ДОКЛАДЫ АКАДЕМИИ НАУК РЕСПУБЛИКИ ТАДЖИКИСТАН
2013, том 56, №3

ИНФОРМАТИКА

УДК 811.111::811.161.1::81’322

Академик АН Республики Таджикистан З.Д.Усманов, Г.М.Довудов*

О МНОГООБРАЗИИ СЛОВОФОРМНЫХ АНАГРАММ

Институт математики им.А.Джураева АН Республики Таджикистан,
Худжандский политехнический институт
Таджикского технического университета им. академика М.С.Осими

Посредством специального кодирования словоформ формируются многообразия анаграмм текстовых коллекций английского и русского языков. Получены статистические данные о количестве различных анаграмм с заданным числом элементов. Анонсирован ряд анаграмм наибольшей мощности.

Ключевые слова: *английский язык – русский язык – словоформа – кодирование – анаграмма – статистика.*

В статье [1] предложено так называемое $\alpha\beta$-кодирование словоформ, предназначенное, в частности, для выявления анаграмм, то есть таких подмножеств на множестве словоформ, элементы которых состоят из одного и того же набора букв. В [2] и [3] эффективность такого кодирования (в смысле возможности осуществления взаимно однозначного соответствия между словоформами и их $\alpha\beta$-кодами) изучена статистическими методами для английского, литовского, русского и таджикского языков, а также искусственного языка эсперанто.

Настоящая работа нацелена на получение предварительной информации о множестве анаграмм, а также их мощности и элементном составе для конкретных текстовых коллекций английского и русского языков.

1. Схема обработки текстовой информации как на английском, так и на русском языке состоит из двух этапов:

- построения списка различных словоформ с частотами их встречаемости;

- кодирования полученных словоформ и формирования списка различных кодов с частотами их встречаемости.

Первый этап – широко известный и не нуждается в пояснении. Что касается второго этапа, то в нём речь идёт о применении $\alpha\beta$-кодирования к произвольной словоформе $W = "\alpha_1\alpha_2 \ldots\alpha_n"$, состоящей из букв α_k $(k = \overline{1, n})$. Результатом преобразования служит цепочка $CW = "\alpha_{s1}\alpha_{s2} \ldots\alpha_{sn}"$ из тех же самых букв, что и в W, но упорядоченных по алфавиту (пример: W = *ласка* $\rightarrow$ CW = *ааклс*).

Адрес для корреспонденции: *Усманов Зафар Джураевич. 734063, Республика Таджикистан, г. Душанбе, пр. Айни, РТ, 299/1, Институт математики АН РТ. E-mail: zafar-usmanov@rambler.ru*

Очевидно, преобразование $W \to CW$ присваивает один и тот же $\alpha\beta$ - код всем словоформам из одной и той же анаграммы и потому разбивает множество $\{W\}$ всех словоформ на непересекающиеся подмножества анаграмм.

2. Коллекция английских текстов “British National Corpus”, доступная по адресу http://ske.fi.muni.cz и подвергнутая исследованию, содержит 96 052 598 словоупотреблений. Обработка данных в соответствии с первым этапом выявила 545 999 словоформ (различных).

3. Коллекция русских текстов представлена Russian web corpus и национальным корпусом, доступными соответственно по адресам http://ske.fi.muni.cz и http://www.ruscorpora.ru.

Размер первого корпуса – 144 413 607 словоупотреблений, среди которых обнаружено 922 284 словоформ (различных). Размер второго корпуса – 187 972 357 словоупотреблений[1], содержит 837 516 словоформ (различных).

4. Итоговые результаты, полученные после выполнения второго этапа обработки коллекций текстов, представлены в табл. 1. В ней m – мощность анаграммы, то есть количество словоформ, входящих в её состав. Как установлено, её значения для British National Corpus варьирует в пределах от 1 до 18, для Russian web corpus – от 1 до 19 и для национального корпуса русского языка – от 1 до 15. При m = 1 мы имеем дело со словоформами, которые находятся во взаимно однозначном соответствии со своими $\alpha\beta$-кодами и лишь для удобства названы “тривиальными” анаграммами. Для прочих значений m ($m \geq 2$) речь идёт уже о реальных анаграммах.

В первом столбце табл. 1 символами n_1, n_2 и n_3 обозначены числа различных анаграмм заданной мощности m (или же число различных кодов анаграмм мощности m) соответственно для British National Corpus, Russian web corpus и национального корпуса русского языка. Таким образом, последующие данные представляют табличную зависимость $n = n(m)$.

Таблица 1

m	1	2	3	4	5	6	7	8	9	10
n_1	429206	29822	7424	3037	1570	982	378	256	145	124
n_2	732608	48916	12365	4907	2497	1336	664	380	265	144
n_3	687841	42921	9920	3562	1638	790	397	205	118	64

Продолжение

m	11	12	13	14	15	16	17	18	19	Итого
n_1	66	30	13	18	5	6	2	2	0	473 086
n_2	108	62	33	18	14	10	3	3	1	804 334
n_3	34	14	9	5	3	0	0	0	0	747 521

[1] На самом-то деле первоначальный размер корпуса состоял из 190 827 174 словоупотреблений. Он был сокращён до размера 187 972 357 за счёт удаления знаков препинания, иностранных слов, чисел и дат, не имеющих отношения к поиску анаграмм.

5. Зависимость *n = n (m)* для трёх рассмотренных текстовых коллекций в графическом виде показана на рис 1. Характерной особенностью поведения всех трёх кривых является строго монотонное уменьшение значений n_k (k = 1, 2, 3) при увеличении значений m.

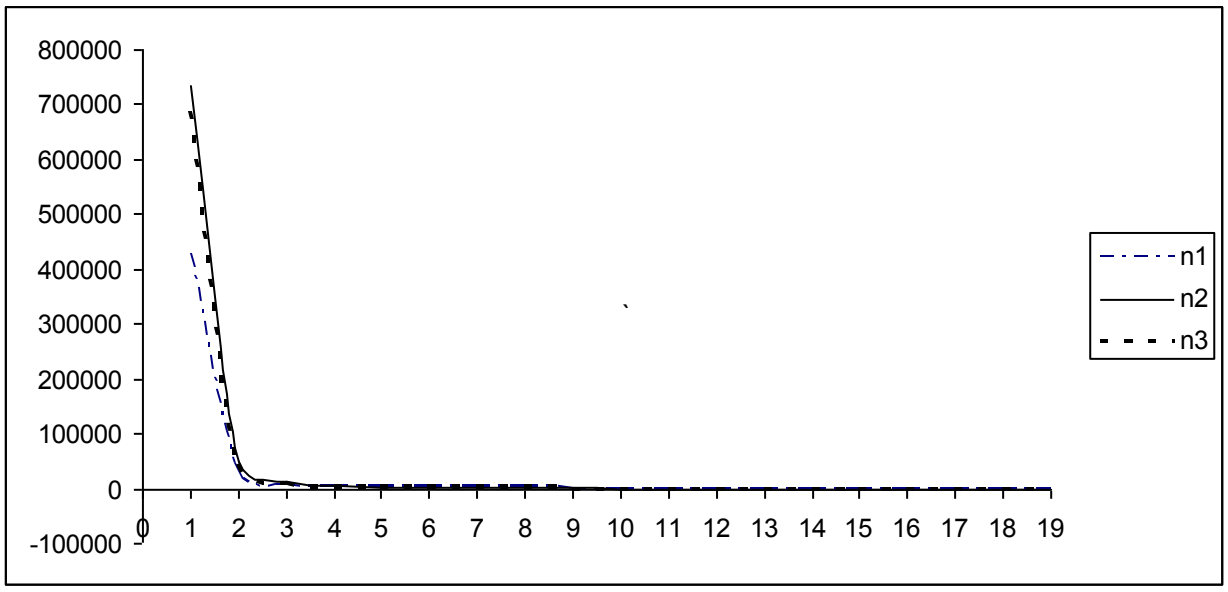

Рис 1. Зависимости n = n (m) для трёх текстовых корпусов.

6. Примеры анаграмм наибольшей мощности $m_{\max}$ приведены в последующих трёх таблицах. В них в первом столбце указывается один и тот же $\alpha\beta$ - код всех тех словоформ, которые составляют анаграмму мощности $m_{\max}$ (см. второй столбец) для соответствующего корпуса текстов. В третьем столбце представляется список словоформ со своими частотами встречаемости. Суммарная частота словоформ предшествует их списку.

British National Corpus

Код анаграммы	$m_{\max}$	Состав словоформ в анаграммах
aelst	18	33460(least 30303; tales 1228; steal 869; slate 614; stale 384; lates 11; sleat 10; astle 8; stael 7; altes 7; tesla 6; leats 4; aslet 3; teals 2; taels 1; stela 1; aelst 1; laste 1)
aerst	18	16676(rates 11553; tears 3949; stare 1107; aster 23; tares 18; artes 6; estar 3; resta 3; aerts 2; tesar 2; tersa 2; staer 2; setar 1; treas 1; stear 1; taser 1; resat 1; astre 1)

Russian web corpus

Код анаграммы	$m_{\max}$	Состав словоформ в анаграммах
аднор	19	22467(народ 22186; ронда 52; дорна 37; андро 34; нодар 30; родна 25; дрона 22; радон 17; доран 16; норда 12; ардон 11; ондар 6; надор 5; одран 3; нардо 3; норад 2; ондра 2; андор 2; дарон 2)

Национальный корпус русского языка

Код анаграммы	$m_{\max}$	Состав словоформ в анаграммах
иорст	15	1176(сирот 855; ситро 72; строи 45; стори 35; сорит 33; сотри 33; рости 25; истор 22; остри 11; тирсо 11; ортис 10; истро 9; сорти 8; ристо 4; стиро 3)
ааимнр	15	8291(марина 7592; ранами 231; нарами 191; армани 96; римана 39; ариман 35; мариан 28; амиран 22; ранима 13; рамина 12; арнима 9; манира 9; имрана 5; армина 5; мирана 4)
аирст	15	2841(расти 2598; сатир 101; истра 34; сарти 20; ритас 16; траси 13; тарси 13; стари 11; истар 7; тирас 6; ситар 5; арист 5; риста 5; тирса 4; тасир 3)

7. Заключение. В списках анаграмм могут присутствовать ошибочные словоформы, причина появления которых обуславливается, очевидно, ошибками, имеющими место в текстовых массивах корпусов, подвергнутых обработке. Даже несмотря на то, что подобных ошибок может оказаться незначительное количество, они будут искажать истинную картину описания множества анаграмм того или иного естественного языка. Именно в этой связи к корпусам и коллекциям текстов, предназначенным для выявления множества анаграмм, следует предъявлять особо высокие требования к недопустимости ошибок в написании слов.

Поступило 04.01.2013 г.

ЛИТЕРАТУРА

1. Усманов З.Д. – ДАН РТ, 2012, т.55, № 7, с. 545-548.
2. Усманов З.Д., Нормантас В. – ДАН РТ, 2012, т.55, № 8, с. 622-625.
3. Усманов З.Д., Нормантас В. – Материалы 16 научно-практ. семинара "Новые информационные технологии в автоматизированных системах". – М., 2013, с. 287 - 292.

З.Ҷ.Усмонов, Г.М.Довудов*

ОИДИ МАҶМӮИ АНАГРАММАИ КАЛИМАҲО

Институти математикаи ба номи А.Ҷураев Академияи илмҳои Ҷумҳурии Тоҷикистон,
****Донишкадаи политехникии Донишгоҳи техникии Тоҷикистон ба номи академик М.С.Осими дар ш. Хуҷанд***

Бо истифодаи кодиронии махсуси калимаҳо маҷмӯи анаграммаҳои матнҳои англисӣ ва русӣ тартиб дода мешавад. Оиди шумораи анаграммаҳои гуногун ва маҷмӯи калимаҳои ба он мутааллиқ маълумотҳои оморӣ ба даст оварда шудааст. Якчанд анаграммаҳои шумораи зиёди элементдошта, оварда шудаанд.

Калимаҳои калидӣ: *забони англисӣ – забони русӣ – шаклҳои калима (парадигмаҳо) – кодиронӣ - анаграмма – омор.*

Z.D.Usmanov, G.M.Dovudova*

ON A SET OF WORDFORM ANAGRAMS

A.Juraev Institute of Mathematics, Academy of Sciences of theRepublic of Tajikistan,
****Khujand's Politecnic Institute of the M.S.Osimy Tajik Technical University***

Thanks to special coding of wordforms, the sets of anagrams to English and Russian corpora are exhaustively described. Statistical data on the number of different anagrams with a specified number of items are received. Some anagrams with the highest number of wordforms are presented for discussion.

Key words: *English – Russian – wordform – coding – anagram – recognition – statistics.*