*Вестник Томского государственного университета. 2014. № 379. С. 17–23*

УДК 81-139

*Г.В. Напреенко*

# ИДЕНТИФИКАЦИЯ ТЕКСТА ПО ЕГО АВТОРСКОЙ ПРИНАДЛЕЖНОСТИ НА ЛЕКСИЧЕСКОМ УРОВНЕ (ФОРМАЛЬНО-КОЛИЧЕСТВЕННАЯ МОДЕЛЬ)

Статья посвящена разработке формально-количественной модели идентификации текста на лексическом уровне. Текст (как персонотекст – носитель образа его автора) содержит проявление свойств языковой личности – языковые характеристики, выраженные, в том числе, количественными показателями. Формально-количественное описание лексических единиц в той или иной степени способно отразить авторскую принадлежность текста. Статья направлена на проверку одного из таких способов квалификации авторской речевой индивидуальности.

**Ключевые слова:** идентификационная лингвистика; лингвоперсонология; персонотекст; формально-количественная модель идентификации текста; квантитативная лингвистика; юрислингвистика.

**1. Идентификация текста по воплощенным в нем языковым особенностям автора как лингвистическая проблема.** Речь и ее продукты – одни из основных средств идентификации личности, так как с их помощью можно проникнуть в сферу внутреннего мира человека. Способность языковой личности (ЯЛ) порождать и воспринимать речь среди прочего реализуется за счет устойчивых речевых образцов, которые являются бессознательными маркерами личности. В предлагаемой статье в данном аспекте рассматриваются частотные характеристики лексических единиц, составляющих текст.

Феномен идентификации реализуется в разных научных дисциплинах. В логике и философии – это установление тождественности неизвестного объекта известному на основании совпадения признаков. Например, Н.Г. Кожина в статье «Аксиологический подход к исследованию идентификации современной личности» [1] в рамках социально-философского анализа рассматривает вопрос, связанный с ценностным восприятием современной личности, выступающим в качестве регулятора процесса аксиологической идентификации. Идентификация в криминалистике представляет собой установление тождества объектов или личности по совокупности их идентификационных признаков. В России учение о криминалистической идентификации было разработано отечественным ученым С.М. Потаповым.

Феномен идентификации стал объектом исследований и в лингвистике, где осуществляются процесс сопоставления признаков, определение сходства или различия объектов. Научную основу идентификационной лингвистики составляет система знаний о закономерностях речевого поведения человека, обусловливающих индивидуальность, устойчивость, вариационность письменной речи. Речевое поведение личности в целом изучает лингвоперсонология как научная дисциплина.

В лингвистике может осуществляться идентификация разных объектов: идентификация собственно текста безотносительно к личности, идентификация текста как ***персонотекста***, идентификация ЯЛ по тексту. При идентификации текста вне обращения к личности, его написавшей, мы имеем дело с таким направлением, как синхронная текстология, важной задачей которой является атрибуция текста. Один из способов атрибуции – стилометрия – исследование стилистики, включающее статистический анализ.

Идентификация ЯЛ представляет собой разновидность такого лингвоперсонологического исследования, как ***портретирование*** ЯЛ. На основе выявленных параметров проводится моделирование ЯЛ, создается портрет конкретной ЯЛ, принадлежащей определенному типу ЯЛ. При портретировании ЯЛ учитываются индивидуальные языковые предпочтения. По определению С.В. Леорды, «речевой портрет – это воплощенная в речи языковая личность» [2]. В портретировании конкретной ЯЛ отражаются различные характеристики личности (гендерные, возрастные, психологические, социальные и др.). Например, так осуществляются портретирование лингвосоционической ЯЛ [3], описание речевого портрета студента-филолога [4] и т.п. При проведении автороведческой экспертизы существуют не только идентификационные, но и диагностические задачи. При так называемой идентификации ЯЛ решаются задачи диагностического типа: например, место проживания автора текста, профессия, пол, возраст, социальное положение и пр.

При идентификации персонотекста объектом становится текст в аспекте проявления в нем свойств ЯЛ. В учебном пособии «Лингвоперсонология и личностно-ориентированное обучение языку» сформулирована лингвоперсонологическая гипотеза: «Язык устроен так, а не иначе еще и потому, что так устроена персонологическая сфера» [5. С. 24]. Каждая личность осуществляет выбор из разнообразия средств и способов для выражения разнообразных коммуникативных намерений и делает этот выбор «в соответствии со своими внутренними личностными потребностями и способностями» [5. С. 12]. Персонотекст как аспектуальная разновидность речевого произведения является репрезентацией языкового потенциала личности, который реализуется автором в тексте через языковые характеристики. Выбор из бесконечного множества формирует индивидуальные предпочтения ЯЛ – ее ***идиостиль***. За результатами идиостиля стоят варианты качества языковой способности, которые в дальнейшем обрабатываются личной речевой практикой и обучением. При идентификации персонотекста задача лингвиста – извлечь из текста и характеризовать варианты способностей ЯЛ, что позволяет в дальнейшем при сопоставлении идентифицировать персонотексты. Таким образом, мы имеем дело с лингвоперсонной идентификацией.

**2. О методах идентификации текста как персонотекста.** Исследователями проблемы идентификации текстов были предложены различные методы определения авторства на орфографическом, лексико-фразеологическом, пунктуационном, синтаксическом,

стилистическом уровнях. Например, нормативно-стилистический анализ текстов проводит А.В. Морозов в авторoведческой экспертизе текста договора [6]. Обычно лингвистическая идентификация проводится в совокупности на нескольких уровнях [7]. Мы предлагаем далее наш вариант методики идентификации персонотекста, осуществляемой на лексическом уровне. Заметим при этом, что лексический уровень текста может быть исследован с точки зрения проявления в нем свойств ЯЛ под разным углом зрения (например, семантическом, стилистическом).

В статье представлен анализ лексического уровня с точки зрения ***частоты*** употребления лексем, составляющих персонотекст, – формально-количественная модель. Персонотекст рассматривается нами сквозь призму частотности составляющих его лексических единиц. «Частотность слова, отмечаемая в частотном словаре, формируется лексическим функционированием слова» [8]. Частотность выступает как особая характеристика слова. Качественная интерпретация количественных характеристик лексем дает определенные данные по распределению лексем в речевом произведении. Сопоставление словников по каждому тексту в данном ракурсе отражает разность авторских предпочтений ЯЛ.

Таким образом, предлагаемое исследование подключено к парадигме квантитативной лингвистики. В настоящее время наблюдается повышенный интерес к формализованным методам анализа текстовой информации на основе слабо контролируемых человеком характеристик текста. Подобные методы возможны в связи с существующей в языке функциональной зависимостью: слова в тексте организованы в соответствии с частотой их появления в тексте, каждому из них присвоен номер ранга и соответствующая частота, которые обратно пропорциональны в списках частотности, согласно закону Ципфа. Функциональные зависимости отражают не только строй языка, они проявляются в тексте индивидуально: в них отражаются способности ЯЛ. Интерпретируя этот закон, исследователи предлагали разные математические модели. «Математика с ее неисчерпаемыми возможностями должна дать основания для более глубокого проникновения в “механизм” языка и “рационального” описания открываемых ею закономерностей» [9. С. 6]. Языковые характеристики текста должны быть настолько формализованы, чтобы быть готовыми к обработке компьютерными программами. Формализация лексического состава языка в широком смысле представляет интерсемиотический перевод, т.е. интерпретацию знаков одной семиотической системы знаками другой. В применении к количественным методам идентификации персонотекста – это перевод лексем текста (вербальных знаков) на количественные характеристики (невербальные знаки), такие как порядок слова в частотном словаре (ранг) и частота его употребления. Эти количественные данные «заменяют» лексические единицы и выступают в роли идентификаторов текста. Формализация способствует автоматизации процесса исследования.

Множество современных исследований посвящено созданию универсальной методики атрибуции текстов и поиску путей совершенствования лингвистической авторoведческой экспертизы. Исследователи отмечают следующие актуальные проблемы и задачи лингвистической идентификации, которые необходимо решить: а) субъективный характер атрибуции текстов; б) объективизация и автоматизация методик; в) создание универсальной методики, т.е. ее применение относительно текстов разных стилей и разного объема; г) возможность и значимость использования математического анализа параметров текста; д) критерий отбора параметров, которые можно считать идентифицирующими, их количество. Далее приведем несколько примеров такого рода исследований.

А.Ю. Хоменко в статье «Алгоритм автоматизации идентификации автора письменного речевого произведения для судебного авторoведения» [10] предлагает методику атрибуции текстов, которая показала свою достоверность только на 50%. В методике сделана попытка совмещения нескольких методологий: анализ ЯЛ автора текста и стилометрического исследования текста. Автор рассматривает условия и приводит рекомендации по совершенствованию методики. В этом отношении статья вносит вклад в исследования по лингвокриминалистике, лингвоперсонологии. Результаты статьи наталкивают на мысль о том, что создание универсальной методики идентификации (атрибуции текстов разных стилей и разного объема) в случае применения стилометрии пока не представляется возможным. Существующие методики релевантны для текстов определенного объема и функционального стиля.

Е.С. Родионова в статье «Лингвистические методы атрибуции и датировки литературных произведений (К проблеме «Мольер – Корнель»)» [11] применяет методы статистики на синтаксическом уровне. Исследуются такие параметры: число элементарных предложений, число сочинительных предложений, число подлежащих и др.

З.И. Резанова, А.С. Романов и Р.В. Мещеряков в статье «О выборе признаков текста, релевантных в авторoведческой экспертной деятельности» дают совокупное представление о существующих параметрах, которые могут выполнять идентифицирующую функцию. Рассматриваются как чисто формальные признаки, так и формально-семантические, оцениваются их релевантность, степень апробации, выделяются слабые стороны. Авторами отмечается существенная проблема: многие методики (в том числе квантитативные) не предназначены для русского языка, именно поэтому есть необходимость создания специальной методики для русского языка или апробации существующих методик для русского языка. Авторы статьи оценивают особенности словаря автора как параметр идентификации таким образом: «Ярких характерных особенностей лексикона у текста, равно как и у автора, может и не быть. 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 справедливо по отношению к единицам всех грамматических классов) и, вследствие этого, не может выступать в качестве надежного идентификатора при вовлечении в экспертизу разножан-

ровых текстов одного автора» [12. С. 6]. Со сказанным согласимся в некоторых пунктах: например, методы идентификации на данный момент, действительно, субъективны; универсальной методики не существует, поэтому наблюдается зависимость идентификации текстов от жанра, к которому они относятся. Хотелось бы отметить следующее. Намеренное моделирование можно распознать с помощью специальных методов, кроме того, есть некоторые слои языка, которые не всегда поддаются сознательному искажению. Наличие или отсутствие «ярких характерных особенностей» у текста или автора, безусловно, зависимо от жанра речевого произведения, при этом в обозрении лингвоперсонологии каждая ЯЛ обладает совокупностью определенных языковых особенностей, реализующихся в речевом произведении.

**3. Лексико-квантитативная методика идентификации персонотекста.** Мы предлагаем формально-количественную модель идентификации персонотекста. В работе ставится задача реализации следующей логики: универсальность – формализация (вслед за ней – автоматизация) – объективность. Объективность касается как результатов, так и процесса исследования. Перспектива исследования – выявление особенностей текста, обусловленных особенностями языковой способности автора.

Выдвигается ***гипотеза***: текст содержит индивидуальные лексические характеристики, которые, выраженные контрастными по отношению друг к другу количественными данными, могут являться идентификаторами персонотекста. Наиболее частотные единицы русского языка содержатся в любом тексте и являются высокочастотными для любого текста. Выпадение их в частотном словаре конкретного текста из этого статуса или, напротив, появление в списке высокочастотных единиц дает основания для утверждения о том, что данный текст нетиповой. Подобная специфичность, если она повторяется в разных персонотекстах, дает основания для их идентификации. Языковые характеристики понимаются как лингвоперсонологические, что позволяет атрибутировать тексты по принадлежности одному автору. В настоящем исследовании под текстом понимается массив дневниковых записей без разрыва их на отдельные элементы.

Интернет-дневники трактуются в настоящей статье как персонотексты – объекты, подвергающиеся идентификации. В структуре работы они являются ***материалом исследования***. ***Объектом*** становятся тексты 4 интернет-дневников и соответственно 4 частотных словарей. Поясним следующее: изначально ясно, что авторов всего два. В качестве эксперимента автором статьи собрано два интернет-дневника, каждый из которых разделен на две части. Названия интернет-дневникам и словникам даны условно и произвольно, тем самым придана анонимность: А, Б, В, Г. В тексте А – 26 805, Б – 15 632, В – 28 992, Г – 19 660 словоупотреблений. В каждом словнике исследуются 100 самых частотных слов текста. Задача – установить, скольким авторам принадлежат персонотексты, атрибутировать каждый конкретному автору.

Лексический уровень текстов с помощью компьютерной программы SimWordSorter представляется в форме словников, тем самым производится интерсемиотический перевод. При составлении словников реализуется статистический анализ. Словники предстают, прежде всего, как способ исследования проявлений свойств ЯЛ в тексте, и статистический метод способствует этому. Словник в нашей схеме идентификации – список слов по убывающей частоте, в котором единицей счета является лексема. Лексема, прежде всего, потому, что мы поставили целью разработать методику идентификации персонотекста на лексическом уровне. Один из этапов работы над словниками – преобразование словоформ (токенов) в лексемы. При токенизации мы имеем дело с «машинным» пониманием словоформ и лексем, когда исключается различие омонимов, многозначных слов. Рабочее определение слова – «набор букв между двумя пробелами, как это принято в машинном переводе» [9. С. 37], или «токен». Для анализа составляются четыре частотно-сопоставительные таблицы. Каждая таблица содержит два словника, в каждом перечислены сто самых частотных лексем и указан ранг каждой лексемы.

Далее следует анализ частотно-сопоставительных таблиц, обозначенный как формально-количественная модель идентификации персонотекстов: вариант идентификации на базе слов-идентификаторов с высокой частотностью, основанный на анализе предпочтений авторов в использовании определенных слов в своей речи. В этом варианте идентификации актуализируется квантитативный аспект, когда интересны анализ количественной наполненности словников и их сопоставление. Исследование делится на два этапа:

**1 этап.** За гипотезу этой части эксперимента принимаем следующее: в словниках, принадлежащих одному автору, лексемы расположены приблизительно на одном ранге / в одной группе. У текстов, написанных другим автором, ранги лексем имеют иной порядок. ЯЛ проявляется в предпочтении использования некоторых слов в большей степени, чем других.

Предлагаемая модель представлена двумя разными способами, которые базируются на разном представлении о ранге:

1) сто самых частотных слов в каждом словнике разделено на 10 групп. В этом случае осуществляется сопоставление лексем между группами в пределах сотни. Под рангом понимается ***группа слов*** (Rг);

2) сопоставление лексем в пределах всей сотни без разделения на группы. В этом случае актуален ***ранг каждого слова*** (R).

За основу метода примем идеальное условие, как, например, «материальная точка» (идеальная модель) или «вакуум» (идеальное условие) в физике. Заключается оно в следующем: существует некоторое идеальное распределение, при котором:

1) у одного автора в разных текстах одно и то же слово преимущественно располагается в пределах одной группы или соседних группах;

2) у одного автора ранг определенного слова совпадает в разных текстах, при этом допустимая максимальная разность – 10 рангов (несмотря на то что в данном способе слова анализируются без разделения на группы по 10 слов, условно примем $\Delta R \leq 10$ в качестве границы, отражающей принадлежность одному или разным авторам).

**Рабочие формулы.**

Для *первого* способа ($R_г$):

$$\Delta R_г = R1_г - R2_г,$$

где $\Delta R_г$ – разность рангов; $R1_г$ – ранг группы первого текста в частотно-сопоставительной таблице; $R2_г$ – ранг группы второго текста.

Если $\Delta R_г = 0, < 2$, т.е. слова находятся в одной или соседних группах, то у персонотекстов один автор.

Если $\Delta R_г \geq 2$, т.е. слова находятся в разных десятках и эта разность равна 2 или больше, то авторы у персонотекстов разные.

Для *второго* способа (R):

$$\Delta R = R1 - R2,$$

где ΔR – разность; R1 – ранг лексемы первого текста в частотно-сопоставительной таблице; R2 – ранг лексемы второго текста.

Если $\Delta R = 0; \leq 10$, т.е. слова находятся на одном ранге либо разность ранга лексемы в разных текстах не более 10, то автор один.

Если $\Delta R > 10$, разница между рангами контрастная. Эта разница отражает принадлежность персонотекстов разным авторам.

После расчетов подсчитывается сумма двух типов лексем: *тип а* – лексемы, подтверждающие то, что тексты написаны одним автором, *тип б* – лексемы, подтверждающие написание текстов разными авторами. По результатам сопоставления рангов каждой частотно-сопоставительной таблицы строятся диаграммы, делаются выводы по количеству лексем двух типов.

Анализ проводится на следующих уровнях по каждому способу:

I. Сопоставление частотных таблиц предположительно разных авторов.

Группа № 1. Частотно-сопоставительная таблица № 1. Словники А – Б.

Группа № 2. Частотно-сопоставительная таблица № 2. Словники В – Г.

II. Сопоставление частотных таблиц предположительно одного автора.

Группа № 1. Частотно-сопоставительная таблица № 3. Словники А – В.

Группа № 2. Частотно-сопоставительная таблица № 4. Словники Б – Г.

III. Сопоставление полученных данных с «абсолютным» показателем – частотным словарем русского языка.

Приведем пример анализа I. Сопоставление частотных таблиц предположительно разных авторов, группа № 1.

Задача: сопоставить словники в рамках частотно-сопоставительных таблиц № 1 и 2 с целью выявить, как одна лексема в разных текстах не столько гипотетически используется разное количество раз, сколько какой ей принадлежит ранг в разных словниках. Если лексема в двух словниках имеет контрастный ранг, это говорит о разности индивидуальных языковых предпочтений авторов. Контрастность определяется указанными выше способами анализа слов. Показатель контрастности может являться критерием идентификации персонотекстов в случае, когда большинство лексем последовательно подтверждают разность авторских языковых способностей. Указывать количество употреблений слова при данном варианте идентификации нет необходимости (как важно это делать при синонимическом способе идентификации), так как актуально исключительно понятие ранга.

Первый способ (единица – ранг группы): $\Delta R_г = R1_г - R2_г$ (рис. 1).

Частотно-сопоставительная таблица № 1 (А – Б).

Все подсчеты ведутся подобным образом (в следующих пунктах будут указаны только выводы):

1. Лексема «я». $\Delta R_г = 1–1 = 0 \rightarrow$ один автор.
2. «как». $\Delta R_г = 1–1 = 0 \rightarrow$ один автор.
3. «но». $\Delta R_г = 1–2 = 1 < 2 \rightarrow$ один автор.
4. «по». $\Delta R_г = 1–3 = 2 = 2 \rightarrow$ разные авторы.
5. «так». $\Delta R_г = 2 – X = X \rightarrow$ разные авторы (X обозначает, что слова нет в анализируемой десятке групп второго текста).
6. «кто». $\Delta R_г = 2–6 = 4 > 2 \rightarrow$ разные авторы.
7. «до». $\Delta R_г = 4–7 = 3 > 2 \rightarrow$ разные авторы.
8. «ну». $\Delta R_г = 5–6 = 1 < 2 \rightarrow$ один автор.
9. «в». $\Delta R_г = 4–7 = 3 > 2 \rightarrow$ разные авторы.
10. «еще». $\Delta R_г = 7–2 = 5 > 2 \rightarrow$ разные авторы.

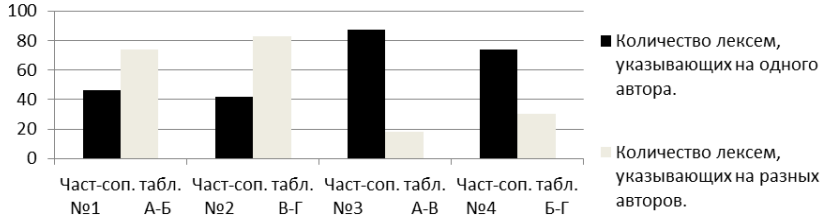


Рис. 1. Общее количество лексем каждого пункта анализа, полученное первым способом

Словники сгруппировались таким образом. Первые два объекта в диаграмме показывают принадлежность персонотекстов разным авторам (лексем *типа б* в 2 раза больше лексем *типа а*). Вторые два объекта – одному автору.

Количественное выражение диаграммы:

Частотно-сопоставительная таблица № 1 А – Б: количество лексем *типа а* – 46, *типа б* – 74; № 2 В – Г: *типа а* – 42, *типа б* – 83; № 3 А – В: *типа а* – 87, *типа б* – 18; № 4 Б – Г: *типа а* – 74, *типа б* – 30.

Второй способ (единица – ранг каждого слова): $\Delta R = R1 - R2$ (рис. 2).

Количественное выражение диаграммы:

Частотно-сопоставительная таблица № 1 А – Б: количество лексем *типа а* – 28, *типа б* – 91; № 2 В – Г: *тип а* – 32, *тип б* – 93; № 3 А – В: *тип а* – 70, *тип б* – 35; № 4 Б – Г: *тип а* – 46, *тип б* –58.

По второму способу варианта идентификации персонотекста выделяется принадлежность разным авторам в двух словниках. Третий столбец показывает принадлежность текстов А и В одному автору. Столбец № 4 не дает определенных выводов, однако приближает к выводу о том, что тексты Б и Г написаны одним автором. Однозначно заключить невозможно, так как количество лексем двух типов примерно одинаково.

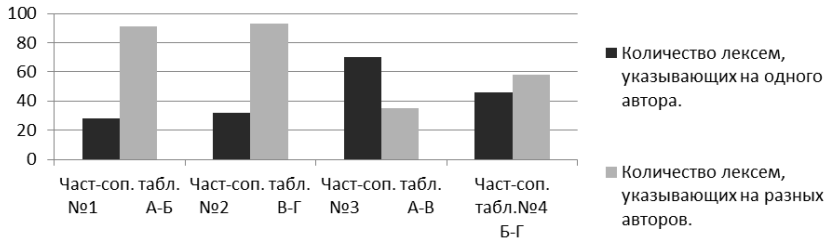


Рис. 2. Общее количество лексем каждого пункта анализа, полученное вторым способом

Оба способа показали возможность использования их при атрибуции текстов. Обязательным является учет описанных условий (жанр текстов, примерное количество словоупотреблений), так как иные условия не проверялись.

**2-й этап.** ***Сопоставление с частотным словарем русского языка*** как с «абсолютным» показателем распределения слов по частотности, которое приведено в современной версии частотного словаря русского языка О.Н. Ляшевской, С.А. Шарова [13]. Выбор этого словаря обусловлен следующим. Основным источником информации о частоте русских слов ранее был словарь русского языка под ред. Л.Н. Засориной (1997 г.). По современным стандартам корпус, на основе которого подсчитана частота слов в этом словаре, мал; список существенно устарел: он соответствует частоте использования слов в период с 1920-х до 1960-х гг. «В результате корпус включает большое число идеологических источников. Слова *советский*, *товарищ* входят в первую сотню русских слов наряду со служебными словами (они встречаются чаще слов *где, здесь*)» [13. С. 5] и т.п. Современная версия частотного словаря существует в электронной форме, что вписывается в задачи нашего исследования.

Предложенным 1-м способом ($\Delta R_г = R1_г - R2_г$) сопоставлены частотный словарь русского языка и словник А, частотный словарь русского языка – словник Б. Словники сопоставлены на предмет того, какой из них приближен к «абсолютному» показателю и, как следствие, является более «стандартным», и какой отличается от стандартного распределения, а значит, является более индивидуализированным. В данном способе во внимание принимается определение «стандартный» / «индивидуализированный».

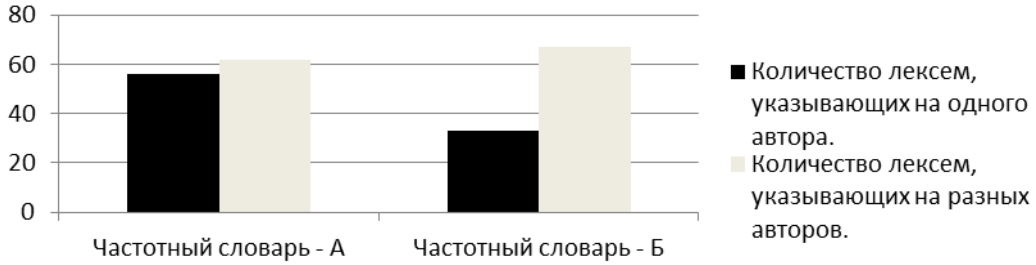


Рис. 3. Результат сопоставления частотного словаря русского языка и словников первым способом

Частотный словарь – словник А: оба типа лексем близки к 50% нахождения каждого типа в тексте. Есть лексемы (около половины), которые встречаются в двух словниках, есть же такие лексемы, которых либо вообще нет в сотне другого словника, либо дистанция между ними контрастная. Так, словник А более «стандартен».

Частотный словарь – словник Б: лексем типа *б* очевидно больше, чем лексем типа *а*. Это говорит о следующем: текст Б в рангово-частотном отношении построен иначе, чем частотный словарь русского языка. Возможны два варианта: 1) одно слово в разных словниках отличается контрастным ранговым показателем; 2) некоторых слов нет в сотне лексем другого словника. Словник Б, таким образом, отклоняется от «стандартного» абсолютного показателя – частотного словаря.

Количественное выражение диаграммы:

Частотный словарь – словник А: лексем типа *а* – 56, типа *б* – 62. Частотный словарь – словник Б: лексем типа *а* – 33, типа *б* – 67.

Согласно частотному словарю русского языка наиболее частотными словами являются предлоги и местоимения. Частотно-сопоставительные таблицы подтверждают это. Предполагаем, что самые частотные слова русского языка (разных частей речи) одинаково частотны в каждом авторском словнике. Сопоставительная таблица по пяти словникам показывает, какие слова одинаково часто присутствуют во всех пяти частотных словарях. Слова, которые есть в каждом из пяти словников на уровне одной группы слов: он, я, быть ($R_г = 1$), она ($R_г = 2$), мы ($R_г = 3$), только ($R_г = 4$).

Большинство слов находится в соседних группах ($\Delta R_г < 2$ и $\Delta R < 10$). Отметим, что основной лексический состав тождествен во всех словниках (например, такие лексемы, как рука, дело, два, жизнь и т.п.).

Следующий этап исследования – ***верифицирующий*** – анализ словников на предмет того, каких лексем нет в том или ином словнике в случае, когда они есть в другом. Методика «любимых» слов – анализ рангового распределения лексем первой сотни текста, отклоняющейся от рангового распределения в частотном словаре русского языка. «Любимые» слова – такие слова, которые автор использует настолько часто, что они вошли в сотню самых частотных слов анализируемого речевого

произведения, когда этих же слов нет в другом словнике. Этот пункт также показывает, насколько каждый из текстов «стандартен».

Примеры лексем словника А, которых нет в частотном словаре русского языка: все ($R_г$ = 2), просто ($R_г$ = 6), можно ($R_г$ = 8), мама ($R_г$ = 8), сегодня ($R_г$ = 10), любить ($R_г$ = 10), тут ($R_г$ = 10), хотя ($R_г$ = 10), работа ($R_г$ = 10) и пр. Всего лексем – 14.

Примеры лексем словника Б, которых нет в частотном словаре русского языка: нравиться ($R_г$ = 6), ничего ($R_г$ = 6), сейчас ($R_г$ = 6), больше ($R_г$ = 7), вообще ($R_г$ = 7), жить ($R_г$ = 7), нужно ($R_г$ = 7), сегодня ($R_г$ = 8), хорошо ($R_г$ = 8), всегда ($R_г$ = 9), много ($R_г$ = 9), музыка ($R_г$ = 9), тогда ($R_г$ = 9), почему ($R_г$ = 10), совсем ($R_г$ = 10) и пр. Всего лексем – 27.

Данный параметр подтверждает вывод о бóльшей «стандартности» текста А и бóльшем отклонении в показателях ранга в тексте Б. Отметим, что с возрастанием ранга группы увеличивается количество слов, отражающее разность их распределения в словнике и частотном словаре русского языка. Отклонения увеличиваются начиная с $R_г$ = 6. Предполагаем, что вторая, третья и последующие группы слов покажут бóльшие результаты относительно перечисленных способов идентификации персонотекста.

**4. Лингвоэкспертные выходы идентификации текста по его авторской принадлежности.** Работа имеет прямой практический выход: идентификация персонотекста вписывается в задачи судебной лингвистической экспертологии, в ее рамках «экспертиза является инструментом решения коллизионных ситуаций между субъектами права» [14. С. 7]. Экспертизу характеризует стремление к объективности: «...юридическая строгость предполагает, прежде всего, максимальную (насколько это возможно) правовую однозначность ответа» [15. С. 16]. Лингвисты-эксперты отмечают, что не существует четких методов проведения экспертизы, поэтому результаты отдалены от требуемой объективности. Следует говорить о «необходимости серьезных научных исследований по разработке и унификации принципов, методов и приемов лингвистической экспертизы текстов, включению ее в общую систему судебных экспертиз и нормативному закреплению ее положения в этой системе» [15. С. 17].

В лингвоэкспертологии данная работа представляет собой аналог идентификационной лингвистической экспертизы, целью которой является установление тождества объектов. В этом смысле объектами являются интернет-дневники. Описываемая ситуация экспертного анализа определяется А.Н. Барановым как «множественная неопределенность»: «...имеется множество текстов или их фрагментов. Необходимо установить, скольким авторам принадлежат тексты, атрибутировать каждый конкретному автору» [16. С. 1]. Для осуществления атрибуции персонотекстов разрабатывается вариант идентификации количественными методами. Так, исследование – эксперимент, в рамках которого решается теоретико-методическая (лингвоперсонологическая) задача идентификации текста, имеющая прямой выход в экспертную деятельность. Лингвоперсонология является методологической базой идентификационной экспертизы. Статистический анализ – базовый метод данной экспериментальной экспертизы.

Формализованный подход к решению задачи идентификации персонотекста позволяет проводить анализ независимо от эксперта, что способствует объективизации исследования. Конструктивная сторона исследования призвана дополнить методы науки, когда обращение к семантике затруднено или исключено (например, при большом объеме текстов). Представленная работа в определенной мере позволила выявить некоторые закономерности лексико-квантитативной структуры текстов, принадлежащих одному или разным авторам. Это дает возможность использования названной методики идентификации персонотекста в судебной авторoведческой экспертизе.

## ЛИТЕРАТУРА


1. *Кожина Н.Г.* Аксиологический подход к исследованию идентификации современной личности // Историческая и социально-образовательная мысль. № 1. 2012. URL: http://cyberleninka.ru/article/n/aksiologicheskaya-identifikatsiya-lichnosti (дата обращения: 10.10.2013).
2. *Леорда С.В.* Речевой портрет современного студента : автореф. дис. … канд. филол. наук. Саратов, 2006. 20 с. URL: http://www.dissforever.ru/order.php?id=714 (дата обращения: 26.09.2013).
3. *Замилова А.В.* Лингвосоционическое портретирование языковой личности (на материале блога) // Актуальные проблемы литературоведения и лингвистики : материалы конф. молодых ученых 1 апреля 2011 г. Томск, 2011. Вып. 12. Т. 1 : Лингвистика. С. 122–127.
4. *Солодянкина Н.В., Хвостова А.В.* Речевой портрет студента-филолога // Русская языковая личность в современном коммуникативном пространстве : материалы Междунар. науч. конф. Бийск, 2012. С. 155–159. URL: http://www2.bigpi.biysk.ru/ff/doc/ff_RYAL.pdf (дата обращения: 15.10.2013).
5. *Лингвоперсонология* и личностно-ориентированное обучение языку : учеб. пособие. Кемерово : КемГУ, 2009. 384 с.
6. *Морозов А.В.* Авторoведческая экспертиза текста договора // Юрислингвистика-5. Юридические аспекты языка и лингвистические аспекты права. Барнаул : Изд-во Алт. ун-та, 2004. С. 290–297.
7. *Голев Н.Д.* Лингвистическое сравнительное и авторoведческое исследование трех текстов // Юрислингвистика-10. Лингвоконфликтология и юриспруденция. Кемерово ; Барнаул : Изд-во Алт. ун-та, 2010. С. 422–431.
8. *Голев Н.Д.* Некоторые проблемы лексической и словообразовательной мотивации на оси частотности. К построению теории квантитативной мотивации. URL: lingvo.asu.ru/golev/articles/z11.html (дата обращения: 15.11.2013).
9. *Ахманова О.С., Мельчук И.А., Падучева Е.В., Фрумкина Р.М.* О точных методах исследования языка (о так называемой «математической лингвистике»). М. : Изд-во Моск. ун-та, 1961. 162 с.
10. *Хоменко А.Ю.* Алгоритм автоматизации идентификации автора письменного речевого произведения для судебного авторoведения // Юрислингвистика. 2013 (в печати). № 13.
11. *Родионова Е.С.* Лингвистические методы атрибуции и датировки литературных произведений (К проблеме «Мольер – Корнель») : автореф. дис. … канд. филол. наук. СПб., 2008. URL: http://cheloveknauka.com/v/56212/a?#?page=1 (дата обращения: 27.11.2013).
12. *Резанова З.И., Романов А.С., Мещеряков Р.В.* О выборе признаков текста, релевантных в авторoведческой экспертной деятельности // Вестн. Том. гос. ун-та. Филология. 2013. № 6 (26). С. 38–52. URL: http://sun.tsu.ru/mminfo/000063105/fil/26/image/26-038.pdf (дата обращения: 22.11.2013).

13. *Ляшевская О.Н., Шаров С.А.* Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М. : Азбуковник, 2009. URL: http://dict.ruslang.ru/freq.php (дата обращения: 20.09.2012).
14. *Голев Н.Д., Матвеева О.Н.* Значение лингвистической экспертизы для юриспруденции и лингвистики // Цена слова: Из практики лингвистических экспертиз текстов СМИ в судебных процессах по искам о защите чести, достоинства и деловой репутации. М. : Галерея, 2002. С. 257–266.
15. *Голев Н.Д.* От редактора: Актуальные проблемы юрислингвистической экспертизы // Юрислингвистика-3. Проблемы юрислингвистической экспертизы : межвуз. сб. науч. тр. Барнаул : Изд-во Алт. ун-та, 2002. С. 5–13.
16. *Баранов А.Н.* Авторизация текста: пример экспертизы // Введение в прикладную лингвистику : учеб. пособие. М. : Эдиториал УРСС, 2001. С. 43–51.

Статья представлена научной редакцией «Филология» 21 января 2014 г.

*Napreenko Galina V.* Kemerovo State University (Kemerovo, Russian Federation). E-mail: vila1991@mail.ru

**AUTHORSHIP IDENTIFICATION OF THE TEXT ON THE LEXICAL LEVEL (FORMAL-QUANTITATIVE MODEL).**

**Key words:** identification linguistics; language personal analysis; personal text; formal-quantitative model of text identification; quantitative linguistics; forensic linguistics.

The article deals with the development of the way to qualify the author's speech individuality – the formal-quantitative model of identification of the personal text on the lexical level. The personal text contains language characteristics, the variants of linguistic personality behaviour that can be expressed through quantitative indicators. In the present article the lexical level is regarded from the point of view of lexeme usage frequency in the personal text. Qualitative interpretation of the formal-quantitative data gives us some particular indicators on the lexemes distribution in speech, depicts the variety of the author's preferences of linguistic personalities that can help to define the author of the text. So, the article states the types of language personal identification. The following hypothesis is formulated in the article: the text contains individual lexical characteristics which, expressed through contrasting qualitative data, can serve as the identifiers of the text. The suggested model is presented at the first stage in two different ways, and they are based on a different interpretation of the term rank. The working formulae are: $\Delta R_г = R1_г - R2_г$ (the rank is understood as a rank of a group of words), $\Delta R = R1 - R2$ (the rank of each word). Every word is analysed by means of these formulae within four frequency-contrastive tables. Based on the results of the lexemes vocabulary rank comparison of each frequency a contrastive table a diagram is made; and a conclusion on lexemes frequency is drawn. These two approaches have shown the possibility of their usage in attribution of the texts. Taking into consideration the conditions described (the genre of the text, approximate number of word usage) is obligatory, because any other conditions have not been verified. The second stage is represented through the comparison of the vocabulary of the texts with the modern frequency dictionary of the Russian language. The dictionary serves as the "absolute" index of the distribution of words, and shows which words are closer to the "absolute index", and, as a result, are more "standard", and which words differ.
Within this study a theoretical and methodological (language personal analysis) problem of text identification is observed, the problem which has a direct access to different expert activities. Language personal analysis becomes a methodological base for identification examination. In the system of forensic linguistics this work serves as an analogue of the linguistic identification examination, the purpose of which is to establish the identity of objects. The present work has identified some regular occurrences in the lexical-quantitative structures of the texts belonging to the same author or different ones. It gives us an opportunity to use the mentioned personal texts identification methods in authorship examination.

## REFERENCES

1. *Kozhina N.G.* Aksiologicheskiy podkhod k issledovaniyu identifikatsii sovremennoy lichnosti // Istoricheskaya i sotsial'no-obrazovatel'naya mysl'. № 1. 2012. URL: http://cyberleninka.ru/article/n/aksiologicheskaya-identifikatsiya-lichnosti (data obrashcheniya: 10.10.2013).
2. *Leorda S.V.* Rechevoy portret sovremennogo studenta : avtoref. dis. … kand. filol. nauk. Saratov, 2006. 20 s. URL: http://www.dissforever.ru/ order.php?id=714 (data obrashcheniya: 26.09.2013).
3. *Zamilova A.V.* Lingvosotsionicheskoe portretirovanie yazykovoy lichnosti (na materiale bloga) // Aktual'nye problemy literaturovedeniya i lingvistiki : materialy konf. molodykh uchenykh 1 aprelya 2011 g. Tomsk, 2011. Vyp. 12. T. 1 : Lingvistika. S. 122–127.
4. *Solodyankina N.V., Khvostova A.V.* Rechevoy portret studenta-filologa // Russkaya yazykovaya lichnost' v sovremennom kommunikativnom prostranstve : materialy Mezhdunar. nauch. konf. Biysk, 2012. S. 155–159. URL: http://www2.bigpi.biysk.ru/ff/doc/ff_RYAL.pdf (data obrashcheniya: 15.10.2013).
5. *Lingvopersonologiya* i lichnostno-orientirovannoe obuchenie yazyku : ucheb. posobie. Kemerovo : KemGU, 2009. 384 s.
6. *Morozov A.V.* Avtorovedcheskaya ekspertiza teksta dogovora // Yurislingvistika-5. Yuridicheskie aspekty yazyka i lingvisticheskie aspekty prava. Barnaul : Izd-vo Alt. un-ta, 2004. S. 290–297.
7. *Golev N.D.* Lingvisticheskoe sravnitel'noe i avtorovedcheskoe issledovanie trekh tekstov // Yurislingvistika-10. Lingvokonfliktologiya i yurisprudentsiya. Kemerovo ; Barnaul : Izd-vo Alt. un-ta, 2010. S. 422–431.
8. *Golev N.D.* Nekotorye problemy leksicheskoy i slovoobrazovatel'noy motivatsii na osi chastotnosti. K postroeniyu teorii kvantitativnoy motivatsii. URL: lingvo.asu.ru/golev/articles/z11.html (data obrashcheniya: 15.11.2013).
9. *Akhmanova O.S., Mel'chuk I.A., Paducheva E.V., Frumkina R.M.* O tochnykh metodakh issledovaniya yazyka (o tak nazyvaemoy «matematicheskoy lingvistike»). M. : Izd-vo Mosk. un-ta, 1961. 162 s.
10. *Khomenko A.Yu.* Algoritm avtomatizatsii identifikatsii avtora pis'mennogo rechevogo proizvedeniya dlya sudebnogo avtorovedeniya // Yurislingvistika. 2013 (v pechati). № 13.
11. *Rodionova E.S.* Lingvisticheskie metody atributsii i datirovki literaturnykh proizvedeniy (K probleme «Mol'er – Kornel'») : avtoref. dis. … kand. filol. nauk. SPb., 2008. URL: http://cheloveknauka.com/v/56212/a?#?page=1 (data obrashcheniya: 27.11.2013).
12. *Rezanova Z.I., Romanov A.S., Meshcheryakov R.V.* O vybore priznakov teksta, relevantnykh v avtorovedcheskoy ekspertnoy deyatel'nosti // Vestn. Tom. gos. un-ta. Filologiya. 2013. № 6 (26). S. 38–52. URL: http://sun.tsu.ru/mminfo/000063105/fil/26/image/26-038.pdf (data obrashcheniya: 22.11.2013).
13. *Lyashevskaya O.N., Sharov S.A.* Chastotnyy slovar' sovremennogo russkogo yazyka (na materialakh Natsional'nogo korpusa russkogo yazyka). M. : Azbukovnik, 2009. URL: http://dict.ruslang.ru/freq.php (data obrashcheniya: 20.09.2012).
14. *Golev N.D., Matveeva O.N.* Znachenie lingvisticheskoy ekspertizy dlya yurisprudentsii i lingvistiki // Tsena slova: Iz praktiki lingvisticheskikh ekspertiz tekstov SMI v sudebnykh protsessakh po iskam o zashchite chesti, dostoinstva i delovoy reputatsii. M. : Galereya, 2002. S. 257–266.
15. *Golev N.D.* Ot redaktora: Aktual'nye problemy yurislingvisticheskoy ekspertizy // Yurislingvistika-3. Problemy yurislingvisticheskoy ekspertizy : mezhvuz. sb. nauch. tr. Barnaul : Izd-vo Alt. un-ta, 2002. S. 5–13.
16. *Baranov A.N.* Avtorizatsiya teksta: primer ekspertizy // Vvedenie v prikladnuyu lingvistiku : ucheb. posobie. M. : Editorial URSS, 2001. S. 43–51.