ИНФОРМАТИКА

УДК 81’322::811.222.8::519.25

А.А.Косимов

ОЦЕНКА ЭФФЕКТИВНОСТИ ИСПОЛЬЗОВАНИЯ УНИГРАММ ПРИ ИДЕНТИФИКАЦИИ ТЕКСТА

Худжандский политехнический институт
Таджикского технического университета им. акад. М.С.Осими

(Представлено академиком АН Республики Таджикистан З.Д.Усмановым 23.11.2016 г.)

Исследованы возможности критерия однородности Н.В.Смирнова и его модификатора распознавать автора текста по частотности буквенных униграмм.

Ключевые слова: *таджикский язык, униграмма, частотность, статистика, эффективность.*

Первые исследования по частоте встречаемости букв в текстах на таджикском языке были предприняты в [1] и связывались с определением “наилучшей” раскладки букв на компьютерной клавиатуре. В [2] изучалась взаимосвязь классической и современной таджикской литературы путем сопоставления распределений частотностей букв различных произведений. Полученный результат – статистическая неразличимость объектов исследований – позволил, с одной стороны, сформировать общую картину частотности букв, свойственную таджикскому языку, а с другой стороны, подсказал необходимость использования иных методов анализа текстов, основанных на частотности букв.

В настоящей работе в качестве исследовательского инструмента тестируются критерий однородности Н.В. Смирнова о принадлежности двух независимых выборок одному закону распределения [3] и сопутствующий ему метод, использованный в [4,5].

1. Информация о коллекции текстов. Выборка текстов, предназначенная для исследовательских целей, так же как и в [2], была представлена произведениями А.Фирдауси, Дж.Руми, С.Айни, М.Турсунзода и Л.Шерали.

2. Обработка литературных данных. В качестве единиц измерения текста используются буквенные униграммы. Напомним, что таджикский алфавит А состоит из 35 букв. Одновременно с А будем использовать также и расширенный алфавит $A^* = A + \text{“}пробел\text{”}$, которому соответствует расширенный набор из 36 униграмм, позволяющий учитывать дополнительную информацию как относительно произведений, так и относительно самих униграмм.

Процесс обработки литературных данных реализуется в 3 этапа.

Этап 1. Вычисления частот встречаемости букв (с учётом и без учёта пробелов) по отдельности для всех упомянутых в п.1 произведений (авторов – 5, у каждого – по 2 произведения, итого – 10 текстов).

Адрес для корреспонденции: *Косимов Абдунаби Абдурауфович. 735700, Республика Таджикистан, Худжанд, ул. Ленина, 226, Худжандский политехнический институт Таджикского технического университета. E-mail: abdunabi_kbtut@mail.ru.*

На основе полученных данных строятся функции распределения $F_{i,n}(\lambda)$ частотности λ униграмм (с учётом и без учёта пробела) для авторских текстов, помечаемых индексом i, $i = 1,\ldots,10$.

Этап 2. Вычисление по формуле

$$D_{n,m}^{(i,j)} = \sup_{\lambda} |F_{i,n}(\lambda) - F_{j,m}(\lambda)| \tag{1}$$

максимального значения взаимного отклонения функций распределения частотностей униграмм i-го и j-го произведений, а по ним и статистики $S_{n,m}^{(i,j)}$ Н.В.Смирнова (см. [3]) по формуле:

$$S_{n,m}^{(i,j)} = \sqrt{\frac{nm}{n+m}}\, D_{n,m}^{(i,j)}, \tag{2}$$

где *n* и *m* – суммарные количества униграмм в i- ом и j - ом произведениях.

Одновременно с (1), (2) применяется другой способ обработки данных. Именно, вместо (1) используется формула

$$d_{p}^{(i,j)} = \sup_{\lambda} |F_{i,p}(\lambda) - F_{j,p}(\lambda)| \tag{3}$$

максимального значения взаимного отклонения функций распределения частотностей униграмм i-го и j-го произведений, а вместо (2) – формула

$$s_{p}^{(i,j)} = \sqrt{\frac{p}{2}}\, d_{p}^{(i,j)}, \tag{4}$$

причём в (3) и (4) *p* – число униграмм (*p* = 35 – для алфавита А и *p* = 36 – для алфавита A^*). Очевидно, что (3) и (4) являются упрощениями формул (1) и (2). Они тестируются здесь для того, чтобы получить представление о перспективности их использования.

Этап 3. Проверка нулевой гипотезы H_0 о том, что пара произведений (авторов), помеченных индексами i и j, являются выборками из одной и той же генеральной совокупности. Если речь идёт о произведениях, то они считаются однородными и могут принадлежать одному и тому же автору. Если же речь идёт об авторах, то их однородность понимается в смысле неразличимости соответствующих функций распределений частотностей униграмм.

Утверждение H_0 проверяется путём тестирования неравенства

$$S_{n,m}^{(i,j)} > K_{\alpha}, \tag{5}$$

в котором K_{α} – квантиль А.Н.Колмогорова уровня значимости α (= 0.05, 0.01, 0.001).

Если (5) выполняется для заданного уровня значимости α, то гипотеза H_0 об "однородности" i и j-объектов отвергается. Справедливой, с уровнем значимости $1-\alpha$, становится конкурирующая (альтернативная) гипотеза H_1, противоречащая H_0: i и j - объекты "не однородны".

Если имеет место неравенство

$$S_{n,m}^{(i,j)} < K_\alpha, \tag{6}$$

то принимается гипотеза H_0 об "однородности" i и j - объектов.

Аналогичные неравенства привлекаются для принятия решений по результатам применения формул (3) и (4) с необходимыми оговорками.

4. Результаты 1-го этапа о распределениях частотности униграмм отдельных произведений и их авторов здесь не показаны: определенное представление о них можно получить из [2], где приводятся списки частотностей униграмм таджикского языка с учётом и без учёта пробела.

Результаты 2-го этапа показаны в таблицах 1 и 2. В каждой ячейке даются два числа – верхнее число, подсчитанное по формулам (1) и (2), и нижнее, подсчитанное по формулам (3) и (4). Отметим, что в ячейках на главной диагонали представлена информация об отношениях между произведениями одного автора, а во всех других ячейках – информация об отношениях между произведениями различных авторов.

Результаты 3-го этапа связаны с проверкой нулевой гипотезы для уровня значимости α = 0.001, которому соответствует квантиль А.Н.Колмогорова со значением K_α = 1.95 . В этом случае для всех статистик Н.В.Смирнова из табл. 1, расположенных на главной диагонали, кроме ячейки [Айни "Одина"-Айни "Аҳмади Девбанд"], выполняется неравенство (6). Согласно критерию Н.В.Смирнова, это эквивалентно утверждению о том, что произведения одного и того же автора (кроме Айни) однородны, то есть подчиняются одному и тому же распределению частотностей униграмм, а произведения Айни оказываются неоднородными.

Таблица 1

Значения статистик $S_{n,m}^{(i,j)}$ Н.В.Смирнова и показателя $s_p^{(i,j)}$ для униграмм без учёта пробела

Авторы и произведения	Фирдауси Беж.&Ман.	Руми Дафтари Аввал	Турсунзода Садои Осиё	Шерали Катибаҳо	Айни Аҳмади Девбанд
Фирдауси Рустам ва Сӯҳроб	1.2321 0.0278	7.2647 0.1307	2.2834 0.1486	3.0900 0.1164	7.4661 0.1999
Руми Дафтари Дуввум	6.8363 0.1301	1.5539 0.0208	3.1030 0.1985	3.3708 0.1203	10.2665 0.2454
Турсунзода Ҳасани Аробакаш	4.1970 0.1112	5.6928 0.1296	1.8178 0.1210	2.6982 0.1084	4.4244 0.1335
Шерали Суханреза	3.4528 0.1215	3.5771 0.1161	1.2437 0.0876	1.1820 0.0548	5.4324 0.2066
Айни Одина	8.9235 0.1804	13.9363 0.2092	2.5448 0.1637	4.1592 0.1512	2.5127 0.0625

Что касается статистик Н.В.Смирнова, расположенных вне главной диагонали табл. 1, то для них, кроме ячейки [“Шерали Суханреза – Турсунзода Садои Осиё”], выполняется неравенство (5). Последнее эквивалентно утверждению о том, что произведения разных авторов не однородны, то есть принадлежат различным распределениям частот встречаемости униграмм, а указанные произведения Шерали и Турсунзода оказываются однородными.

Таким образом, критерий Н.В.Смирнова из 25 случаев лишь в двух случаях даёт ошибочный результат. Следовательно, эффективность применение критерия оценивается в 92%.

Идентификация авторства по формулам (3), (4) также оказывается вполне приемлемой, если вместо (5) и (6) воспользоваться их аналогами:

$$s_p^{(i,j)} > k, \tag{7}$$

$$s_p^{(i,j)} < k, \tag{8}$$

полагая при этом, что $k = 0.07$.

В таком случае с выполнением неравенства (7) будем связывать неоднородность i и j произведений, а с неравенством (8) – напротив, их однородность. Применяя это правило к нижнему ряду чисел табл. 1, устанавливаем, что (7) выполняется для всех ячеек, расположенных вне главной диагонали, а (8) нарушается лишь в одной ячейке на пересечении произведений Турсунзода. В рассматриваемом случае эффективность метода составляет 96 % и оказывается даже выше, чем для критерия Н.В.Смирнова.

Таблица 2

Значения статистик $S_{n,m}^{(i,j)}$ Н.В.Смирнова и показателя $s_p^{(i,j)}$ для униграмм с учетом пробела

Авторы и произведения	Фирдауси Беж.&Ман.	Руми Дафтари Аввал	Турсунзода Садои Осиё	Шерали Катибаҳо	Айни Аҳмади Девбанд
Фирдауси Рустам ва Сӯҳроб	1.2491 0.0258	6.7251 0.1108	2.8395 0.1708	2.7163 0.0939	8.2945 0.2044
Руми Дафтари Дуввум	6.4865 0.1131	1.2442 0.0152	3.8365 0.2269	3.5817 0.1173	10.4503 0.2303
Турсунзода Ҳасани Аробакаш	4.4826 0.1091	6.7612 0.1415	1.5287 0.0941	2.5825 0.0953	4.2710 0.1188
Шерали Суханреза	3.5615 0.1152	3.9132 0.1168	1.5990 0.1041	1.3403 0.0571	5.1573 0.1807
Айни Одина	9.5962 0.1782	14.7053 0.2029	2.2189 0.1320	3.9900 0.1331	2.4532 0.0563

Анализ табл. 2 (с учетом пробела), хотя и повторяет результаты анализа табл.1 (без учета пробела), тем не менее проявляет более высокую чувствительность в распознавании авторства текста: для ячейки “Айни Одина – Айни Аҳмади Девбанд” при учете пробела значение статистики Н.В.Смирнова понизилось со значения 2.5127 до 2.4532, а для ячейки “Шерали Суханреза – Турсунзода Садои Осиё”, напротив, повысилось с 1.2437 до 1.5990.

Что касается метода, основанного на формулах (7), (8), то его эффективность остается прежней, на уровне 96%. Единственная ошибка по-прежнему связана с ячейкой на пересечении произведений Турсунзода.

5. Заключение. Из полученных результатов извлекается следующее статистическое

Утверждение. *Критерий Н.В.Смирнова и его модификатор позволяют по частотности знаков таджикского алфавита (букв с пробелами и без них) с достаточно высокой степенью эффективности идентифицировать произведения поэтов классической таджикско-персидской литературы, а также различных авторов современной таджикской поэзии и прозы.*

Высказанное утверждение опирается на результаты обработки ограниченного по объёму материала, который, тем не менее, как по составу авторов, так и по списку использованных произведений представляет собой представительную выборку из генеральной совокупности изучаемой предметной области.

Сделанный вывод согласуется с аналогичными результатами для русского языка [6].

Поступило 23.11.2016 г.

Л И Т Е Р А Т У Р А

1. Усманов З.Д., Солиев О.М. Проблема раскладки символов на компьютерной клавиатуре. – Душанбе: Ирфон, 2010, 104 с.
2. Усманов З.Д., Косимов А.А. Частотность букв таджикской литературы. – Доклады Академии наук Республики Таджикистан, 2015, т.58, № 2, с. 112-115
3. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики.– М.: Наука: Гл. ред. физ-мат. литературы, 1983, 416 с.
4. Усманов З.Д., Косимов А.А. Частотность биграмм таджикской литературы. – Доклады Академии наук Республики Таджикистан, 2016, т.59, № 1-2, с. 28-32.
5. Усманов З.Д., Косимов А.А. О распознавании авторства таджикского текста. – Доклады Академии наук Республики Таджикистан, 2016, т.59, № 3-4, с. 114-119.
6. Романов А.С., Шелупанов А.А., Мещеряков Р.В. Разработка и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста. – Томск: В-Спектр, 2011, 188 с.

А.А.Қосимов

БАҲОДИҲИИ САМАРАБАХШӢ ИСТИФОДАБАРИИ УНИГРАММА ДАР МОНАНДКУНИИ МАТНҲО

Донишкадаи политехникии Донишгоҳи техникии Тоҷикистон ба номи М.С.Осими дар ш. Хуҷанд

Таҳқиқот оиди имконияти истифодабарии меъёри ягонагии Н.В.Смирнов ва ба шакли дигар даровардашудаи он, барои шинохти муаллифи матн бо басомади вохӯрии ҳарфҳои унограмма гузаронида шуд.

Калимаҳои калидӣ: *забони тоҷикӣ, унограмма, басомади вохӯрӣ, омор, самаранокӣ.*

A.A.Kosimov

EVALUATION OF UNIGRAMM USE EFFICIENCY FOR A TEXT IDENTIFICATION

Khujand's Polytechnic Institute of the M.S.Osimi Tajik Technical University

Efficiency of N.V.Smirnov's uniformity criterion and his modifier for identification of the author of a text by means of letter unigram frequencies are investigated.

Key words: *Tajik language, unigram, frequency, statistics, efficiency.*