемой с привлечением количественных методов анализа, разрабатываемых в теории информации и интеллектуальном анализе данных, является и необходимость привлечения к статистическому анализу значительного по объему корпуса текста. Вследствие этого данный прием используется и апробируется, как правило, при определении авторов художественных текстов. Однако с приведенными выше ограничениями прием может быть использован и используется в стилометрии. Фактором, стимулирующим использование данного приема специалистами в области информационных систем, является и то, что при обработке текста в аспекте частотности корпус анализируемых текстов не нуждается в специальной лингвистической разметке.

Во-вторых, в автороведческой экспертизе, осуществляемой на основе последовательного применения количественных методов, при обращении к лексическому уровню текста широко распространено *выявление частотности слов*. Так, в качестве признаков стали использоваться слова (или признаки слов), сгруппированные по частоте встречаемости в тексте. Наиболее типичными для текста являются группы слов, встретившиеся в нем один (hapax legomena) или два раза (dis legomena), – такие слова, как правило, преобладают в большинстве текстов, и чаще всего это знаменательные слова. Неболь-

шая по составу группа служебных слов характеризуется значительной повторяемостью в тексте, и по этой причине, частотность служебного слова может значительно варьироваться в разных текстах.

Например, мерой, учитывающей описанное выше распределение, является оценка, предложенная Юлом (G.U. Yule) в работе [12]:

$$K = 10^4 (\sum i^2 V_i - N) / N^2 ,$$

где V_i – количество слов, встретившихся в тексте i раз, N – количество слов в тексте.

Оценку K можно интерпретировать как коэффициент повторения слова в тексте: чем чаще повторяется произвольно выбранное слово, тем выше коэффициент K. Соответственно, чем выше значение K, тем беднее словарь автора.

Позже были предложены другие оценки разнообразия авторского словаря, однако в работе [13] показано, что постоянными для одного автора являются лишь меры K, предложенные Юлом, и D [14]:

$$D = \sum V_i (v_i / N)((i-1)/(N-1)),$$

но и они не обладают достаточной различительной способностью в случае большого количества идентифицируемых авторов.

Словари авторов можно сравнивать напрямую, например, рассчитывая статистическое расстояние между ними по частоте слов, как это сделано в работе О. Хрулева [15].

Достоинства данного приема мы усматриваем в его апробированности (его результативность проверена на материале разных языков), а также в том, что членение текста на лексические единицы не требует специальной подготовки программиста.

Однако при применении данной методики необходимо осознавать ее ограниченность при решении ряда задач автороведческой экспертизы. Очевидно, что частотность слов в текстах одного автора, относимых к разным стилям и жанровым формам, будет значительно различаться. Вследствие этого, как показывает анализ, эффективно данная методика работает при анализе текстов одного стиля. Недостатком подхода, основывающегося на оценке разнообразия словаря автора, является то, что он по умолчанию основывается на предположении о том, что каждый человек использует на протяжении жизни и в разных коммуникативных ситуациях сформировавшийся словарь. При этом не учитывается, что словарный запас в течение жизни человека изменяется и растет по мере получения новых знаний, варьируется в разных дискурсивных практиках. Таким образом, в составе единиц символьного уровня выделяется группа ***контентно-специфических*** признаков.

С учетом этого аспекта единиц символьного уровня также проводится автороведческая экспертиза. Группа чувствительных к содержанию признаков, вовлекаемых в стилометрический анализ, который основан на применении формально-количественных методов, включает ключевые слова и фразы по

определенной тематике [45, 46] или определенные *N*-граммы ключевых слов [47]. Кроме того, могут подсчитываться специфические сокращения и аббревиатуры или слова и фрагменты текста, написанные на языке, отличном от оригинального языка текста, а также эмотиконы [50].

Основной принцип, на котором базируются приемы автороведческой экспертизы, использующие количественный анализ тематически ограниченных ключевых слов, состоит в том, что если слово часто встречается в текстах одного класса, но редко в текстах другого, то оно, возможно, более значимо для качественного разделения двух классов, чем слово, встречающееся в малом количестве текстов, но во многих классах. Очевидным недостатком привлечения данной группы текстовых признаков является то, что их использование эффективно только в текстах по определенной тематике, однако это не исключает их действенности при решении частных задач автороведческой экспертизы.

* * *

Слабость приемов, основанных на анализе словаря, заключается в том, что при этом, как правило, не учитывается словоизменение. Автор, индивидуальной особенностью которого является употребление определенного слова и его частота, использует его в речи в разных формах. Учет этого фактора особенно актуален для флективного русского языка, тенденция к аналитичности грамматической структуры которого проявлена в значительно меньшей степени, чем в английском языке, на материале которого проведено большее количество стилометрических исследований. Но и исследования, проведенные на материале текстов английского языка с учетом данного фактора [16, 9], строились на количественном анализе не слов, а последовательностей символов, из которых они состоят. Полученные *N*-граммы символов учитывают особенности текста, недоступные при анализе с помощью ограниченного словаря.

Для более полного учета словоизменения необходимо применение алгоритмов стемминга для выделения неизменяющейся части слова, а также проведение морфологического анализа с помощью специальных инструментов для определения грамматического класса, нормальной формы слов и т.д. Эти приемы количественного анализа основываются на совмещении признаков лексического и грамматического уровней.

1.2. Единицы грамматики текста

Вовлечение в практику автороведческой экспертизы, основанной на последовательном применении методов количественного анализа, единиц грамматики текста видится нам весьма эффективным, так как данный уровень текста генерируется подсознательно во время его создания и не контролируется автором направленно. Следствием этого является и значительная жанровая и дискурсивная независимость грамматических текстовых характеристик. Грамматика текста реализуется как функциональное соотнесение морфологических признаков слов и их синтаксических позиций, определяемых структурой синтаксического целого – высказывания (предложения).

В числе *морфологических признаков* наиболее часто для целей определения авторства используются распределения частей речи [17, 18, 50], реже – более полный набор грамматических классов и их сочетаний [19, 20]. Ряд работ выстраивается на комбинации данных признаков, A. Argamon-Engleson и M. Koppel, например, анализируют *N*-граммы частей речи как более информативную группу характеристик, чем одиночные части речи слов [21, 22].

Группа *синтаксических признаков* включает характеристики словосочетаний и предложений, способов их образования и употребления.

Основная сложность применения данного приема связана с необходимостью привлечения специальных лингвистических знаний программистом. Тексты, с которыми может работать программист, нуждаются в предварительной специальной грамматической разметке. С проблемами приходится сталкиваться уже на этапе определения границ предложения из-за неоднозначностей, существующих в языке. Омонимия, неполнота используемого морфологического словаря и отсутствие эффективных алгоритмов морфологического анализа обусловливают неоднозначность определения грамматического класса слова. Это приводит к невозможности автоматического выделения даже устойчивых синтаксических конструкций. По сути, задача может быть решена только с опорой на разработанные специальные корпуса текстов.

Вследствие этого в практике автороведческого анализа часто используются собственно формальные признаки предложения: его длина [23, 24, 25] и знаки пунктуации, репрезентирующие его смысловое и функциональное членение [26, 27, 50]. Очевидно, что чем сложнее предложение, тем оно длиннее и тем больше знаков препинания в нем используется; употребление вопросительных и восклицательных знаков может свидетельствовать об эмоциональной окраске речи и под.

К этой же группе признаков относится анализ синонимических конструкций [28]. Выбор того или иного слова при формировании фразы в соотнесении с особенностями ее построения носят индивидуальный характер и могут использоваться как признаки для идентификации автора текста.

Другой способ заключается в ограничении группы анализируемых признаков набором служебных (функциональных) слов, этот способ активно используется отечественными и зарубежными исследователями [28, 29, 30]. В данном случае идея состоит в том, что слова из этих подмножеств встречаются в любом фрагменте текста намного чаще, чем специфичные слова автора, и являются более информативным признаком. Использование данного способа позволяет их легко выделить из потока слов благодаря высокой встречаемости[1].

При работе с функциональными словами следует учитывать различие грамматических структур языков. Так, в английском, французском, болгарском и других языках большинство синтаксических связей выражается с помощью предлогов и других служебных слов, а также порядком слов в предложении. В отличие от них русский язык (а также немецкий, латинский,

[1] Некоторые исследователи используют модификацию данного метода – выявляются наиболее часто встречающиеся слова корпуса, которые помимо непосредственно функциональных слов языка включают наиболее часто употребляемые слова автора [26, 31, 32, 33, 50].

польский и др.) является флективным и синтетичным, следовательно, служебные слова играют меньшую роль в выражении синтаксических связей.

Альтернативным способом, учитывающим лексическую и синтаксическую информацию и анализирующим слово в контексте его окружения, является использование *N*-грамм слов [34, 35, 36]. При этом можно также учитывать позиции слов в предложении для построения дальнодействующих *N*-граммных моделей [37].

1.3. Идиосинкразические признаки

К особой группе характеристик текста, с опорой на который ведется авторoведческая экспертиза, осуществляемая специалистами-информатиками, относятся идиосинкразические признаки**:** орфографические и грамматические ошибки и другие текстовые аномалии, связанные с нарушением норм употребления единиц всех языковых уровней. Обычно подобные характеристики извлекаются программистом в предварительной работе с текстом с помощью словарей и средств проверки орфографии, грамматики, а также путем сравнения текста с эталоном.

Подход, основанный на использовании текстовых аномалий на всех лингвистических уровнях, предложен в работах [38, 39, 40]. Эксперт в области криминалистической лингвистики К. Часки (C.E. Chaski), в частности, утверждает, что выделенные таким образом значимые отличия, являются ключевыми для идентификации и пронизывают все уровни от фонетики и морфологии до синтаксиса, семантики и дискурса включительно, и подтверждает это серией экспериментов на собственном корпусе [27]. Точность при использовании в качестве признаков орфографических ошибок составила от 65 до 78%, ошибок пунктуации – 75%, грамматических ошибок, связанных с неправильным образованием морфологических форм слов, – от 72 до 92%. Данные подтверждаются также в работах [41].

Стоит отметить, что ошибки пунктуации, построения предложений, грамматические ошибки и т.д. активно используются отечественными экспертами-криминалистами при проведении авторoведческой экспертизы. Так, например, А.Ю. Комиссаров в своей диссертационной работе описывает методику идентификации исполнителя текста по орфографическим ошибкам [42]. Для оценки им предлагается использовать отношение количества совпавших в двух документах ошибок к общему количеству ошибок в исследуемом тексте. Если полученное значение превышает 65%, то выносится категорически положительное решение о тождестве исполнителя текста. При этом автор отмечает, что надежность методики снижается с повышением уровня орфографических навыков.

Однако следует сказать, что использование в качестве диагностических признаков пунктуационных и орфографических (а также в значительной степени и аномалий на других уровнях) ограничено текстами так называемой естественной письменной речи. Они утрачивают свою репрезентативность в текстах, подвергающихся редакторской и корректорской правке. Таким образом, широкий спектр художественных текстов не может атрибутироваться с опорой на данные признаки. С осторожностью данные приемы должны использоваться и при анализе текстов различных жанров интернет-коммуникации в связи с введением элементов автоматической правки текста, а также осмыслен-

ным коверканием слов и использованием интернет-жаргона (в данном случае эксперт имеет дело с сознательно моделируемым коммуникантом аспектом текстовой структуры, что, несомненно, должно учитываться).

2. Следующая группа признаков относится нами к разряду **собственно текстовых**, опосредствованно соотносимых с языковыми элементами.

2.1. Структура и графическое оформление текста

Структурные признаки в целом тяготеют к группе контентно-специфических.

В меньшей степени зависят от типа, канала коммуникации, в пределах которого порождаются тексты, такие структурные признаки, как разделение текста на фрагменты (главы, пункты, абзацы), наличие у фрагментов заголовков и их стиль.

Некоторые структурные признаки, выделяемые в автороведческой экспертизе, релевантны для текстов ограниченного набора дискурсов. Так, например, цитирование автором других источников и способы указания ссылок на них, как правило, выделяются и просчитываются в автороведческой экспертизе научных текстов. Следующие структурные признаки выделяются в автороведческой экспертизе текстов электронной коммуникации. Для многих видов текстов, веб-страниц, электронных презентаций, текстов, подготовленных для печати в издательских системах, структура и разметка играют важную роль при определении авторства [10, 11, 43]. Подобные системы позволяют контролировать форматирование, включая выбор параметров шрифта, оформление фигур, таблиц и анимации, цветовую гамму документа и его отдельных элементов.

Характеристики печатных документов, используемых в автороведческих экспертизах, могут включать количество пробелов, предшествующих знакам пунктуации, отступы, позиции табуляции и т.д.

Подобные признаки присущи текстам исходных кодов компьютерных программ [44]: оформление комментариев, конструкций условий и циклов, количество пробелов в отступах при переходе к следующему уровню вложенного кода и т.д.

Данные типы признаков являются сугубо формальными и достаточно легко выделяются в предварительной обработке анализируемого текстового массива.

Однако в последнее время данные признаки утрачивают актуальность, так как большинство современных сред разработки поддерживают автоматическое форматирование текста. Разметка гораздо чаще подвергается правке корректорами, редакторами или модераторами веб-сайтов перед публикацией, чем состав словаря, морфологические или синтаксические характеристики текста. Различные инструменты для создания документов одного формата обладают разными возможностями, которые могут отсутствовать в других инструментах. Перевод документа из одного формата в другой может привести к изменению его разметки. Стоит также учитывать знание автором конкретной среды создания документов и ее возможностей.

2.2. Метаданные документа

Другой плодотворной областью исследований является использование **метаданных** – дополнительных служебных данных, не относящихся непо-

средственно к содержанию текста. Учитывая современный уровень развития судебной компьютерной экспертизы, отметим, что данные характеристики являются наиболее проработанными в аспекте идентификации автора. Например, заголовок электронного письма, относящегося к делу, содержит адрес отправителя и другую информацию. Даже в том случае, когда информация скрыта намеренно, эксперт может извлечь некоторые полезные для себя данные, которых будет достаточно для определения следующего звена в цепи передачи сообщения, на основе которого можно продолжить анализ.

Подобные метаданные обычно присутствуют и в составных документах. Так, всем известный редактор Microsoft Word включает в документ информацию о дате создания, версии программы, имени пользователя, истории изменений, имени оригинального файла, авторе правок и т.д. [48, 49]. С практической точки зрения наличие или отсутствие имени автора в таком документе при проверке студенческой работы на плагиат порой бывает намного полезнее, чем статистический анализ документа. А информация о том, что документ был создан в Microsoft Word в 10 часов утра, может стать существенным доводом против предположения о том, что он написан ярым сторонником операционных систем Linux. Возможно также целенаправленное добавление метаданных, содержащих информацию об авторе, в документ при помощи методов стеганографии.

Однако не стоит преувеличивать информативность метаданных. Они должны рассматриваться с тем же уровнем доверия, что и форматирование и разметка текста.

Заключение

1. Анализ существующих в настоящее время работ по идентификации автора текста, осуществляемых с привлечением количественных методов анализа, разрабатываемых в теории информации и интеллектуальном анализе данных, свидетельствует о значительной эффективности использования в качестве релевантных для современных индоевропейских языков таких признаков, как биграммы и триграммы символов и слов, функциональные (служебные) слова, распределение слов по частям речи, наиболее частотные слова, знаки пунктуации, распределение длины слова и длины предложения. Как видим, это прежде всего *собственно языковые единицы практически всех уровней языковой системы*, признаки как сугубо формальные, так и формально-семантические.

2. К числу наиболее часто используемых в качестве диагностических параметров в авторoведческой экспертизе относятся *признаки формального членения текста*, в составе последних – признаки низших уровней языковой системы, что мотивируется как объективными факторами (это слабо- или нерефлексируемые говорящим аспекты порождения речи), так и субъективными – простотой их вычленения, не требующей специальной лингвистической обработки текста, относительно незначительным объемом привлекаемых текстов для установления статистических закономерностей.

3. Относительно меньшая частотность и результативность использования в качестве диагностических параметров языковых единиц *формально-смыслового* членения речи определяется взаимодействием разного рода причин:

– единицы высших – лексического и лексико-синтаксического – срезов текста в большей степени по сравнению с единицами формального структурирования могут быть рефлексируемыми и, следовательно, намеренно моделируемыми;

– при работе с единицами грамматического уровня требуется предварительная квалифицированная разметка текстов, создание лингвистически размеченных корпусов текстов;

– при анализе единиц высших уровней языковой системы для установления статистических закономерностей необходимы более объемные корпусы текстов, нежели при выявлении статистики использования единиц символьного формального уровня.

4. В составе выделяемых в практике автороведческой экспертизы следует различать признаки текста, которые могут выступать в качестве диагностирующих только в определенных типах текста (идиосинкразические признаки не работают в текстах, подвергающихся редакторской и корректорской правке; использование метаданных применимо к компьютерным текстам и т.д.), и признаки контекстно независимые.

5. Вопрос о составе контекстно независимых признаков текста нуждается в серьезной многоаспектной проработке.

Во-первых, многие из обсуждавшихся в данной статье результатов получены при исследовании текстов английского и ряда других индоевропейских языков. В связи с этим актуализируется вопрос о релевантности данных признаков для русскоязычных текстов, особенно остро этот вопрос стоит при обращении к единицам грамматического уровня. В частности, особенностью русского языка в сравнении с английским, на материале которого получена большая часть результатов, является его флективность, а следовательно, и более сложное словообразование, высокая степень морфологической и синтаксической омонимии, что в совокупности обусловливает необходимость дополнительной проверки функционирования данных признаков в качестве диагностирующих. В целом в настоящее время недостаточно внимания уделяется идентификации автора на основе *сочетаний признаков русскоязычного текста*.

Во-вторых, автороведческая экспертиза русскоязычных текстов формировалась и апробировала методологию преимущественно на материале художественных текстов. В настоящее время спектр текстов, вовлекаемых в автороведческую экспертизу, существенно расширяется, что актуализирует вопрос и о составе максимально контекстно устойчивых признаков текста, и признаков, релевантных при анализе текстов определенного типа.

В-третьих, в состав контекстно независимых признаков включаются единицы морфологического и синтаксического членения текста, эффективное использование которых возможно только на основе предварительной лингвистической разметки текста, что актуализирует необходимость создания специализированных корпусов русскоязычных текстов, следовательно, более последовательного и направленного взаимодействия лингвистов и специалистов в области теории информации и интеллектуального анализа данных.

Литература

1. *Rudman J.* The state of authorship attribution studies: Some problems and solutions // Computers and the Humanities. – 1998. – Vol. 31. – P. 351–365.

2. *Хмелев Д.В.* Распознавание автора текста с использованием цепей А.А. Маркова // Вестн. МГУ. – Сер. 9: Филология. – 2000. – № 2. – С. 115–126.

3. *Шевелев О.Г.* Методы автоматической классификации текстов на естественном языке: учеб. пособие. – Томск: ТМЛ-Пресс, 2007. – 144 с.

4. *Benedetto D.* Language Trees and Zipping / D. Benedetto, E. Caglioti, V. Loreto // Phys. Rev. Lett. – 2002. – Vol. 88, №4. – P. 487-490.

5. *Hoorn J.* Neural network identification of poets using letter sequences / J. Hoorn, S. Frank, W. Kowalczyk et al. // Literary and Linguistic Computing. – 1999. – Vol. 14, № 3. – P. 311–338.

6. *Kjell B.* Authorship attribution of text samples using neural networks and Bayesian classifiers / B. Kjell // IEEE International Conference on Systems, Man and Cybernetics. San Antonio, TX, 1994.

7. *Kjell B.* Authorship determination using letter pair frequencies with neural network classifiers // Literary and Linguistic Computing. – 1994. – Vol. 9, № 2. – P. 119–124.

8. *Peng F.* Augumenting Naive Bayes Text Classifier with Statistical Language Models / F. Peng, D. Schuurmans, S. Wang // Information Retrieval. – 2004. – Vol. 7, № 3–4. – P. 317–345.

9. *Peng F.* Language independent authorship attribution using character level language models / F. Peng, D. Schuurmans, S. Wang et al. // Proceedings of the 10th conference on European chapter of the ACL. – 2003. – Vol. 1. – P. 267–274.

10. D*e Vel O.* Mining e-mail content for author identification forensics / O. De Vel, A. Anderson, M. Corney et al. // ACM SIGMOD. – NY : ACM, 2001. – Rec. 30. – № 4. – P. 55–64.

11. *Zheng R.* A framework for authorship analysis of online messages: Writing-style features and techniques / R. Zheng, J. Li, Z. Huang et al. // Journal of the American Society for Information Science and Technology. – 2006. – Vol. 57, № 3. – P. 378–393.

12. *Yule G.U.* The Statistical Study of Literary Vocabulary. – Cambridge University Press, 1944. – 306 p.

13. *Tweedie F.J.* How Variable may a Constant be? Measures of Lexical Richness in Perspective / F.J. Tweedie, H. Baayen // Computers and the Humanities. – Springer, 1998. – Vol. 32, № 5. – P. 323–352.

14. *Simpson E.H.* Measurement of Diversity / E.H. Simpson // Nature. – Macmillan Publishers Ltd, 1949. – № 163. – P. 688.

15. *Хрулев О.* Определение автора по тексту на естественном языке [Электронный ресурс]. – Режим доступа: www.geshtalt.ru/ psycholingvist_author.php.

16. *Juola P.* What can we do with small corpora?: Document categorization via cross-entropy [Electronic resource] // Proceedings of an Interdisciplinary Workshop on Similarity and Categorization, Edinburgh, UK. – 1997. – URL: http://www.mathcs.duq.edu/~juola/papers.d/identification.ps.

17. *Argamon S.* Routing documents according to style [Electronic resource] / S. Argamon, M. Koppel, G. Avneri // Proceedings of the 1st International Workshop on Innovative Information. – 1998. – URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.52.688&rep=rep1&type=pdf.

18. *Baayen R.H.* Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution / R.H. Baayen, H.V. Halteren, F.J. Tweedie // Literary and Linguistic Computing. – 1996. – Vol. 11. – P. 121–131.

19. *Кукушкина О.В.* Определение авторства текста с использованием буквенной и грамматической информации / О.В. Кукушкина, А.А. Поликарпов, Д.В. Хмелев // Проблемы передачи информации. – 2001. – Т. 37, вып.2. – С. 96–109.

20. *Stamatatos E.* Computer-based authorship attribution without lexical measures / E. Stamatatos, N. Fakotakis, G. Kokkinakis // Computers and the Humanities. – 2001. – Vol. 35, № 2. – P. 193–214.

21. *Argamon-Engleson A.* Style-based text categorization: What newspaper am I reading / A. Argamon-Engleson, M. Koppel, G. Avneri // Proceedings of the AAAI Workshop of Learning for Text Categorization. – 1998. – P. 1–4.

22. *Koppel M.* Automatically categorizing written texts by author gender / M. Koppel, S. Argamon, A.R. Shimoni // Literary and Linguistic Computing. – 2002. – Vol. 17, № 4. – P. 401–412.

23. *Argamon S.* Style mining of electronic messages for multiple authorship discrimination: first results / S. Argamon, M. Saric, S.S. Stein // Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – NY : ACM, 2003. – P. 475–480.

24. *Kruh L.* A basic probe of the Beale cipher as a bamboozlement. P. 1 / L. Kruh // Cryptologia. – 1982. – Vol. 6, № 4. – P. 378–382.

25. *Morton A.Q.* Literary Detection: How to Prove Authorship and Fraud In Literature and Documents / A.Q. Morton. – New York : Scribner's, 1978. – 221 p.

26. *Baayen R.H.* An experiment in authorship attribution / R.H. Baayen, H.V. Halteren, A. Neijt et al. // Proceedings of JADT 2002. – Universit´e de Rennes, St. Malo, 2002. – P. 29–37.

27. *Chaski C.E.* Empirical evaluations of language-based author identification // Forensic Linguistics. – 2001. – Vol. 8, № 1. – P. 1–65.

28. *Mosteller F.* Inference and Disputed Authorship: The Federalist / F. Mosteller, D.L. Wallace. – Reading, MA : Addison-Wesley, 1964 – 287 p.

29. *Сысуев В.* Проект «Пси Офис» [Электронный ресурс]. – 2002. – Режим доступа: http://psy-two.narod.ru/embedded.html.

30. *Green T.R.G.* The necessity of syntax markers: Two experiments with artificial languages / T.R.G. Green // Journal of Verbal Learning and Verbal Behavior. – 1979. – Vol. 18. – P. 481–96.

31. *Burrows J.* "An ocean where each kind...": Statistical analysis and some major determinants of literary style / J.F. Burrows // Computers and the Humanities. – 1989. – Vol. 23, №4. – P. 309–321.

32. *Halteren H.* New machine learning methods demonstrate the existence of a human stylome / H. Halteren, R.H. Baayen, F. Tweedie et al. // Journal of Quantitative Linguistics. – 2005. – Vol. 12, № 1. – P. 65–77.

33. *Hoover D.L.* Delta prime? / D.L. Hoover // Literary and Linguistic Computing. – 2004. – Vol. 19, № 4. – P. 477–495.

34. *Nowson S.* Identifying more bloggers: Towards large scale personality classifiation of personal weblogs [Electronic resource] / S. Nowson, J. Oberlander. – 2007. – URL: http:// nowson.com/ papers/NowOberICWSM07.pdf

35. *Oakes M.* Text categorization: Automatic discrimination between US and UK English using the chi-square text and high ratio pairs / M. Oakes // Research in Language. – 2003. – Vol. 1. P. 143–156.

36. *Yu B.* English usage comparison between native and non-native english speakers in academic writing [Electronic resource] / B. Yu, Q. Mei, C. Zhai // Proceedings of ACH/ALLC. – 2005. – URL: http://mustard.tapor.uvic.ca/cocoon/ach_abstracts/xq /xhtml.xq?id=207.

37. *Эл Л.С.* Вывод и оценка параметров дальнодействующей триграммной модели языка / Л.С. Эл, С.В. Протасов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Междунар. конф. «Диалог», Бекасово, 4–8 июня 2008 г. – М., 2008. – Вып. 7 (14). – С. 443–448.

38. *Chaski C.E.* Multilingual Forensic Author Identification through N-Gram Analysis [Electronic resource] / C.E. Chaski // Proceedings of the 8th Biennial Conference on Forensic Linguistics/Language and Law, July 2007, Seattle, WA. – 2007. – URL: http:// www. allacademic. com/meta/p177064_index.html

39. *Foster D.* Author Unknown: Adventures of a Literary Detective / D. Foster. – London : Owl Books, 2000. – 320 p.

40. *Koppel M.* Exploiting stylistic idiosyncrasies for authorship attribution / M. Koppel, J. Schler // Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico, 2003. – 2003. – P. 69–72.

41. *Grant T.* Identifying reliable, valid markers of authorship: A reponse to Chaski / T. Grant, K. Baker // Forensic Linguistics. – 2001. – Vol. 8, № 1. – P. 66–79.

42. *Комиссаров А.Ю.* Криминалистическое исследование письменной речи с использованием ЭВМ: дис. … канд. юрид. наук. – М., 2001. – 225 с.

43. *Abbasi A.* Identification and comparison of extremist-group Web forum messages using authorship analysis / A. Abbasi, H. Chen // IEEE Intelligent Systems. – 2005. – Vol. 20, № 5. – P. 67–75.

44. *Oman W.P.* Programming style authorship analysis / W.P. Oman, R.C. Cook // Proceedings of the 17th Annual ACM Computer Science Conference. – NY, 1989. – P. 320–326.

45. *Elliot W.* Was the Earl of Oxford the true Shakespeare? / W. Elliot, R. Valenza // Notes and Queries. – 1991. – Vol. 38. – P. 501–506.

46. *Martindale C.* On the utility of content analysis in author attribution: The federalist / C. Martindale, D. McKenzie // Computers and the Humanities. – 1995. – Vol. 29. – P. 259–270.

47. *Diederich J.* Authorship attribution with support vector machines / J. Diederich, J. Kindermann, E. Leopold // Applied Intelligence. – Springer Netherlands, 2003. – Vol. 19, №1–2. – P. 109–123.

48. *Migletz J.* Automated metadata extraction [Electronic resource] / J. Migletz. – 2008. – URL: http://simson.net/clips/students/ 08Jun_Migletz.pdf.

49. The risks of metadata and hidden information [Electronic resource]. – 2007. – URL: http: // www.stg.srs.com/eds/docdet/archive/BitformFortune100Study.pdf.

50. *Романов А.С., Шелупанов А.А., Мещеряков Р.В.* Разработка и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста. – Томск: В-Спектр, 2011. – 188 с.