

Оценка устойчивости во времени частотных словарей авторов в задачах идентификации текстов

77-30569/274006

12, декабрь 2011

Павлов Ю. Н., Тихомирова Е. А.

УДК. 004.021

МГТУ им. Н.Э. Баумана

pavlov@bmstu.ru

elizarti@mail.ru

Одним из известных методов идентификации авторов по написанным ими текстам является метод частотных словарей. Идея метода чрезвычайно проста и впервые упоминается в работе А.А. Маркова [1], опубликованной в 1913 году: по достоверно установленным текстам автора (достаточно большого объема) проводится статистический анализ. Результатом подобного анализа является оценка частоты появления отдельных слов в текстах автора (частотный словарь). Подобный анализ приводит к тому, что каждый автор будет представлен уникальным набором этих частот, позволяющим в принципе отличить работы одного автора от другого. При этом в качестве критерия близости текстов служит, так или иначе, введенное понятие «расстояния» между соответствующими текстами, которое и является, в конечном счете, определяющим фактором при вынесении решения о принадлежности текста тому или иному автору. И хотя ещё А.А. Марков указал в своей работе о не очень большой надёжности этого метода, метод применяется для идентификации авторских текстов и в настоящее время.

Исследователи, занимающиеся автороведческой экспертизой, под частотой появления того или иного слова понимают среднее значение, полученное по возможно большему числу публикаций автора, руководствуясь правилом «чем больше выборка, тем точнее оценка среднего». Однако, такое усреднение вызывает сомнение в его правомерности, поскольку совершенно очевидно, что все произведения автора написаны в разное время и временной интервал между публикациями порой составляет десятки лет. Например, свой первый роман «Бедные люди» 25-летний Ф.М. Достоевский опубликовал в 1846 году, а последний - «Братья Карамазовы» в 1881 году, когда ему было уже 60 лет. В некотором

смысле это два разных человека, поскольку с течением времени писатель (как и всякий другой человек) приобретает жизненный опыт, у него расширяется кругозор и образованность, изменяются вкусы и суждения и, наконец, сам словарь русского языка претерпевает за это время небольшие, но изменения.

Целью публикуемой работы является попытка оценить устойчивость авторских словарей во времени на примере двух авторов: Ф.М. Достоевского и М. Горького и попытка оценить степень их близости.

Перу Ф.М. Достоевского принадлежит 29 произведений, написанных им с 1846 по 1881 годы. Общий объём слов в них составляет 28481 слов. Максим Горький написал 31 произведение за 45 лет, общим объёмом 29164 слов. Число совпадающих слов, встречающихся и у Ф.М. Достоевского, и у М. Горького равно 15605.

Попытка автоматизации поиска слов в текстах наталкивается на проблему изменчивости слов, вследствие употребления их в различных формах. Имеется в виду спряжения, склонения, единственные и множественные числа. Решение этой проблемы возможно при условии выделения в словах их основ. При этом под основой слова понимаем ее неизменяемую часть. Существует, так называемый, алгоритм Портера для выделения основ в словах английского языка, существуют аналоги и для других языков, в том числе для русского. Однако в силу того, что английский и русский языки являются языками разного типа, алгоритм Портера не может учесть всей сложности и особенности русского языка. В связи с этим авторами статьи была произведена модификация вышеуказанного алгоритма Портера для получения более достоверных результатов, что подтверждено соответствующим моделированием. Программа статистической обработки написана на языке C# и позволяет определить, в конечном итоге, относительные частоты появления слов в текстах автора.

Под относительной частотой появления выбранного слова понимается отношение абсолютного числа появления слова в литературном произведении исследуемого автора к общему числу слов в конкретном произведении этого автора.

На рисунке 1 и 2 соответственно, показаны относительные частоты 100 наиболее часто встречающихся слов в шести произведениях Ф.М. Достоевского и М. Горького.

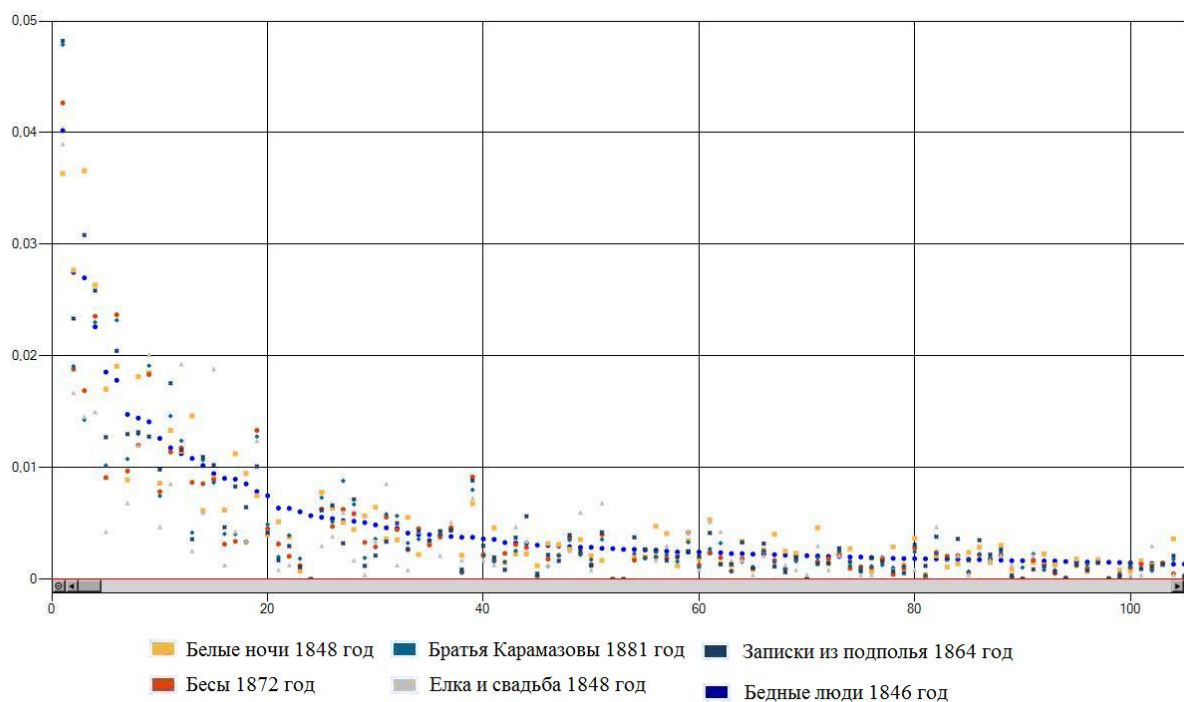


Рис.1. Относительные частоты слов в произведениях Ф.М. Достоевского

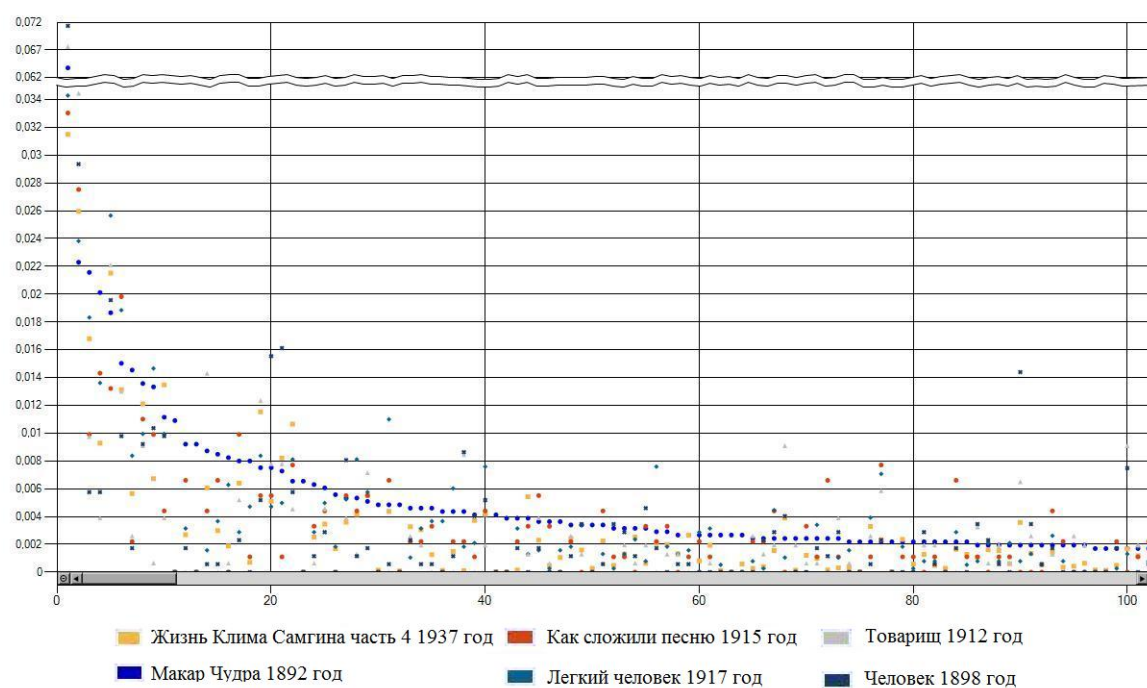


Рис. 2. Относительные частоты слов в произведениях М. Горького

На рисунках заметны значительные различия относительных частот одинаковых слов в произведениях этих авторов.

В качестве примера на рисунке 3 (а, б) соответственно показано, как менялись во времени относительные частоты появления слов «человек» и «общий» в произведениях Ф.М. Достоевского. Здесь по горизонтальной оси отложены в хронологическом порядке названия произведений, а по вертикальной - относительные частоты появления этих слов. Там же показаны средние значения частот этих слов. Видим, что разброс относительно среднего очень большой и убывает со временем.

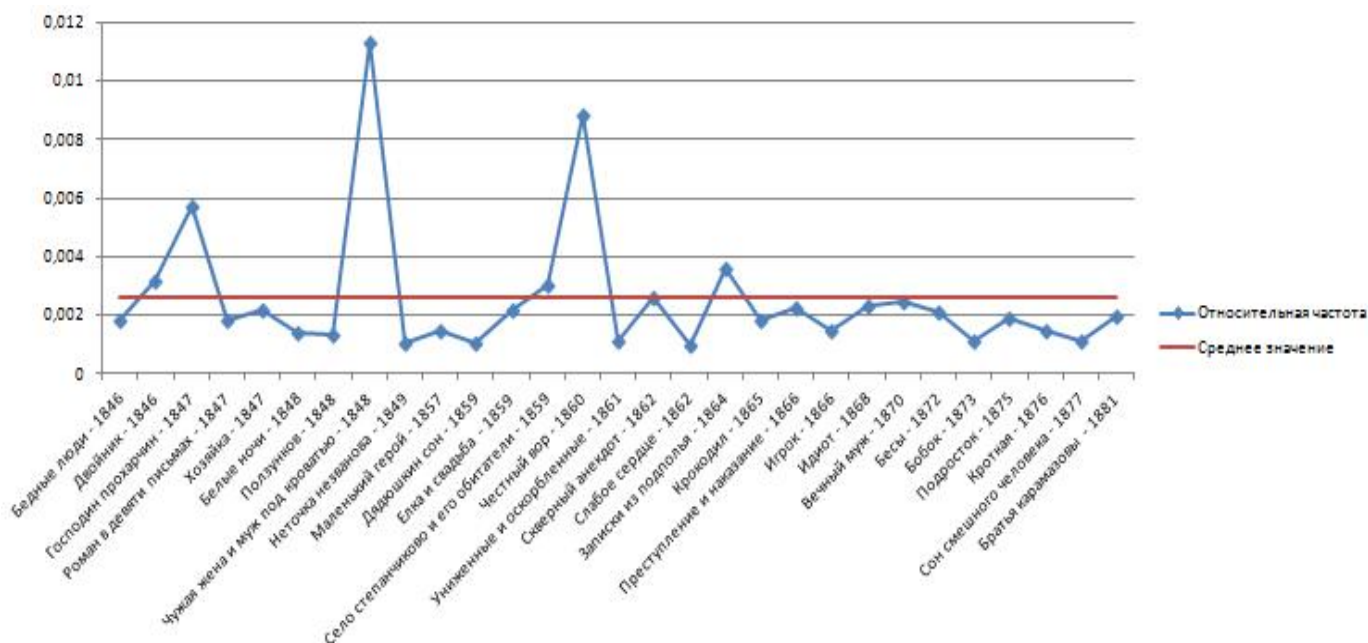


Рис. 3а. Относительная частота появления слова «человек» в романах Ф.М. Достоевского в хронологическом порядке

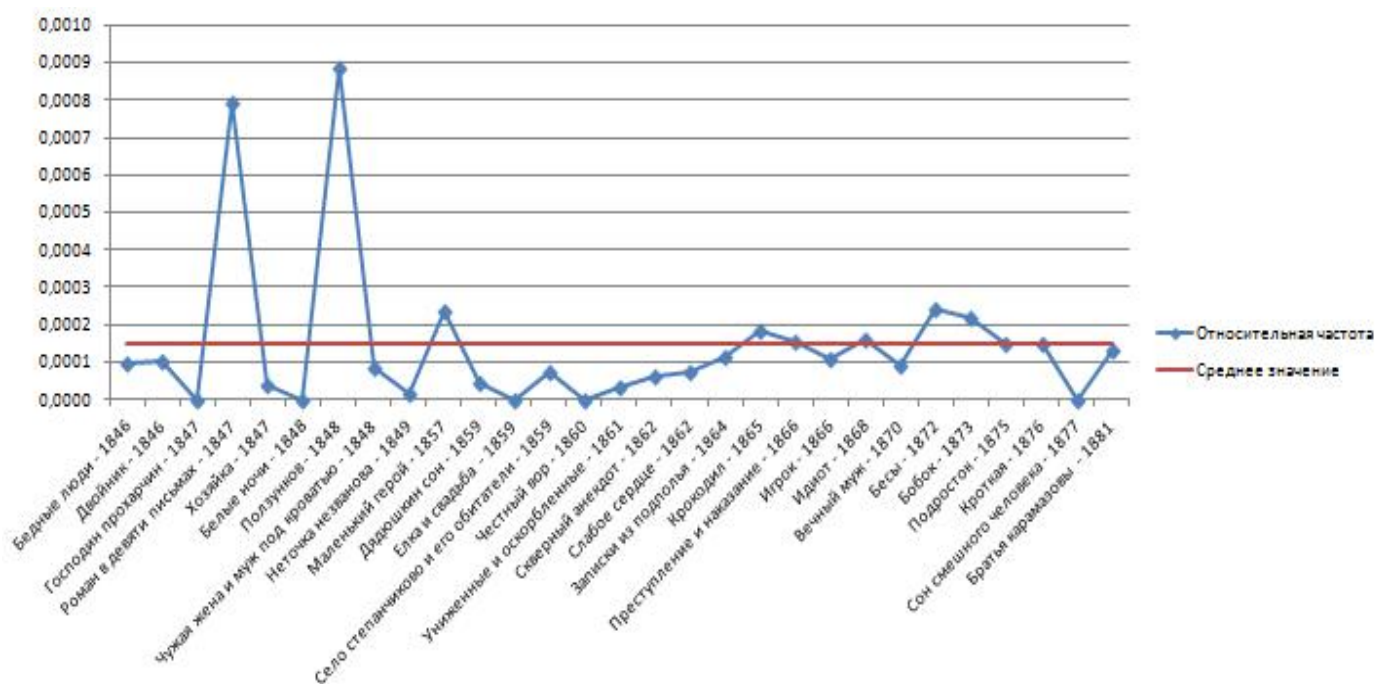


Рис. 3б. Относительная частота появления слова «общий» в романах Ф.М. Достоевского в хронологическом порядке

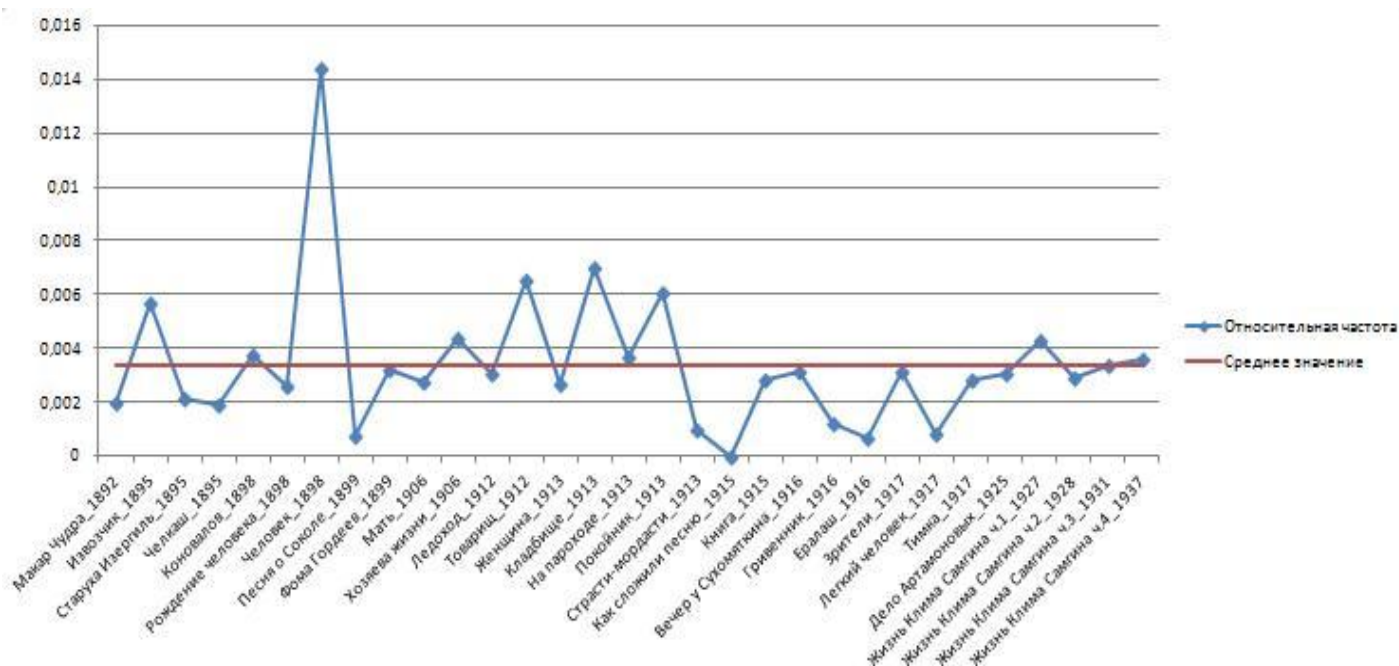


Рис. 4а. Относительная частота появления слова «человек» в романах М. Горького в хронологическом порядке

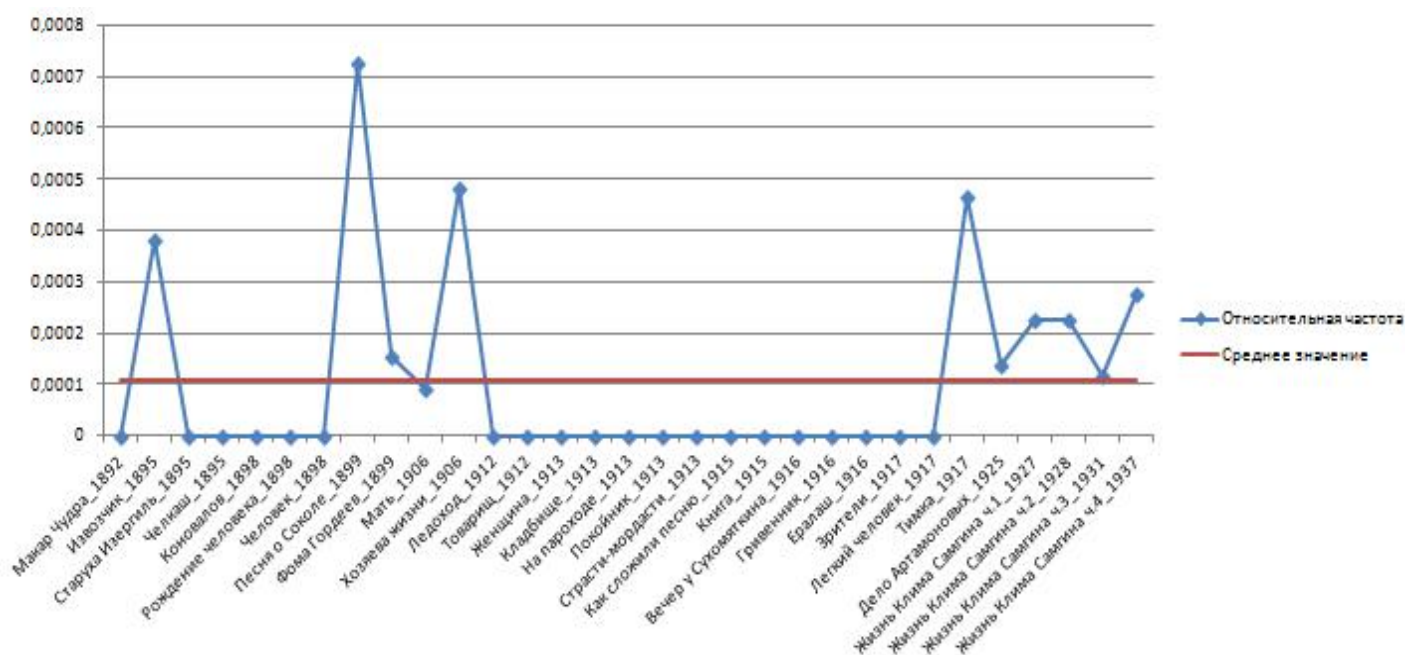


Рис. 46. Относительная частота появления слова «общий» в романах М. Горького в хронологическом порядке

На рисунке 4 (а, б) показаны относительные и средние частоты для слов «человек» и «общий» в произведениях М. Горького. Видим, что разброс относительно среднего значения так же очень большой. И хотя средние значения относительных частот слов «человек» и «общий» в произведениях этих авторов различны, уместно задать вопрос, различимы ли они в статистическом смысле при таких больших дисперсиях, то есть не принадлежат ли их средние к одной совокупности. При этом коэффициенты вариации, определяемые по следующему соотношению:

$$C_x = \frac{s_x}{\bar{x}} \cdot 100\%$$

где $\bar{x}$ - среднее значение выборки,

$$s_x = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n(n-1)}}$$

составляют для слов «человек» и «общий» в произведениях Ф.М. Достоевского 16 % и 26 %, соответственно, а для произведений М. Горького -14 % и 31 %, что значительно больше критической оценки 5 %.

Для проверки гипотезы о равенстве средних значений относительных частот слов «человек» и «общий» в произведениях Ф.М. Достоевского и М. Горького воспользуемся критерием согласия Крамера-Уэлча. При этом нулевая гипотеза H_0 звучит так: «средние значения выборок слова «человек» в произведениях Ф.М. Достоевского и М. Горького совпадают». Затем проверим эту же гипотезу для слова «общий».

Для этого воспользуемся следующим соотношением:

$$T_{\text{экс}} = \frac{|\bar{x} - \bar{y}|}{\sqrt{n_1 \cdot D_x + n_2 \cdot D_y}} \cdot \sqrt{n_1 \cdot n_2}$$

Здесь обозначено:

D_1 – дисперсия относительных частот появления слова «человек» в произведениях Ф.М. Достоевского;

D_2 – дисперсия относительных частот появления этого же слова в произведениях М. Горького;

n_1 - объем выборки в произведениях Ф.М. Достоевского;

n_2 - объем выборки в произведениях М. Горького.

Вычислим экспериментальную величину критерия $T_{\text{экс}}$ и сравним её с табличным $T_{\text{т}}$ для уровня значимости $\alpha = 0,1$.

По следующему соотношению:

$$D = \frac{\sum_i (x_i - \bar{x})^2}{n - 1}$$

вычисляем дисперсии для слова «человек». Дисперсия для Ф.М. Достоевского $D_1 = 5,4\text{E}-06$, для М. Горького $D_2 = 7,0\text{E}-06$. Объем выборки для Ф.М. Достоевского $n_1 = 29$, для М. Горького $n_2 = 31$.

Подставив эти значения в соотношение для $T_{\text{экс}}$, получим $T_{\text{экс}}$ для слова «человек» равно 1,27. Поскольку табличное значение [3] $T_{\text{т}} = 1,65$ больше $T_{\text{экс}}$, то нулевая гипотеза не отвергается.

Для слова «общий» соответствующие вычисления дают следующие результаты:

$$D_1 = 4,2E-08;$$

$$D_2 = 3,4E-08;$$

$$T_{\text{экс}} = 0,80;$$

$$T_t = 1,65 \text{ для того же уровня значимости } \alpha = 0,1.$$

Поскольку табличное значение [3] критерия T_t больше экспериментального $T_{\text{экс}}$, то нулевая гипотеза о равенстве средних значений частоты появления слова «общий» в произведениях Ф.М. Достоевского и М. Горького не отвергается, а это значит что слова «человек» и «общий» в частотных словарях Ф.М. Достоевского и М. Горького статистически неразличимы.

Число совпадающих, то есть присутствующих, как в произведениях Ф.М. Достоевского, так и М. Горького составляет 15605 слов. Для каждого из этих слов был проведен подобный статистический анализ: по критерию согласия Крамера-Уэлча проверялась гипотеза «среднее значение частоты появления данного слова в произведениях Ф.М. Достоевского совпадает со средним значением частоты появления того же слова в произведениях М. Горького» для уровней значимости $\alpha = 0,05$, $\alpha = 0,1$, $\alpha = 0,2$.

В таблице 1 представлены несколько слов из 15605, проверенные по критерию согласия Крамера-Уэлча, и различимость их при разных уровнях значимости. При этом 59 % неразличимых слов у Ф.М. Достоевского и М. Горького при уровне значимости критерия согласия Крамера-Уэлча $\alpha = 0,05$ находятся в первых 10 тысячах слов словаря Шарова [2].

Таблица 1

Слова, проверенные по критерию согласия Крамера-Уэлча

Слово	$\alpha = 0,05$	$\alpha = 0,1$	$\alpha = 0,2$
фонарь	неразлично	различно	различно
птица	неразлично	различно	различно
девица	неразлично	различно	различно
добрый	неразлично	различно	различно
лес	неразлично	различно	различно

иногда	неразлично	неразлично	различно
старуха	неразлично	неразлично	различно
чай	неразлично	неразлично	различно
потолок	неразлично	неразлично	различно
соседка	неразлично	неразлично	неразлично
сердце	неразлично	неразлично	неразлично
буфет	неразлично	неразлично	неразлично

Таким образом, в соответствии с вышеизложенным, можно сказать, что при уровне значимости критерия согласия Крамера-Уэлча $\alpha = 0,05$ число статистически неразличимых слов в произведениях Ф.М. Достоевского и М. Горького составляет 85 %, при уровне значимости $\alpha = 0,1$ число неразличимых слов составляет 78 %, а для уровня значимости $\alpha = 0,2$ число неразличимых слов равно 65 %.

По-видимому, такой большой процент неразличимых слов и служит причиной малой ненадёжности метода частотных словарей. Более того, утверждение многих авторов, что при идентификации необходимо брать 5-10 тысяч первых слов частотного словаря русского языка, можно подвергнуть сомнению, поскольку, как это видно из представленных результатов большое число неразличимых слов как раз лежит в этих пределах (по крайней мере, у авторов, тексты которых подвергались статистической обработке авторами статьи).

Литература:

1. Марков А.А. Пример статистического исследования над текстом "Евгения Онегина", иллюстрирующий связь испытаний в цепь // Изв. Имп. акад. наук. 1913. № 3. С. 153-162.
2. Шаров С.А. Частотный словарь русского языка //artint.ru. URL: <http://www.artint.ru/projects/frqlist.asp>.
3. Алгоритмика, статистика и теория вероятностей//matstats.ru URL. <http://matstats.ru/kramer.html>

Frequency dictionaries time stability rating in texts identification

77-30569/274006

12, December 2011

Павлов Ю. Н., Tihomirova E.A.

Bauman Moscow State Technical University

pavlov@bmstu.ru

elizarti@mail.ru

The article presents the time stability assessment results of F. Dostoevsky's and M. Gorky's vocabularies. The resulting statistical characteristics suggest that frequency dictionaries are unstable. That is why informativity of identifying the authors of the frequency dictionary is poor. The authors also show that a large percentage of words found by these authors are statistically indistinguishable. This fact, in the authors' opinion, will help to improve authorship examination quality.

Publications with keywords: [frequency dictionary](#), [authorship examination](#), [statistical estimation](#)

Publications with words: [frequency dictionary](#), [authorship examination](#), [statistical estimation](#)

Reference

1. Markov A.A., An example of statistical study on the text of the "Eugene Onegin", illustrating the connection test in the chain, Izv. Imp. akad. Nauk - Proceedings of the Imperial Academy of Sciences 3 (1913) 153-162.
2. Sharov S.A., <<http://www.artint.ru/projects/frqlist.asp>>.
3. <<http://matstats.ru/kramer.html>>.