БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ ACADÉMIE BULGARE DES SCIENCES
ИЗВЕСТИЯ НА МАТЕМАТИЧЕСКИЯ ИНСТИТУТ
BULLETIN DE L'INSTITUT DE MATHÉMATIQUES
Том (Tome) XIII

ИЗСЛЕДВАНИЯ НА НЯКОИ ЧЕСТОТИ В ПИСМЕНИЯ БЪЛГАРСКИ ЕЗИК

Петър Бърнев, Димитър М. Добрев и Румяна Киркова

Изследванията на честоти на различни образувания (букви, разделители, двойки от букви, тройки от букви и пр.) в писмения език могат да имат различни приложения. От особен интерес са такива изследвания за лингвистиката, както и за някои приложни области: при набора на печатарски текстове, при определяне на оптималното разположение на символите в различни клавиатури, за създаване на оптимални стенографски символи и др.

Първи изследвания за честотата на буквите в писмения български език са публикувани в [1], където се привежда таблица, получена чрез непосредствено преброяване на една извадка от около 20% от романа „Под игото“ на Иван Вазов.

В настоящата работа се излагат някои от резултатите, получени чрез обработка с автоматична сметачна машина на целия текст на същия роман, обхващащ 696 081 символа. За целта бе създадена методика и бяха съставени съответни програми за машината „Минск-2“ в Математическия институт с Изчислителен център на БАН. Разработената методика дава възможност да се пресмятат честоти на последователности от n символа, $n=1, 2, 3, \ldots$, срещани в текста. При това текстът може да се разглежда като непрекъснат или като разчленен на отделни думи.

За да могат да се извършат необходимите изчисления, а също така за да се подготви текстовата информация за удобно използване за други изследвания, бяха направени редица програми със стандартен характер. Ще спомена само някои от основните процедури.

Текстът се перфорира във вид, в който е бил напечатан. След въвеждането му в машината той се линеаризира програмно, т. е. подрежда се в един единствен ред, като се премахват всички символи, имащи значение само за напечатването — между думите се оставя само по един разделител, който може да бъде шпация, препинателен знак или символ за нов ред, и се премахват тиретата за пренасяне. Така обработеният текст се подлага на серия от програми, в които се пресмятат честотите на различни образувания от символи (т. е. букви и разделители) и букви. При това се отчитат не само абсолютните честоти, а и честотите в зависимост от позицията на съответните образувания. Отделна програма определя честотите на буквите, които се намират на последна позиция в

Таблица 1

Честоти в проценти на букви в различни позиции

Бук	I	II	III	IV	V	VI	VII	VIII	IX
а	12,55	10,36	1,34	21,15	6,81	14,91	13,92	4,25	25,81
б	1,81	1,49	4,41	0,84	2,24	1,83	0,70	6,40	0,07
в	4,37	3,62	6,03	1,62	4,56	3,88	5,08	4,86	2,92
г	1,83	1,51	3,27	0,99	3,04	2,44	0,95	2,55	0,22
д	3,62	2,98	6,99	1,62	4,57	4,06	2,15	5,32	0,96
е	9,09	7,51	2,10	14,45	6,78	10,65	8,72	1,61	18,28
ж	0,73	0,61	0,55	0,16	1,48	0,83	0,68	0,17	0,11
з	2,27	1,87	3,55	2,68	3,69	1,55	1,20	2,38	1,08
и	8,12	6,69	7,28	7,95	5,95	6,70	9,21	4,80	14,75
й	0,85	0,70	0,01	0,02	4,06	0,31	0,79	0,09	2,47
к	3,76	3,08	6,10	0,74	5,57	4,09	4,05	8,91	1,25
л	3,16	2,60	1,37	2,91	5,84	2,84	4,11	1,22	1,12
м	2,51	2,07	4,49	1,30	4,16	1,97	1,72	5,49	1,78
н	6,14	5,06	9,68	1,24	4,69	4,82	7,97	7,76	2,04
о	9,19	7,60	5,27	16,02	4,02	10,58	12,13	7,27	14,45
п	2,79	2,28	8,96	0,73	2,55	1,48	0,68	5,35	0,13
р	4,50	3,72	2,33	7,06	7,31	4,45	4,16	3,89	0,59
с	4,59	3,78	11,83	1,69	4,64	2,99	2,27	7,82	1,66
т	7,10	5,86	7,12	4,65	4,53	6,83	7,92	12,87	4,37
у	1,77	1,46	1,60	3,59	1,73	1,58	1,51	0,58	1,33
ф	0,17	0,14	0,24	0,04	0,14	0,51	0,09	0,53	0,04
х	0,90	0,75	0,94	0,14	1,13	0,70	1,13	1,61	0,64
ц	0,52	0,43	0,39	0,03	0,51	0,74	0,48	0,65	0,10
ч	1,52	1,26	2,26	0,42	1,52	2,49	1,90	1,90	0,12
ш	1,26	1,04	0,31	0,03	2,34	0,82	2,60	0,30	0,50
щ	0,64	0,53	0,87	0,30	1,41	0,48	0,41	0,51	0,10
ъ	1,88	1,55	0,02	5,02	1,85	1,93	1,54	0,00	0,06
ь	0,01	0,01	0,00	0,01	0,00	0,01	0,02	0,01	0,01
ю	0,13	0,11	0,12	0,19	0,09	0,04	0,17	0,61	0,06
я	2,20	1,81	0,54	2,38	2,77	3,44	1,70	0,24	3,97

Таблица 2

Честота в проценти на препинателните знаци

Препинателен знак	Честота спрямо всички символи	Честота спрямо препинателните знаци
Интервал	15,074	74,070
Точка	1,913	9,398
Запетая	1,582	7,776
Въпросителна	0,192	0,942
Удивителна	0,184	0,905
Двоеточие	0,010	0,510
Кавички	0,007	0,339
Тире	0,416	2,043
Нов ред	0,740	3,636
Точка и запетая	0,006	0,319
Отворена скоба	0,00065	0,031
Затворена скоба	0,00067	0,032

думите. Основните затруднения и тук са свързани с необходимостта от избягване на голяма памет, която би била необходима при директното изчисляване. Така например за пресмятане на честотата на четворките би била необходима памет от 810 000 позиции само за съхраняване на окончателния резултат. За да се съкрати тази памет, бяха съставени специални програми, които предварително дават груба оценка за честотата на отделните съчетания, така че в подробните изчисления да могат да се вземат пред вид само най-често срещаните съчетания.

Таблица 3

Честота в проценти на най-срещаните двойки от символи в текста (Знакът „—“ означава произволен символ, който не е буква)

Двойка	Честота	Двойка	Честота	Двойка	Честота
А—	4,55	ПО	0,79	ЕД	0,48
Е—	3,22	ТЕ	0,78	ЧЕ	0,48
И—	2,59	—М	0,78	ВИ	0,48
О—	2,36	—Б	0,78	ЛЕ	0,46
—С	2,08	Т—	0,77	ТИ	0,45
НА	1,71	НО	0,77	ВЕ	0,44
—Н	1,68	ОТ	0,73	Й—	0,43
—П	1,56	НЕ	0,73	ЛА	0,42
ТО	1,45	ОВ	0,73	АЗ	0,42
—И	1,24	СЕ	0,73	ОЙ	0,41
—Т	1,24	ЕН	0,67	АВ	0,41
—Д	1,22	ПР	0,64	—Р	0,41
ТА	1,20	ЗА	0,63	СИ	0,39
КА	1,09	—З	0,63	ВО	0,39
—К	1,05	ШЕ	0,62	ДЕ	0,39
—В	1,04	РЕ	0,59	—Ч	0,39
РА	0,99	ИТ	0,58	—Е	0,39
—О	0,97	—Г	0,57	МА	0,39
АТ	0,94	РИ	0,56	БЕ	0,39
ВА	0,91	ЛИ	0,52	ОС	0,39
СТ	0,91	В—	0,52	АШ	0,38
ДА	0,85	АН	0,51	ДО	0,38
НИ	0,84	ГО	0,50	ИЗ	0,37
КО	0,80	ЕТ	0,49	ХА	0,37

С помощта на създадените програми бяха съставени следните таблици:

1) честота на буквите спрямо всички символи;

2) честота на буквите спрямо общия брой букви;

3) честота на буквите от фиксирана позиция в думата спрямо броя на буквите от тази позиция (разглеждат се от 1 по 20 позиции);

4) честота на последните букви на думите спрямо общия брой на думите в текста;

5) честота на главните букви спрямо общия им брой;

6) честота на буквите с ударение;

7) честота на препинателните знаци;

8) честота на двойките букви, влизащи в думите;

Таблица 4

Честота в проценти на най-срещаните двойки от букви вътре в думите

Двойка	Честота	Двойка	Честота	Двойка	Честота
НА	2,66	ТИ	0,69	ЯТ	0,45
ТО	2,22	ВЕ	0,69	ЕЛ	0,45
ТА	1,85	ЛА	0,65	ТР	0,44
КА	1,71	АЗ	0,64	НЯ	0,42
РА	1,52	ОЙ	0,63	КИ	0,41
АТ	1,44	АВ	0,62	ОД	0,41
ВА	1,41	СИ	0,61	ИЕ	0,40
СТ	1,39	ВО	0,61	СК	0,40
ДА	1,31	ДЕ	0,61	ВЪ	0,38
НИ	1,29	МА	0,60	ЛО	0,37
КО	1,24	БЕ	0,59	АС	0,37
ПО	1,24	ОС	0,59	ЕР	0,37
ТЕ	1,20	АШ	0,58	ЕС	0,37
НО	1,18	ДО	0,58	ОБ	0,37
ОТ	1,13	ИЗ	0,57	МО	0,37
НЕ	1,13	ХА	0,57	ИЛ	0,36
ОВ	1,13	ОР	0,57	ДН	0,36
СЕ	1,13	ЕШ	0,57	ПА	0,35
ЕН	1,02	ИН	0,57	СЪ	0,33
ПР	0,99	АК	0,57	ОК	0,33
ЗА	0,98	ИЯ	0,55	ИС	0,32
ШЕ	0,96	АМ	0,55	ИВ	0,32
РЕ	0,90	АД	0,53	ЪТ	0,32
ИТ	0,89	АЛ	0,53	ИЧ	0,31
РИ	0,87	ЪЛ	0,52	РЪ	0,31
ЛИ	0,83	РО	0,51	МУ	0,31
АН	0,78	АР	0,51	ГЕ	0,31
ГО	0,76	ОГ	0,51	БО	0,31
ЕТ	0,76	МЕ	0,49	ТВ	0,31
ЕД	0,75	ДИ	0,48	АХ	0,30
ЧЕ	0,75	ОЛ	0,48	ЧА	0,29
ВИ	0,75	МИ	0,48	БА	0,29
ЛЕ	0,72	ГА	0,46	ИК	0,29

9) честота на двойките символи от текста;
10) честота на двойките букви от фиксирана позиция;
11) честота на тройките букви, влизащи в думите;
12) честота на тройките символи от текста;
13) честота на четворките букви, влизащи в думите;
14) честота на четворките символи от текста;
15) честота на дължините на думите;
16) честота на някои избрани, често срещащи се думи.

Някои от получените резултати се дават със съответни обяснения по-долу. Приложените таблици са на базата на целия материал. В хода на работата бяха направени и изследвания над по-малки извадки, които показаха, че се получава стабилизация на резултатите. В приведените таблици всички знаци с изключение на последните са надеждни.

В табл. 1 се дават честотите на буквите, както следва: в първа колона спрямо всички букви, във втора колона спрямо всички символи; трета до седма колона съдържат честотите на буквите от 1-ва до 5-а позиция; в осма колона са дадени честотите на главните букви, а в деветa колона — честотите на последните букви в думите.

Таблица 5

Честота в проценти на най-срещаните тройки от букви вътре в думите

Тройка	Честота	Тройка	Честота	Тройка	Честота
АТА	0,440	АХА	0,115	ПРА	0,087
ИТЕ	0,352	ИНА	0,110	ДНА	0,087
ЕТЕ	0,252	ИЯТ	0,110	ВИЯ	0,085
АШЕ	0,236	КАЗ	0,106	РАТ	0,080
ЕТО	0,219	РАД	0,102	СКА	0,080
КАТ	0,216	СТР	0,102	ИКА	0,080
ТОЙ	0,205	РЕД	0,101	АВИ	0,080
ОВА	0,194	ЛЕД	0,101	ДИН	0,080
ЕНИ	0,192	ТОВ	0,101	АРИ	0,079
СТА	0,184	ПОД	0,100	ВАН	0,078
ОТО	0,172	СТИ	0,100	ЕДИ	0,078
ОСТ	0,171	ОГН	0,099	ЛЕН	0,077
ПРЕ	0,171	СТО	0,098	ТРА	0,076
ПРИ	0,164	ЯНО	0,097	ИЦА	0,076
АТО	0,158	АНА	0,095	ДЕН	0,075
РАЗ	0,158	ЗНА	0,095	САН	0,075
НОВ	0,143	ЕСТ	0,095	ЧЕР	0,075
ЕДН	0,142	КОЛ	0,094	ОКО	0,074
НИЕ	0,139	СТВ	0,094	ВАТ	0,074
АВА	0,137	ГНЯ	0,093	ЕЧЕ	0,073
ВАШ	0,128	НЯН	0,091	НИТ	0,073
НАТ	0,127	РАВ	0,091	ТАН	0,072
КАК	0,125	НИЯ	0,090	СКИ	0,072
ОВЕ	0,124	АЗА	0,089	ЕТЕ	0,072
ПРО	0,124	КОЙ	0,088	ВОР	0,071
БЕШ	0,122	ВЪР	0,088	ОГА	0,071
АНИ	0,115	ТЕЛ	0,087	ГЛЕ	0,071

В табл. 2 се дават честотите на препинателните знаци.

В табл. 3 се дават честотите на първите 72 най-често срещани двойки символи спрямо всички двойки символи от текста (знакът „—“ в таблицата означава произволен разделител).

В табл. 4 се привеждат честотите на първите 99 най-често срещани двойки от букви, влизащи в думи, спрямо общия брой такива двойки.

В табл. 5 и 6 се дават честотите на първите 99 най-често срещани тройки (респективно четворки) от букви, влизащи в думи.

Накрая табл. 7 съдържа честотите на думите с дължина от 1 до 14 букви.

С помощта на пресметнатите честоти на символи, двойки символи и тройки символи се получават [2] следните стойности (в десетични единици), характеризиращи ентропията на писмения български език:

$H_1 = 1{,}277$ — при предположение, че появите на отделните символи са независими;

Таблица 6

Честота в проценти на най-срещаните четворки от букви вътре в думите

Четворка	Че тота	Четворка	Честота	Четворка	Честота
КАШЕ	0,0356	НЕНО	0,0044	АНАШ	0,0024
КАТЕ	0,0282	ДАТЕ	0,0043	КАВЕ	0,0024
НАТЕ	0,0267	ВЕТА	0,0042	ЛЕТА	0,0023
КАХА	0,0228	ЛАНО	0,0042	НАТО	0,0023
ОСТЕ	0,0162	РЕНО	0,0042	НАШЕ	0,0022
ВАТЕ	0,0139	ТАВО	0,0042	АНИЯ	0,0022
РАВО	0;0123	ЕНИЯ	0,0041	ДАТИ	0,0021
ТАТЕ	0,0122	ВАНО	0,0041	ЗАНА	0,0021
ЛЕНО	0,0110	ДАНО	0,0038	ВЕНЕ	0,0021
ЛЕДЕ	0,0106	РЕДЕ	0,0037	ЛЕНЕ	0,0020
ТАНЕ	0,0104	РАЗА	0,0036	ЕНЕШ	0,0020
РАТЕ	0,0093	РАВИ	0,0035	КАНЕ	0,0020
ДЕНО	0,0089	РАТО	0,0035	ТЕНО	0,0020
ДЕШЕ	0,0084	ВЕДО	0,0034	ОСТИ	0,0019
ВАНИ	0,0082	ОТАВ	0,0034	ЕНОЙ	0,0019
ЕНОВ	0,0081	ТАНИ	0,0034	ДЕТА	0,0018
ИНАШ	0,0079	РАНО	0,0034	МАНО	0,0018
ОСТО	0,0076	ЕНАШ	0,0033	МАТА	0,0018
ЧЕШЕ	0,0074	ЧЕТИ	0,0033	ЗАШЕ	0,0018
ВЕТИ	0,0070	ЧЕТЕ	0,0033	КАКО	0,0018
ТАНО	0,0065	ВЕНО	0,0032	НЕНА	0,0018
НЕШЕ	0,0063	НАВЕ	0,0032	ЕТАВ	0,0017
ВАНЕ	0,0058	АНОЙ	0,0031	РАТИ	0,0017
ЛАВЕ	0,0058	ДЕТИ	0,0031	НЕТИ	0,0017
ДАВЕ	0,0052	РАНЕ	0,0030	ТАМА	0,0017
ТАКО	0,0052	ВАНА	0,0029	КАНО	0,0016
ЧЕНО	0,0052	НАКО	0,0028	ДАНЕ	0,0016
АНОВ	0,0052	СЕТО	0,0027	ТЕНЕ	0,0016
МАТЕ	0,0051	НЕНЕ	0,0026	НАВИ	0,0015
ЛАТЕ	0,0051	МАНЕ	0,0026	РАТА	0,0015
АНИТ	0,0049	ВЕТЕ	0,0025	ЗАТЕ	0,0014
ТАВЕ	0,0049	НАКА	0,0025	КАНИ	0,0014
АНЕШ	0,0045	РАКО	0,0025	РЕНЕ	0,0014

Таблица 7

Честота на думите с различна дължина

Брой на символите в думата	Честота спрямо общия брой на думите
1	0,088
2	0,218
3	0,081
4	0,124
5	0,118
6	0,112
7	0,098
8	0,068
9	0,039
10	0,024
11	0,015
12	0,009
13	0,004
14	0,002

$H_2 = 1{,}004$ — при предположение, че вероятността за поява на даден символ зависи само от предидущия символ;

$H_3 = 0{,}946$ — при предположение, че вероятността за поява на даден символ зависи само от предидущите два символа.

ЛИТЕРАТУРА

1. Penkov, B., A. Obretenov, B. Sendov, T. Kirpikova, T. Djoukanov. Frequences of letters in written Bulgarian. C. R. Acad. Bulg. Sci., **15** (1962), No. 3, 243—244.
2. Яглом, А., И. Яглом. Вероятност и информация. С., 1961.

Постъпила на 26. II. 1970 г.

ИССЛЕДОВАНИЕ НЕКОТОРЫХ ЧАСТОТ В ПИСЬМЕННОМ БОЛГАРСКОМ ЯЗЫКЕ

Петр Бырнев, Димитр М. Добрев и Румяна Киркова

(Резюме)

Приводятся результаты исследования частоты букв (табл. 1), знаков препинания (табл. 2), групп из двух, трех и четырех букв (табл. 3, 4, 5 и 6) и частоты слов разной протяженности (табл. 7). Исследования проведены на базе полного текста романа „Под Игом“ Ивана Вазова.

A STUDY OF SOME FREQUENCES IN THE WRITTEN BULGARIAN

Petâr Bârnev, Dimitâr M. Dobrev, Rumjana Kirkova

(Summary)

The paper comprises the results from a study of the frequences of occurrence of letters (Table 1), punctuation marks (Table 2), combinations of two, three and four letters (Tables 3, 4, 5 and 6) as well as those of words of different length (Table 7). The study is based on the whole text of the novel "Under the Yoke" by Ivan Vazov.