*Философия образования*

**В.Н. Фокина**, *кандидат социологических наук*
**А.В. Лукьянова**, *кандидат технических наук*
**М.Е. Широкова**, *кандидат социологических наук*

# АНАЛИЗ ПОДХОДОВ К УСТАНОВЛЕНИЮ КРИТЕРИЕВ ПОСТРОЕНИЯ ИНДИВИДУАЛЬНОГО ЛЕКСИЧЕСКОГО ПРОФИЛЯ С ЦЕЛЬЮ ОПРЕДЕЛЕНИЯ АВТОРСТВА ТЕКСТОВ

Рассмотрены возможности определения авторства текстов на основе критериальной системы построения индивидуального лексического профиля. Показаны возможности составления «ядра» словаря автора методами математической статистики и возможности оценки авторства на основе различных статистических критериев.

***Ключевые слова**: плагиат, определение авторства, индивидуальный лексический профиль, идиостиль, частотный словарь, словоупотребление.*

В настоящее время проблема установления авторства при написании рефератов, курсовых и дипломных работ, а также диссертаций во все мире приобрела поистине огромные масштабы. По различным данным в плагиате замешаны до 70% студентов, а скандалы с разоблачением ученых и политиков «списавших» свои докторские диссертации уже давно воспринимаются как нечто обыденное.

Причем количество сайтов, предлагающих студентам и аспирантам услуги по написанию различного вида работ, постоянно увеличивается. В настоящее время в Интернете существует 24 млн ссылок, перейдя по которым можно скачать, например, курсовую работу. Большинство работ «кочуют» с сайта на сайт без изменений, или с незначительными правками.

По данным социологических исследований ГУ ВШЭ, чаще всего скачивают рефераты, эссе и курсовые студенты четвертых курсов вузов – 52%. Реже первокурсники – их 47%. Покупают готовые работы от 3 до 7% студентов [1].

Поэтому естественно, что в системе образования на всех уровнях возникает вопрос, как оценить самостоятельность выполнения обучающимся письменных работ.

Традиционно одним из способов определения уровня самостоятель-

ности обучающегося при подготовке текстов является определение объема заимствования в тексте работы с использованием автоматизированных программных средств, которые широко представлены, в том числе в Интернете в свободном доступе. Вузы для оценки уровня самостоятельности написания текстов широко используют как разработки сторонних организаций (самая известная из них – система «Антиплагиат» компании Форексис используется МГУ и ВАК), так и собственные разработки, например, «АУРА-Текст» СПбГУ.

Основная проблема этих и других подобных программных средств состоит в базе текстов, загруженных в данные ПО в качестве основы для сравнения и выявления степени оригинальности анализируемой работы. Как правило, эти базы представляют собой «черный ящик», содержание которого является коммерческой тайной разработчиков ПО.

Кроме того, существенная проблема данных программных средств состоит в отсутствия в этих системах учета легальных заимствований – правильно оформленных цитат. Системы просто выдают долю заимствованного текста, включая в нее все цитаты. При этом строго формализованных единых требований и правил по оформлению печатных научных работ и, в частности, цитат, в настоящее время практически не существует. Имеется также множество регламентов оформления студенческих работ в различных образовательных организациях.

СГА ставил перед собой задачу не только борьбы с антиплагиатом в студенческих и аспирантских работах, но создания таких критериев оценки самостоятельности работы обучающегося над текстом, при которых ему будет выгоднее самостоятельно написать работу, чем пользоваться чужими текстами.

Для этого необходимо определить индивидуальный стиль изложения обучающегося, который назван авторами статьи «*индивидуальный лексический профиль*»[1]. Индивидуальный лексический профиль обучающегося может стать кодом для идентификации подготовленных им текстов, по которому, вероятно, станет возможным реально оценить степень самостоятельности при написании текстов различной тематической направленности.

Вопросам исследования стилей различных авторов и определения авторства посвящено значительное количество работ отечественных и

[1] Под индивидуальным лексическим профилем будем понимать перечень наиболее часто встречающихся слов с частотами их употребления.

зарубежных авторов. В языкознании существует определенный термин для выражения индивидуального авторского стиля – идиостиль[1]. Главным вопросом в изучении идиостиля является проблема установления авторства текста и поиск объективных критериев выявления индивидуально-авторских признаков, позволяющих с достаточной достоверностью определить автора исследуемого произведения.

Основой проведения статистического анализа текста являются частотные словари. Частотный словарь включает в себя те слова или другие лингвистические единицы (словоформы[2], словосочетания), которые зарегистрированы составителем в обследованных им текстах (или тексте). При этих словах, словоформах и т. д. указываются частоты их употребления в данных текстах (тексте) [2].

В связи с развитием вычислительной техники и возможностями одновременной обработки больших объемов данных, значительную роль в исследовании текстов в настоящее время играют формально-количественные, статистические методы, которые решают практические задачи по атрибуции литературных произведений, авторство которых вызывает дискуссии (работы Г.В. Ермоленко, Л.В. Милова, М.Ю. Мухина, А.А. Поликарпова, Д.В. Хмелева, М.В. Копотева, В.А. Плунгяна, Г. Хьетсо, А.Я. Шайкевича и др.) [3]. В зарубежной лингвистике стилистические исследования, включающие статистический анализ, обычно относятся к направлению, получившему название «стилометрия» (R.H. Baayen, J.F. Burrows, T.N. Corns, D.I. Holmes, D.L. Hoover, H. Love).

Особенности идиостилей и признаки авторских концептуальных систем М. Булгакова, В. Набокова, А. Платонова и М. Шолохова описаны М.Ю. Мухиным в его многочисленных публикациях и обобщены в монографии [4]. Идиостиль указанных авторов рассматривается им в двух аспектах: во-первых, индивидуально-авторская частотная лексика, и, во-вторых, авторские лексические биграммы (пары слов, извлеченные из одного фразового контекста), регулярно встречающиеся в разных произведениях одного автора и нехарактерные для творчества других писателей.

В работах Д.В. Хмелева [5, с. 115–126], О.В. Кукушкиной и А.А. По-

[1] Идиостиль – система содержательных и формальных лингвистических характеристик, присущих произведениям определенного автора, которая делает уникальным воплощенный в этих произведениях авторский способ языкового выражения.

[2] Словоформа – одна из косвенных форм слова, полученная из нормальной формы слова с помощью склонения или спряжения.

ликарпова [6] для определения авторства текста используется метод, основанный на формальной математической модели встречаемости последовательности элементов текста как реализации цепи Маркова. В качестве элементов текста используются последовательности букв и последовательности грамматических классов слов. По тем произведениям автора, которые достоверно им созданы, вычисляется матрица переходных частот употребления пар элементов (букв, грамматических классов слов и т.п.). Она служит оценкой матрицы вероятности перехода из элемента в элемент. Матрица переходных частот строится для каждого автора. Для каждого автора оценивается вероятность того, что именно он написал анонимный текст (или фрагмент текста). Автором анонимного текста полагается тот, у которого вычисленная оценка вероятности больше (т. е. используется принцип максимального правдоподобия) [6].

В результате опробования метода на 385 текстах 82 писателей установлено, что частоты употребления пар букв и пар грамматических классов в тексте на русском языке являются достаточно устойчивой характеристикой автора и их можно использовать, чтобы решать проблемы спорного авторства текста. Однако, несмотря на высокую точность полученных авторами результатов (73% точных определений) данный метод является трудоемким и требует предварительной обработки текстов и объемных расчетов.

К тому же рассмотренные методики оценки индивидуальных авторских стилей и определения авторства не могут использоваться в образовательной сфере, так как процесс оценивания студенческих работ должен быть «встроен» в образовательный процесс с установленными периодами представления результатов аттестации письменных работ обучающихся.

Например, численность обучающихся в СГА составляет около 70 тыс. человек, в течение семестра каждый обучающийся в соответствии с учебным планом должен написать не менее трех письменных творческих работ, средний объем одной работы – 15 страниц. Таким образом, общий объем обрабатываемых текстов составляет: 70 тыс. чел. × 3 работы/чел. × 15 страниц/1 работа = 3,15 млн страниц, что при среднем объеме 300 слов на 1 странице составляет 945 млн слов. Проводить дополнительную предварительную подготовку текстов работ такого объема для их последующего анализа с использованием действующих методик практически нереально. Таким образом, описанные ранее методики практически невозможно

применить в установленные сроки, определенные рамками учебного процесса.

Указанные причины обусловили необходимость формирования собственных подходов к определению индивидуального лексического стиля авторов с использованием статистических методов и средств информатизации.

В СГА разработано собственное программное обеспечение – интеллектуальный робот «Живой язык» (далее – ИР ЖИЯЗ). Данный ИР позволяет в автоматизированном режиме проводить лемматизацию[1] загруженных в него текстов, преобразуя употребляемые в тексте слова в различных словоформах к нормальной форме слова, а также проводить подсчет относительных частот употребления слов в анализируемых текстах (обозначим ее как ЧМ – частота на миллион словоупотреблений[2]).

Возможности, которыми обладает ИР ЖИЯЗ, позволяют в автоматизированном режиме формировать частотные словари каждого анализируемого текста. Анализ словарей, сформированных по текстам разных статей одного и того же автора, показал, что у каждого автора есть слова, которые он употребляет чаще других. Причем эти слова он использует практически в каждом тексте (если это текст достаточно большого объема), значения относительных частот употребления слов сопоставимы (разброс частоты составляет не более 25%). Такие слова, которые по нашему предположению, будут определять индивидуальный авторский стиль, названы детерминантами. Перечень слов-детерминант со средними относительными частотами их употребления определим как *индивидуальный лексический профиль*.

Наиболее яркими лексическими профилями, на наш взгляд, должны обладать известные ученые, признанные авторитеты в различных областях знаний. Учитывая, что СГА – вуз гуманитарный, то для выбора слов-детерминант с целью формирования индивидуального лексического профиля и выработки критериев для его формирования выбраны следующие авторы – Дмитрий Сергеевич Лихачев, искусствовед и филолог, академик РАН, Леонид Иванович Абалкин – доктор экономических наук, академик РАН и Джангир Аббасович Керимов – доктор юридических наук, член-корреспондент РАН.

[1] Лемматизация – это преобразование словоформ к базовой (нормальной) форме слова.

[2] Частота на миллион словоупотреблений (относительная частота употребления слова) рассчитывается как отношение количества употреблений слова в тексте к общему количеству словоупотреблений в тексте.

Было сделано предположение, что значение ЧМ каждого слова – случайная величина. Для дальнейшего исследования необходимо определить, какому закону подчиняется распределение ЧМ каждого слова. Исследования по определению закона распределения были проведены на основе корпуса текстов Д.С. Лихачева, объем которых был достаточен для таких исследований – 482 статьи общим объемом 1,25 млн словоупотреблений и 29507 слов.

Для установления закона распределения, которому подчиняются частотные характеристики слов, корпус текстов Д.С. Лихачева был разделен на 35 подкорпусов. В результате для каждого слова из корпуса текстов получили 35 значений ЧМ (в нашем случае – исследуемой случайной величины), по которым можно построить графики плотности распределения.

Для построения графиков случайным образом отобрали 6 слов, попавших в каждый из 35 подкорпусов, т. е. в каждом случае имеющих ЧМ, отличные от нуля:

1) полный;
2) свой;
3) их;
4) мочь;
5) сильный;
6) некоторый.

Итак, имеем 6 случайных величин, для каждой из которых известно:

- 35 значений ЧМ;
- математическое ожидание, рассчитанное как среднее значение ЧМ;
- поле (диапазон) рассеивания $R = ЧМ_{max} – ЧМ_{min}$;
- среднеквадратичное отклонение σ.

Был применен следующий алгоритм для построения графика плотности распределения.

1. В качестве диапазона построения графика (ось абсцисс) взят диапазон рассеивания.

2. В качестве шага для определения расчетных точек по шкале абсцисс принято значение σ – среднеквадратичное отклонение.

3. Подсчитали число значений ЧМ, соответствующих значениям ЧМ на данном шаге. Пользуясь описанным выше алгоритмом, построили графики плотности распределения ЧМ для 6 слов (рис. 1–2).

Анализ полученных графиков показал, что ЧМ всех отобранных случайным образом слов имеет распределение, близкое к нормальному, т. е. значение ЧМ каждого слова будет группироваться возле среднего значения.

Таким образом, мы вправе принимать за ЧМ каждого слова его математическое ожидание, или среднее значение ЧМ по различным подкорпусам. В этом случае можно пренебречь возможностью отклонения значения ЧМ слова от его матожи-

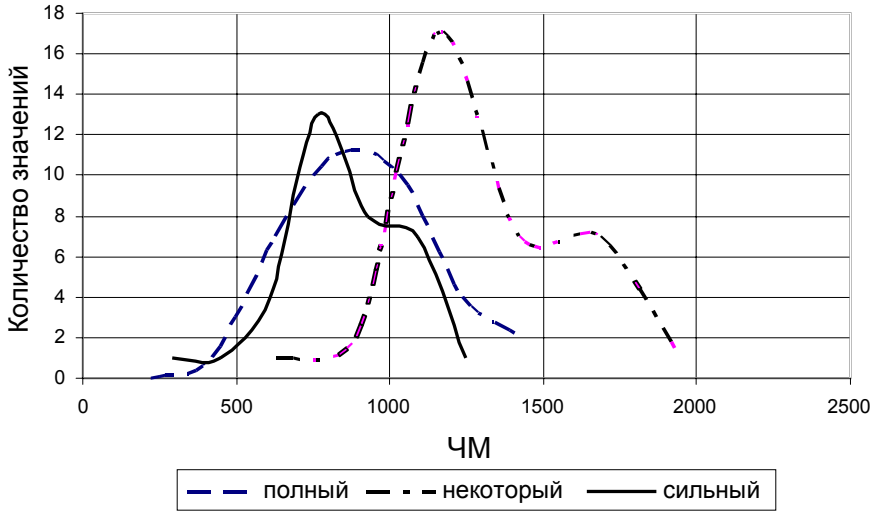


**Рис. 1.** Плотность распределения ЧМ слов «полный», «некоторый», «сильный»

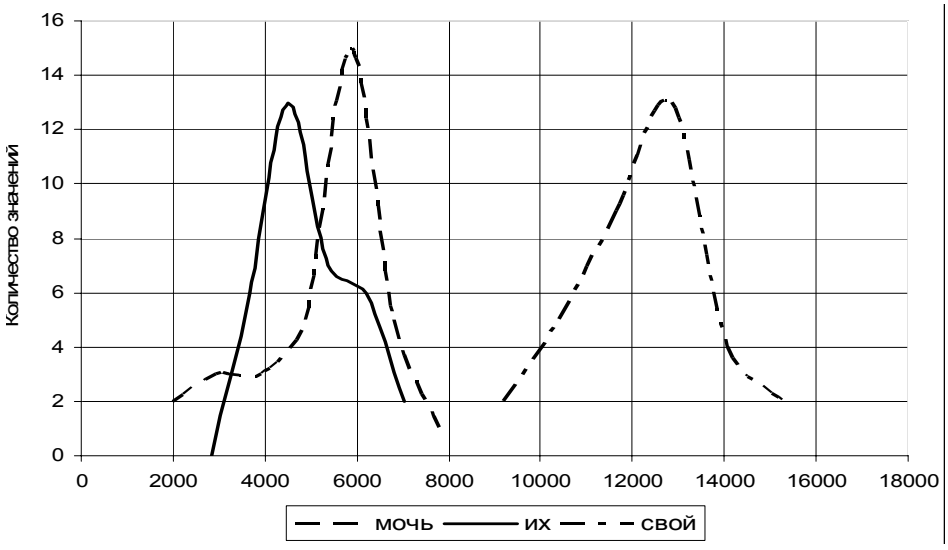


**Рис. 2.** Плотность распределения ЧМ слов «мочь», «их», «свой»

дания, превышающего 3σ, и в качестве меры рассеивания (дисперсии) значений ЧМ принять $\sigma^2$.

Для проведения исследований по выявлению слов-детерминант, определяющих авторский стиль текстов указанных ученых по каждому из них были сформированы по десять подкорпусов текстов, содержащих статьи, заметки, воспоминания,

главы монографий. Все подкорпуса имеют приблизительно равный объем по 4,3 Мб каждый (~125000 словоупотреблений).

Исследование начали с изучения авторского стиля Д.С. Лихачева. В результате анализа слов, выделенных из 10 подкорпусов, установлено, что 3133 слова входят в каждый из 10 подкорпусов. Из этих слов и выделяли в дальнейшем слова-детерминанты.

Для установления слов-детерминант, которые по нашему предположению должны составлять ядро словаря Д.С. Лихачева и присутствовать в каждой его статье, причем с близким значением ЧМ, нами использован следующий подход.

В математической статистике одним из важных показателей степени разбросанности исследуемого параметра является коэффициент вариации. Чем больше значение коэффициента вариации, тем относительно больший разброс и меньшая выравненность исследуемых значений. Если коэффициент вариации меньше 10%, то изменчивость вариационного ряда принято считать незначительной, от 10 до 20% относится к средней, больше 20% и меньше 33% к значительной и если коэффициент вариации превышает 33%, то это говорит о неоднородности информации и необходимости исключения самых больших и самых маленьких значений.

Поэтому слова-детерминанты, определяющие индивидуальный лексический профиль, выявляли путем расчета коэффициента вариации $V$, который характеризует относительную меру отклонения измеренных значений от математического ожидания:

$$V = \sigma/A \times 100\%,$$

где $V$ – коэффициент вариации; σ – корень квадратный из дисперсии; $A$ – математическое ожидание.

Учитывая, что ЧМi – дискретные величины, то математическое ожидание рассчитывалось как среднеарифметическое значение ЧМ по подкорпусам текстов.

По результатам расчета коэффициента вариации получено 3 перечня слов с разными значениями коэффициента вариации:

- слова, коэффициент вариации ЧМ которых не превышает 10%, количество слов – 3;
- слова, коэффициент вариации ЧМ которых не превышает 20%, количество слов – 38;
- слова, коэффициент вариации ЧМ которых не превышает 33%, количество слов – 334.

Таким образом, для исследуемых подкорпусов текстов Д.С. Лихачева 38 слов (1,2% от всех слов, встречающихся во всех подкорпусах текстов) имеют среднюю изменчивость ЧМ,

296 слов (9,4% от всех слов, встречающихся во всех подкорпусах текстов) имеют значительную изменчивость ЧМ и только три слова из всех подкорпусов имеют незначительную изменчивость ЧМ.

В результате для дальнейшего исследования в качестве детерминант выбрали 142 слова, коэффициент вариации которых лежит в интервале от 0 до 25%.

Аналогичный подход использовали и при выявлении детерминант Л.И. Абалкина и Д.А. Керимова. Из подкорпусов выделяли слова, которые встречаются во всех подкорпусах. В текстах Л.И. Абалкина в 13066 слов во всех подкорпусах совпадающих слов 1033; в текстах Д.А. Керимова в 9700 слов во всех подкорпусах совпадающих слов 748.

По каждому слову из выделенной совокупности рассчитано среднеарифметическое значение ЧМ в исследуемых подкорпусах (математическое ожидание), дисперсия и коэффициент вариации.

В результате расчетов установлено, что по текстам Л.И. Абалкина количество слов, коэффициент вариации ЧМ которых не превышает 25%, равно 59; по текстам Д.А. Керимова – 36.

Сравнительный анализ слов, вошедших в детерминанты Л.И. Абалкина и Д.А. Керимова, с детерминантами Д.С. Лихачева показал, что часть слов во всех трех перечнях совпадает. К таким совпадающим словам относятся: *этот, тот, свой, такой, мочь, другой, их, новый, весь, самый, первый, многий, иметь, число, идти, часть, мера, главный, общий, сам, целый, сторона, место, положение, причина, широкий, сильный.* Таким образом, из 59 детерминант Л.И. Абалкина с детерминантами Д.С. Лихачева совпало 28 слов, слово *время* совпало с детерминантами Д.А. Керимова. То есть количество детерминант Л.И. Абалкина равно: 59–28–1 = 30.

Из 36 детерминант Д.А. Керимова с детерминантами Д.С. Лихачева совпало 13 слов: *этот, тот, их, весь, такой, каждый, сила, следует, последний, какой, место, причина, род;* слово *время* совпало с детерминантами Л.И. Абалкина. Тогда, количество детерминант Д.А. Керимова равно: 36 – 28 – 1 = 22 слова.

Таким образом, для дальнейшего исследования в качестве детерминант Л.И. Абалкина будем использовать 30 слов, в качестве детерминант Д.А. Керимова – 22 слова.

С целью выработки критериальной системы оценки авторства и формирования индивидуального лексического профиля выделенные из генеральной совокупности корпусов текстов Д.С. Лихачева, Л.И. Абалкина и Д.А. Керимова детерминанты этих авторов проана-

лизированы с позиции следующих критериев:

• **процент детерминант из генеральной совокупности**, проявившихся в исследуемом тексте (критерий № 1);

• **процент детерминант из генеральной совокупности**, ЧМ которых в исследуемом тексте попадает в зону эквивалентности (критерий №2);

• **среднее отклонение ЧМ в исследуемом тексте** от ЧМ генеральной совокупности (в долях σ) у проявившихся детерминант (критерий № 3);

• **средняя разница рангов детерминант** генеральной совокупности и в исследуемом тексте (критерий № 4).

С целью установления возможности использования критериев для оценки индивидуального авторского стиля был проведен анализ «поведения» детерминант в произвольно выбранных статьях автора, не вошедших в подкорпуса его текстов, а также в статьях и подкорпусах текстов других авторов. Средний объем одной статьи составляет 700 словоупотреблений, одного подкорпуса ~ 58000 словоупотреблений.

Рассмотрим далее полученные результаты отдельно по каждому критерию.

В качестве первого критерия установления авторства нами выбран *процент детерминант, проявившихся в исследуемом тексте*.

Критерий рассчитывается как отношение количества детерминант, выделенных из корпуса, употребляемых в исследуемом тексте, к общему количеству детерминант.

Для детерминант Д.С. Лихачева значение этого критерия в его статьях равно 97,5%, 94,15% и 94,17% (табл. 1). Как видно из табл.1, в статьях Л.И. Абалкина среднее значение критерия равно 39,72% (от 20,8 до 64,2%), в статьях Д.А. Керимова – 36,25%.

Однако анализ значений критерия №1 для детерминант Д.С. Лихачева на текстах большего объема – по подкорпусам текстов Л.И. Абалкина и Д.А. Керимова показывает, что средние значения (95,2 и 94,8% соответственно) близки к значению этого критерия в статьях Д.С. Лихачева.

Таким образом, с увеличением объема анализируемого текста словарь употребляемых автором слов расширяется, вероятность употребления слов с большой частотой существенно возрастает. Учитывая, что в перечне детерминант Д.С. Лихачева 25% слов (35 из 120) имеют высокую частоту употребления и могут быть отнесены к общеупотребимым, то возможности применения данного критерия для установления авторства текста, вероятно, существенно зависят от объема анализируемого текста и не могут дать стабильный достоверный результат.

Значения критерия №1 для детерминант Л.И. Абалкина показаны в табл. 2, для детерминант Д.А. Керимова – в табл. 3.

*Таблица 1*

**Процент детерминант Д.С. Лихачева, проявившихся в произвольно выбранных статьях Д.С. Лихачева и подкорпусах Л.И. Абалкина и Д.А. Керимова (критерий № 1)**

| № п/п | Исследуемые тексты | Значение критерия №1, рассчитанное по текстам различных авторов, % | | |
|---|---|---|---|---|
| | | Д.С. Лихачев | Л.И. Абалкин | Д.А. Керимов |
| 1 | Статья № 1 | 97,5 | 34,17 | 34,17 |
| 2 | Статья № 2 | 94,15 | 20,83 | 38,33 |
| 3 | Статья № 3 | 94,17 | 64,17 | – |
| 4 | Среднее значение по статьям | 95,28 | 39,72 | 36,25 |
| 5 | Подкорпус № 1 | – | 93,3 | 95,8 |
| 6 | Подкорпус № 2 | – | 96,7 | 93,3 |
| 7 | Подкорпус № 3 | – | 95 | 95 |
| 8 | Подкорпус № 4 | – | 93,3 | 95,8 |
| 9 | Подкорпус № 5 | – | 97,5 | 94,2 |
| 10 | Среднее значение по подкорпусам | – | 95,2 | 94,82 |

*Таблица 2*

**Процент детерминант Л.И. Абалкина, проявившихся в произвольно выбранных статьях Л.И. Абалкина и подкорпусах Д.А. Керимова и Д.С. Лихачева (критерий № 1)**

| № п/п | Исследуемые тексты | Значение критерия № 1, рассчитанное по текстам различных авторов | | |
|---|---|---|---|---|
| | | Л.И. Абалкин | Д.А. Керимов | Д.С. Лихачев |
| 1 | Подкорпус/статья № 1 | 60% | 93% | 83% |
| 2 | Подкорпус/статья № 2 | 30% | 100% | 90% |
| 3 | Подкорпус/статья № 3 | 77% | 97% | 93% |
| 4 | Среднее значение по подкорпусам/статьям | 56% | 97% | 89% |

*Таблица 3*

**Процент детерминант Д.А. Керимова, проявившихся в произвольно выбранных статьях Д.А. Керимова и подкорпусах Л.И. Абалкина и Д.С. Лихачева (критерий № 1)**

| № п/п | Исследуемые тексты | Значение критерия №1, рассчитанное по текстам различных авторов | | |
|---|---|---|---|---|
| | | Д.А. Керимов | Л.И. Абалкин | Д.С. Лихачев |
| 1 | Подкорпус/статья № 1 | 41% | 91% | 91% |
| 2 | Подкорпус/статья № 2 | 59% | 91% | 100% |
| 3 | Подкорпус/статья № 3 | – | 95% | 95% |
| 4 | Среднее значение по подкорпусам/статьям | 50% | 92% | 95% |

Как видно из табл. 2 и 3, около 50% детерминант проявляются в статьях их авторов (минимальное значение – 30% в статье Л.И. Абалкина № 2 и максимальное значение – 77% в статье Л.И. Абалкина № 3). При этом, в подкорпусах других авторов проявляются более 90% детерминант.

Как было показано ранее при исследовании «поведения» детерминант Д.С. Лихачева, значение этого критерия существенно зависит от объема статьи и словаря анализируемого текста. В нашем случае объем словаря статьи Л.И. Абалкина № 2 составляет 286 слов, а статьи № 3 – 1274 слова, средний объем словаря подкорпусов ~ 57000 слов. Вероятно, в тексте небольшого объема, написанном на узкоспециализированную тему все детерминанты просто «не успевают» проявиться, особенно когда они не относятся к повседневным словам (например, *концепция, позиция, фундаментальный* – детерминанты Д.А. Абалкина), *специфический, обретать, свидетельствовать* – детерминанты Д.А. Керимова).

Таким образом, результаты исследований показали, что критерий №1 не может дать стабильного результата на текстах различного объема, а следовательно, его не возможно использовать в качестве критерия для выявления индивидуального лексического профиля автора.

Рассмотрим далее второй критерий, выбранный нами – *процент детерминант из генеральной совокупности, ЧМ которых в исследуемом тексте попадает в зону эквивалентности.*

В качестве зоны эквивалентности нами выбрана зона отклонения величины ЧМ слов-детерминант

в подкорпусах текстов от среднего значения ЧМ на величину среднеквадратического отклонения.

Границы зоны эквивалентности ЧМ слов-детерминант Д.С. Лихачева рассчитывали по следующей формуле:

$$ЧМ_{min} = ЧМ_{ср.} - n \times \sigma,$$
$$ЧМ_{max} = ЧМ_{ср.} + n \times \sigma,$$

где $ЧМ_{min}$ – минимальное значение зоны эквивалентности; $ЧМ_{max}$ – максимальное значение зоны эквивалентности; $ЧМ_{ср.}$ – среднеарифметическое значение ЧМ по подкорпусам текстов (матожидание); σ – среднеквадратичное отклонение ЧМ от матожидания по подкорпусам; *n* – коэффициент, учитывающий долю σ для изменения размеров зоны эквивалентности.

Для изучения частотных характеристик детерминант в статьях Д.С. Лихачева, не входящих в корпус, критерий №1 был рассчитан для двух зон: 2,5 и 3σ (табл. 4).

Результаты показали, что среднее значение критерия для ширины зоны 2,5σ составляет 60,3%, причем по одной из статей это значение составляет 55%. Такое значение критерия, близкое к 50% говорит о том, что вероятность попадания ЧМ детерминанты из произвольной статьи в зону эквивалентности 2,5σ близка к 50%, т. е. это событие с равной вероятностью может произойти или не произойти. Поэтому для дальнейших исследований нами выбрана ширина зоны эквивалентности ±3×σ.

Значение критерия № 2, рассчитанное по статьям других авторов (табл. 5) показывает, что в статьях Л.И. Абалкина процент детерми-

*Таблица 4*

**Процент детерминант Д.С. Лихачева, попавших в зону эквивалентности (критерий № 2), в произвольно выбранных статьях Д.С. Лихачева для разной ширины зоны**

| № п/п | Исследуемые тексты | Процент детерминант, попавших в зону эквивалентности, % | |
|---|---|---|---|
| | | ширина зоны 2,5σ | ширина зоны 3σ |
| 1 | Статья Д.С. Лихачева № 1 | 60,83 | 73,33 |
| 2 | Статья Д.С. Лихачева № 2 | 55,83 | 63,33 |
| 3 | Статья Д.С. Лихачева № 3 | 64,17 | 74,17 |
| 4 | Среднее значение по статьям Д.С. Лихачева | 60,28 | 70,3 |

*Таблица 5*

**Процент детерминант Д.С. Лихачева, попавших в зону эквивалентности в произвольно выбранных статьях Д.С. Лихачева и подкорпусах Л.И. Абалкина и Д.А. Керимова (критерий № 2)**

| № п/п | Исследуемые тексты | Значение критерия № 2, рассчитанное по текстам различных авторов, % | | |
|---|---|---|---|---|
| | | Д.С. Лихачев | Л.И. Абалкин | Д.А. Керимов |
| 1 | Статья № 1 | 73,3 | 10 | 14,2 |
| 2 | Статья № 2 | 63,3 | 4,2 | 13,3 |
| 3 | Статья № 3 | 74,2 | 37,5 | – |
| 4 | Среднее значение по статьям | 70,3 | 17,2 | 13,75 |
| 5 | Подкорпус № 1 | – | 55,0 | 56,7 |
| 6 | Подкорпус № 2 | – | 63,3 | 59,2 |
| 7 | Подкорпус № 3 | – | 57,5 | 56,7 |
| 8 | Подкорпус № 4 | – | 56,7 | 61,7 |
| 9 | Подкорпус №5 | – | 70,0 | 53,3 |
| 10 | Среднее значение по подкорпусам | – | 60,5 | 57,5 |

нант, ЧМ которых попали в зону эквивалентности 3σ, варьируется от 4,17 до 37,5% (среднее значение – 17,2%).

В статьях Д.А. Керимова этот процент составляет в среднем 13,75%. Таким образом, полученные значения критерия №2 по детерминантам Д.С. Лихачева в статьях Л.И. Абалкина и Д.А. Керимова, значения ЧМ которых попадают в зону эквивалентности, существенно отличаются от аналогичных значений в статьях самого Д.С. Лихачева. То есть на текстах малого объема (не более 1 тыс. словоупотреблений) наблюдаются существенные различия значений критерия в текстах автора детерминант от текстов других авторов.

Однако последующий анализ этого критерия для текстов большего объема (подкорпусов текстов) показывает, что средние значения критерия по подкорпусам Л.И. Абалкина (60,5%) и Д.А. Керимова (57,5%) не значительно отличаются от среднего по статьям Д.С. Лихачева – 70,3%.

Аналогичные результаты получены при изучении критерия №2 на детерминантах Л.И. Абалкина (табл. 6) и Д.А. Керимова (табл. 7).

Анализ табл. 6 и 7 показывает, что значения среднего процента детерминант, ЧМ которых попадают в зону

*Таблица 6*

**Процент детерминант Л.И. Абалкина, попавших в зону эквивалентности в произвольно выбранных статьях Л.И. Абалкина и подкорпусах Д.А. Керимова и Д.С. Лихачева (критерий № 2)**

| № п/п | Исследуемые тексты | Значение критерия № 2, рассчитанное по текстам различных авторов, % | | |
|---|---|---|---|---|
| | | Л.И. Абалкин | Д.А. Керимов | Д.С. Лихачев |
| 1 | Подкорпус/статья № 1 | 7 | 57 | 43 |
| 2 | Подкорпус/статья № 2 | 10 | 60 | 33 |
| 3 | Подкорпус/статья № 3 | 60 | 47 | 23 |
| 4 | Среднее значение по подкорпусам/статьям | 26 | 55 | 33 |

*Таблица 7*

**Процент детерминант Д.А. Керимова, попавших в зону эквивалентности в произвольно выбранных статьях Д.А. Керимова и подкорпусах Л.И. Абалкина и Д.С. Лихачева (критерий № 2)**

| № п/п | Исследуемые тексты | Значение критерия № 2, рассчитанное по текстам различных авторов, % | | |
|---|---|---|---|---|
| | | Д.А. Керимов | Л.И. Абалкин | Д.С. Лихачев |
| 1 | Подкорпус/статья № 1 | 18 | 23 | 5 |
| 2 | Подкорпус/статья № 2 | 27 | 18 | 27 |
| 3 | Подкорпус/статья № 3 | – | 32 | 27 |
| 4 | Среднее значение по подкорпусам/статьям | 23 | 24 | 20 |

эквивалентности, соизмеримы в статьях самих «авторов» детерминант и в исследуемых текстах других авторов.

Таким образом, использование данного критерия для установления авторства текста, как и для детерминант Д.С. Лихачева, не дает однозначного результата.

*Данный критерий – «процент детерминант Д.А. Керимова, попавших в зону эквивалентности» также как и первый не может быть использован для выявления индивидуального стиля автора.*

Критерий № 3. *Среднее отклонение ЧМ в исследуемом тексте от ЧМ*

*генеральной совокупности (в долях σ) у проявившихся детерминант.*

В качестве дополнительного критерия для установления авторства мы попробовали критерий, близкий по смыслу критерию 3.1, но учитывающий отклонение ЧМ в исследуемом тексте от среднеарифметического в корпусах текстов как в меньшую, так и в большую сторону.

Значения этого критерия для детерминант Д.С. Лихачева в статьях Л.И. Абалкина и Д.А. Керимова практически в 3 раза больше, чем в статьях самого Д.С. Лихачева за исключением статьи № 3 Л.И. Абалкина (табл. 8).

При увеличении объема исследуемого текста значения данного критерия для всех авторов сопоставимы. Разница между минимальным значением критерия в подкорпусе № 5 Л.И. Абалкина 2,76σ (строка 9, столбец 4 таблицы 13) и максимальным значением 2,46σ в статье № 1 Д.С. Лихачева (строка 1, столбец 3 таблицы 13) составляет 0,3σ.

Значения критерия № 3.2 для детерминант Л.И. Абалкина показаны в табл. 9, для детерминант Д.А. Керимова – в табл. 10.

Анализ табл. 9 и 10 показывает, что значение критерия в статьях «авторов» детерминант нестабильны и, вероятно, как и для критерия 1, в значительной мере зависят от объема анализируемого текста. Например,

*Таблица 8*

**Среднее отклонение ЧМ в исследуемом тексте от ЧМ генеральной совокупности (в долях σ) у проявившихся детерминант (критерий № 3)**

| № п/п | Наименование критерия | Значение критерия № 3, рассчитанное по текстам различных авторов, доли σ | | |
|---|---|---|---|---|
| | | Д.С. Лихачев | Л.И. Абалкин | Д.А. Керимов |
| 1 | Статья № 1 | 2,46 | 10,43 | 11,10 |
| 2 | Статья № 2 | 2,08 | 23,53 | 10,39 |
| 3 | Статья № 3 | 2,22 | 3,97 | – |
| 4 | Среднее значение по статьям | 2,25 | 12,6 | 10,75 |
| 5 | Подкорпус № 1 | – | 3,4 | 3,12 |
| 6 | Подкорпус № 2 | – | 3,16 | 3,65 |
| 7 | Подкорпус № 3 | – | 3,85 | 3,49 |
| 8 | Подкорпус № 4 | – | 3,35 | 3,01 |
| 9 | Подкорпус № 5 | – | 2,76 | 3,87 |
| 10 | Среднее значение по подкорпусам | – | 3,3 | 3,43 |

*Таблица 9*

**Среднее отклонение ЧМ в статьях Л.И. Абалкина и подкорпусах Д.А. Керимова и Д.С. Лихачева от ЧМ в подкорпусах Л.И. Абалкина (в долях σ) у проявившихся детерминант Л.И. Абалкина (критерий № 3)**

| № п/п | Исследуемые тексты | Значение критерия № 3, рассчитанное по текстам различных авторов | | |
|---|---|---|---|---|
| | | Л.И. Абалкин | Д.А. Керимов | Д.С. Лихачев |
| 1 | Подкорпус/статья № 1 | 8,11 | 2,72 | 3,40 |
| 2 | Подкорпус/статья № 2 | 5,01 | 2,92 | 4,11 |
| 3 | Подкорпус/статья № 3 | 1,79 | 3,83 | 3,80 |
| 4 | Среднее значение по подкорпусам/статьям | 4,97 | 3,16 | 3,77 |

*Таблица 10*

**Среднее отклонение ЧМ в статьях Д.А. Керимова и подкорпусах Л.И. Абалкина и Д.С. Лихачева от ЧМ в подкорпусах Д.А. Керимова (в долях σ) у проявившихся детерминант Д.А. Керимова (критерий № 3)**

| № п/п | Исследуемые тексты | Значение критерия № 3, рассчитанное по текстам различных авторов | | |
|---|---|---|---|---|
| | | Д.А. Керимов | Л.И. Абалкин | Д.С. Лихачев |
| 1 | Подкорпус/статья № 1 | 7,8 | 3,84 | 3,82 |
| 2 | Подкорпус/статья № 2 | 2,51 | 2,90 | 3,06 |
| 3 | Подкорпус/статья № 3 | – | 2,31 | 2,74 |
| 4 | Среднее значение по подкорпусам/статьям | 5,16 | 3,02 | 3,21 |

минимальное суммарное отклонение ЧМ детерминант Л.И. Абалкина в его статьях наблюдается для статьи № 3 – 1,79σ (объем словаря статьи – 1274 слова). Максимальное суммарное отклонение ЧМ детерминант Л.И. Абалкина в статье №1 – 8,11σ (объем словаря статьи 286 слов).

Аналогичные результаты наблюдаются на детерминантах Д.А. Керимова (см. табл. 10). Среднее отклонение ЧМ в статье № 1 равно 7,8σ (объем словаря статьи – 399 слов), в статье №2, объем словаря которой составляет 547 слов, среднее отклонение ниже в 3 раза – 2,51σ.

При этом, среднее отклонение ЧМ в подкорпусах других авторов – величина стабильная (разброс от 2,31σ до 4,11σ) и имеет средние значения ниже, чем в статьях «авторов» детерминант. Это характерно как для детерминант Л.И. Абалкина, так и для детерминант Д.А. Керимова.

*Таким образом, данный критерий не дает однозначных результатов и, следовательно, не может быть использован для установления авторства на текстах достаточно большого объема (около 50 тыс. словоупотреблений).*

Критерий № 4. *Средняя разница рангов детерминант в генеральной совокупности и в исследуемом тексте.* Для расчета значения критерия первоначально определяли ранг каждой детерминанты в подкорпусах Д.С. Лихачева и рассчитывали среднее значение ранга детерминанты каждого слова. Далее по каждому слову находили его ранг в исследуемом тексте (статье, подкорпусе) и рассчитывали разницу между рангом детерминанты в подкорпусе Д.С. Лихачева и рангом этой детерминанты в исследуемом тексте. Сумма модулей разниц рангов всех детерминант даст нам значение критерия.

Значения критерия № 4 для детерминант Д.С. Лихачева в исследуемых статьях и подкорпусах тек-

*Таблица 11*

**Средняя разница рангов детерминант Д.С. Лихачева в генеральной совокупности и в исследуемых текстах (критерий № 4)**

| № п/п | Наименование критерия | Значение критерия № 4, рассчитанное по текстам различных авторов | | |
|---|---|---|---|---|
| | | Д.С. Лихачев | Л.И. Абалкин | Д.А. Керимов |
| 1 | Статья №1 | 246 | 345 | 362 |
| 2 | Статья №2 | 208 | 363 | 335 |
| 3 | Статья №3 | 222 | 348 | – |
| 4 | Среднее значение по статьям | 225 | 352 | 349 |
| 5 | Подкорпус № 1 | – | 434 | 452 |
| 6 | Подкорпус № 2 | – | 360 | 381 |
| 7 | Подкорпус № 3 | – | 413 | 332 |
| 8 | Подкорпус № 4 | – | – | 346 |
| 9 | Подкорпус № 5 | – | – | 331 |
| 10 | Среднее значение по подкорпусам | – | 402 | 368 |

стов Д.С. Лихачева, Л.И. Абалкина и Д.А. Керимова показаны в табл. 11.

Анализ табл. 11 показывает, что значения данного критерия для детерминант Д.С. Лихачева в текстах Д.С. Лихачева отличаются от значений в текстах Л.И. Абалкина и Д.А. Керимова не менее чем на 100 ед.

Кроме того, значения критерия устойчивы как на отдельных статьях, так и на подкорпусах текстов. Например, среднее значение критерия №4 в статьях Л.И. Абалкина равно 352, а в его подкорпусах даже выше – 402. Аналогично по текстам Д.А. Керимова – среднее значение в статьях (349) сопоставимо со средним значением в подкорпусах – 368.

Значения критерия №4 для детерминант Л.И. Абалкина в исследуемых статьях Л.И. Абалкина и подкорпусах текстов Д.С. Лихачева и Д.А. Керимова показаны в табл. 12.

Анализ табл. 12 показывает, что средние значения этого критерия в статьях Л.И. Абалкина отличаются от значений, полученных на подкорпусах Д.С. Лихачева и Д.А. Керимова не менее чем в 2 раза. Исключение составляют данные, полученных по статье 3 Л.И. Абалкина (237) и подкорпусу №2 Д.А. Керимова (292). Вероятно, такие неявные отличия могут быть связаны с близкими тематиками исследуемых текстов.

Значения критерия №4 «Средняя разница рангов детерминант генеральной совокупности и в исследуемом тексте» для детерминант Д.А. Керимова в исследуемых статьях Д.А. Керимова и подкорпусах текстов Д.С. Лихачева и Л.И. Абалкина показаны в табл. 13.

*Таблица 12*

**Средняя разница рангов детерминант Л.И. Абалкина в генеральной совокупности и в исследуемых текстах (критерий № 4)**

| № п/п | Исследуемые тексты | Значение критерия № 4, рассчитанное по текстам различных авторов | | |
|---|---|---|---|---|
| | | Л.И. Абалкин | Д.А. Керимов | Д.С. Лихачев |
| 1 | Подкорпус № 1 | 144 | 514 | 715 |
| 2 | Подкорпус № 2 | 156 | 292 | 417 |
| 3 | Подкорпус № 3 | 237 | 505 | 533 |
| 4 | Среднее значение по подкорпусам | 179 | 437 | 555 |

*Таблица 13*

**Средняя разница рангов детерминант Д.А. Керимова в генеральной совокупности и в исследуемых текстах (критерий № 4)**

| № п/п | Исследуемые тексты | Значение критерия № 4, рассчитанное по текстам различных авторов | | |
|---|---|---|---|---|
| | | Д.А. Керимов | Л.И. Абалкин | Д.С. Лихачев |
| 1 | Подкорпус № 1 | 220 | 795 | 2172 |
| 2 | Подкорпус № 2 | 169 | 816 | 1052 |
| 3 | Подкорпус № 3 | 181 | 507 | 818 |
| 4 | Среднее значение по подкорпусам | 190 | 706 | 1347 |

Анализ табл. 13 показывает существенные различия между средней разницей рангов детерминант Д.А. Керимова в его статьях и подкорпусах других авторов. Максимальное значение для статей Д.А. Керимова – 220 ед. отличается от минимального – 507 в подкорпусе Л.И. Абалкина №3 более чем в 2 раза.

*Таким образом, проведенные исследования показали, что значение критерия № 4 – средняя разница рангов детерминант не зависит от объема анализируемого текста и критерий может использоваться для установления авторства по корпусам текстов различного объема.*

В результате проведенного исследования установлено, что: однозначные результаты для установления авторства текста получены для всех детерминант (Д.С. Лихачева, Л.И. Абалкина и Д.А. Керимова) только по критерию № 4 – «Средняя разница рангов детерминант генеральной совокупности и в исследуемом тексте».

Таким образом, полученные данные позволяют расширить подходы к оценке уровня самостоятельности обучающегося при написании курсовых, дипломных и других творческих письменных работ.

Возможно идеи, изложенные авторами в данной статье, о создании для каждого человека своего индивидуального лексического профиля, сегодня покажутся кому-то несколько утопическими. Но такой же утопией в свое время казались работы Жюля Верна о полетах на воздушном шаре, покорении подводных глубин и т. д., которые уже давно стали объективной реальностью.

**Литература**

1. Ивойлова И. Украденные мысли // Российская газета. № 4830 от 20.09.2009.

2. Алексеев П.М. Частотные словари. СПб.: Изд-во СПбУ, 2001.

3. Мухин М.Ю. Лексическая статистика и идиостиль автора: корпусное идеографическое исследование: на материале произведений М. Булгакова, В. Набокова, А. Платонова и М. Шолохова: Дис. ... д-ра филол. наук. Екатеринбург, 2011.

4. Мухин М.Ю. Лексическая статистика и концептуальная система автора: М. Булгаков, В. Набоков, А. Платонов, М. Шолохов. Екатеринбург: Изд-во Урал. ун-та, 2010.

5. Хмелев Д.В. Распознавание автора текста с использованием цепей А.А. Маркова. Вестник МГУ. Серия 9: Филология. 2000. № 2.

6. Кукушкина О.В., Поликарпов А.А., Хмелев Д.В. Определение авторства текста с использованием буквенной и грамматической информации // Проблемы передачи информации. Т. 37. 2001. № 2.

## PHILOSOPHY OF EDUCATION

**Fokina V.N.**, *PhD in Sociology*
**Lukyanova A.V.**, *PhD in Technical Sciences*
**Shirokova M.E.**, *PhD in Sociology*

### Analysis approach to setting criteria for constructing individual lexical profile to determine the authorship of texts

The article discusses the possibility of determining the authorship of texts on the basis of the criteria of the construction of individual lexical profile. The possibilities of making the "core" of the dictionary's methods of mathematical statistics and the possibility of evaluation of authorship on the basis of various statistical criteria.

***Key words:*** *plagiarism, the definition of authorship, individual lexical profile, idiostyle, frequency dictionary, word usage.*