

ЭНТРОПИЯ СЕРБОХОРВАТСКОГО ЯЗЫКА : ИССЛЕДОВАНИЕ НА ОПУСЕ П.П. НЕГОША

ДР. ЗДРАВКО ВУКЧЕВИЧ

Университет Черногории
Подгорица, Джорджа Вашингтона, бб. Подгорица, Черногория
e-mail: mzvukcevic@yahoo.com, <http://www.pmf.ac.me/>

Ключевые слова: информатика, теория информации, математическая статистика и лингвистика, энтропия

Аннотация: Появление теории информации и быстрое развитие информационных технологий позволили вычисление функциональной нагрузки фонем и энтропии конкретной языковой системы. Язык сводится к серии символов, которые считаются и анализируются как последовательность повторения монограммных, диграммных, триграммных и нграммных секвенций.

В настоящей работе рассчитана энтропия поэтического текста нулевого, первого, второго и третьего уровня (H_0 , H_1 , H_2 и H_3). Расчёт энтропии выполнен на основе собраний сочинений, а именно 1.368.098 графем текста. В обследовании были приняты во внимание 30 символов сербохорватского кириллического письма однозначно связанных с тем же числом фонем, расстояние между словами (бленк) и расстояние в полусложных словах (-).

1 ЭНТРОПИЯ И ИНФОРМАЦИЯ

Энтропия – это понятие, использующееся в термодинамике, которое вошло в теорию информации, чтобы обозначить уровень дезинтеграций в процессе информации.

Снижение неопределенности (*uncertainty reduction*), т.е. уменьшение числа возможных альтернатив, представляет собой ключ для квантифицирования перенесенной информации. Если число альтернатив большое, то информация, отклоняющая эти альтернативы, будет давать больше, будет побольше в количественном смысле (Рис.1). Отклонение неизвестности, т.е. уменьшение альтернатив всегда является бинарным, причем неопределенность в каждой ситуации превращения неизвестности в информацию уменьшается на половину. Согласно основному уравнению Шеннона-Вивера, количество информации какой-нибудь системы является альтернативой

$$H = -\sum p_i \cdot \log_2 p_i$$

т.е. оно тождественно сумме вероятностей появления каждой из возможных альтернатив.

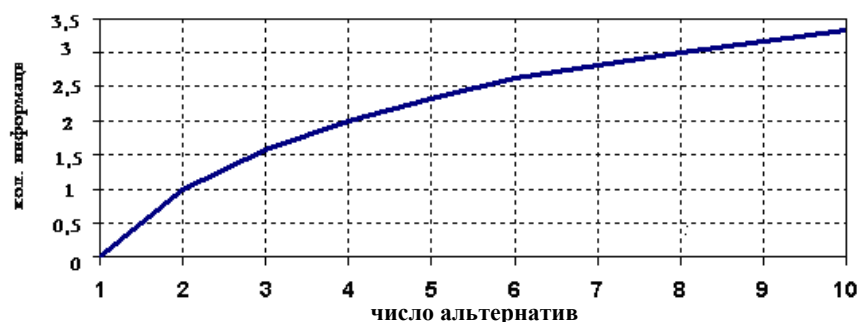


Рис. 1. Соотношение альтернатив и количество информации

Энтропия всегда обратно пропорциональна вероятности результата. Она максимальна, когда все вероятности в совокупности одинаковы, т. е. когда совокупность наименее упорядочена, а она минимальна и равна нулю, когда одна из вероятностей имеет числовое значение 1 и когда остальные равны нулю. Если энтропия определена для совокупности знаков, вероятности которых неопределены, то такая энтропия всегда больше энтропии взаимообусловленных символов. При открытии одного знака, мы принимаем информацию о последующих знаках, так что и неопределенность совокупности меньше. Такая энтропия обычно называется условной энтропией.

Основные положения теории информации могут применяться для разговорного языка или письменного текста с целью исследования статистических свойств конкретного языка. Анализируемый язык рассматривается как совокупность дискретных источников, взаимообусловленных и подчиненных принципам вероятности. Такой язык лишается всех своих индивидуальных свойств и семантических значений и переводится на язык символов с количественно-вероятностными характеристиками. Тем самым, текст утрачивает все качественные характеристики и представляется в виде групп знаков и секвенций монограммных, диграммных, триграммных и n -граммных структур. Таким образом, достигается возможность осуществления дескрипции языка путем методов математической статистики.

Поскольку коммуникации между людьми чаще всего осуществляются путем использования языка, то, с этой точки зрения, энтропию можно дефинировать и как меру неупорядоченности или неопределенности некоторой совокупности знаков, т.е. «меру информативности языковых единиц». Если языковые единицы независимы и не обусловлены своим местом в системе знаков, то их энтропия самая большая и равна единице ($H = 1$). Так как их выбор произволен, вероятность того, что необусловленные единицы формируют высказывание из n элементов, равна нулю, а также и количество информации, содержащейся в них.

2 ИССЛЕДОВАТЕЛЬСКИЙ КОРПУС

Исследования данной работы проводились на материале корпуса опуса литературных произведений Петра II Петровича Негоша, согласно совместному изданию группы югославских издателей 1974 года. Корпус произведений обследован интегрально, в его составе 242 646 слов, т.е. 1.368 098 знаков, включая расстояние между словами и дефис в полусложных словах. Исследования энтропии сербскохорватского языка до сих пор вычислялись на основании тридцати графем кириллического текста, связанного однозначно с одинаковым числом фонем, тогда как в последних работах равноправным 31 знаком считали и расстояние между словами. В этой работе впервые в исследовании энтропии сербскохорватского текста вводится и 32 знак – расстояние в полусложных словах.

2.1 Энтропия монограммы

При вычислении энтропии на нулевом уровне (H_0) для совокупности $n = 30$ элементов, т. е. 30 однозначно связанных фонем и букв, предполагалось, что их появление не взаимообусловлено, т. е. условие одинаковой возможности их появления, дает максимально возможную энтропию – $H_0 = H_{max}$. Она вычисляется по формуле $H = \log_2 n$ и составляет 4,90 бита/букву [5, 7].

Имея в виду, что расстояние между словами и полусложный знак (-) должны учитываться как равноправные знаки письменной или разговорной реализации, то тогда максимальная энтропия будет немного больше и будет составлять: $H_0 = H_{ma} = \log_2 32 = 5$ битов.

Однако, так как известно, что частотность появления некоторых знаков в совокупности графем неодинакова, то при вычислении энтропии на первом уровне $-H_1$ учитывается относительная частотность этих знаков. С помощью статистического анализа целого корпуса от 1.368.098 знаков (1.124.797 фонем или букв, 242 645 расстояний между словами и 659 полусложных расстояний), получены данные о частотности появления каждого из знаков в литературных произведениях П. П. Негоша (Таблица 1)

Ранг	Знак	Абсолютная частотность	Ранг	Знак	Абсолютная частотность
1	-	242645	17	L	29999
2	A	132427	18	G	22512
3	E	106457	19	B	20219
4	O	103413	20	[18453
5	I	100828	21	Z	15291
6	R	56891	22	^	10869
7	N	56247	23]	10372
8	S	54063	24	C	10258
9	U	53618	25	H	10073
10	T	50125	26	W	6479

11	V	43986	27	Q	6137
12	D	42707	28	@	5859
13	M	41230	29	\	3931
14	K	40974	30	F	1057
15	J	39523	31	-	659
16	P	30279	32	X	517

Таблица 1. Абсолютная частотность знаков корпуса

Реляцией $p_i = \frac{f_i}{N}$, в которой p_i обозначены вероятности появления, f_i – абсолютная частотность и N – полная совокупность знаков в тексте, получены суммы относительной частотности для знаков совокупности (Таблица 2)

Средняя информация для каждого знака языка Негоша вычислена по формуле Шеннона для среднего значения энтропии

$$H = -\sum_{i=1}^n p_i \log_2 p_i,$$

причем реляция $p_i \log_2$ определяет p_i парциальную энтропию каждого особого знака.

Ранг	Знак	Относительная частотность	$p_i \log_2 p_i$	Ранг	Знак	Относительная частотность	$p_i \log_2 p_i$
1	-	0,1774	0,4426	17	L	0,0219	0,1208
2	A	0,0968	0,3261	18	G	0,0165	0,0975
3	E	0,0778	0,2867	19	B	0,0148	0,0899
4	O	0,0756	0,2816	20	[0,0135	0,0838
5	I	0,0737	0,2773	21	Z	0,0112	0,0725
6	R	0,0416	0,1908	22	^	0,0079	0,0554
7	N	0,0411	0,1893	23]	0,0076	0,0534
8	S	0,0395	0,1842	24	C	0,0075	0,0529
9	U	0,0392	0,1832	25	H	0,0074	0,0522
10	T	0,0366	0,1748	26	W	0,0047	0,0366
11	V	0,0322	0,1594	27	Q	0,0045	0,0350
12	D	0,0312	0,1561	28	@	0,0043	0,0337
13	M	0,0301	0,1523	29	\	0,0029	0,0243
14	K	0,0299	0,1516	30	F	0,0008	0,0080
15	J	0,0289	0,1477	31	-	0,0005	0,0053
16	P	0,0221	0,1217	32	X	0,0004	0,0043

Таблица 2. Парциальная энтропия знаков корпуса

Полученная энтропия первого порядка H_1 для $n = 32$ знака составляет 4. 7908. По сравнению с максимальной энтропией H_0 , видно снижение энтропии H_1 ($H_0 > H_1$) до 0,21 бита по знаку, что меньше, чем ожидалось.

До сих пор, в исследованиях энтропии сербскохорватского языка печатного текста получены значительно меньшие значения энтропии, потому что и исследования проводились с меньшим числом знаков исследуемых образцов, что не должно быть единственной причиной наблюдаемых различий. Энтропия обычно вычислялась в совокупности 31 знака [5, 18], а очень часто и в совокупности только 30 фонем, т. е. букв [3, 7]. И энтропия языка Негоша, вычисленная для 30 фонем, составила бы только 4. 3429 бита (Приложение 1), что очень близко к ранее полученным результатам. Так, например, энтропия языка Андрича, выраженная через H_1 составляет 4,29 бита [15].

На основании уравнения Шеннона для среднего значения энтропии вычислено максимальное количество информации корпуса произведений Негоша, для $N = 1.368.098$, по формуле $I_0(N) = N \log_2 32 = 6.840.490$ бита.

Общее количество информации, содержащееся в отдельных знаках (x_i) корпуса, вычислены по формуле $I(x_i) = \log \frac{1}{p_i} = -\log p(x_i)$ (Таблица 3)

Ранг	Знак	Количество информации	Ранг	Знак	Количество информации
1	-	2,495	17	L	5,513
2	A	3,369	18	G	5,922
3	E	3,684	19	B	6,079
4	O	3,726	20	[6,211
5	I	3,762	21	Z	6,481
6	R	4,588	22	^	6,984
7	N	4,605	23]	7,040
8	S	4,662	24	C	7,059
9	U	4,673	25	H	7,079
10	T	4,772	26	W	7,734
11	V	4,957	27	Q	7,797
12	D	5,003	28	@	7,862
13	M	5,054	29	\	8,430
14	K	5,063	30	F	10,289
15	J	5,113	31	-	10,967
16	P	5,500	32	X	11,289

Таблица 3. Количество информации по знаку корпуса

На рисунке 2 показаны графики функций, аппроксимирующих данные по относительной частотности и количеству информации языка Негоша, которые ясно показывают обратную зависимость относительной частотности и количества информации и имеют определенный знак совокупности. Самое большое количество информации содержат знаки с наименьшей частотностью появления (полусложное расстояние, аффрикаты), тогда как наиболее частотные знаки (расстояние между словами и гласные) дают наименьшую информацию.

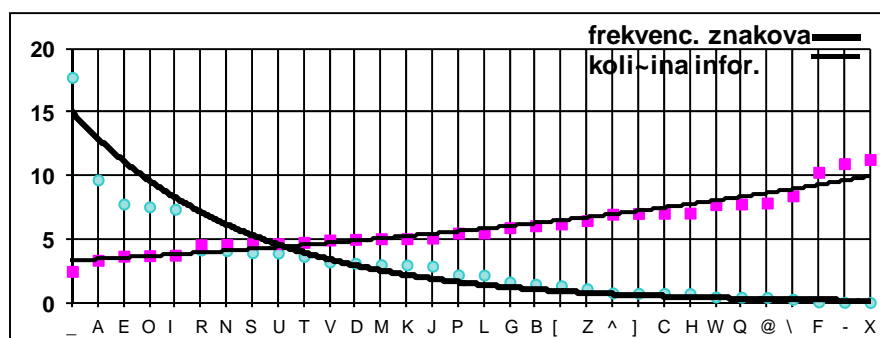


Рис. 2. Соотношение частотности и количества информации для совокупности 32 знаков

2.2. Энтропия диграммы

Чтобы определить точное количество информации, которую имеет вторая, третья ... n -фонема, проводится более сложный анализ текста, т. е. вычисляется энтропия на втором, третьем ... n -ом уровне - $H_2, H_3...H_n$. Это условные энтропии, так как фонемы, для которых вычисляется энтропия, находятся в прямой зависимости от фонем, предшествующих им. Итак, мера неопределенности n -ой фонемы зависит от влияния $n - 1$ фонемы. Условная энтропия определяется для первых трех фонем, тогда как определение энтропии для $n > 3$ трудно выводится из-за большого числа комбинаций букв, даже при помощи компьютера. Учитывая, что взаимовлияния фонем в тексте различного порядка, то и оценка энтропии какого-либо порядка проводится путем элиминации энтропии нижней n -граммы от высшей, согласно формуле

$$H_n = H_{n-gram} - H_{n-1gram}.$$

На уже известном корпусе из $N = 1.368.098$ знаков и $n = 32$ символов (30 букв и двух расстояний: полусложного и расстояния между словами), сформированы таблицы частотности комбинаций из двух слов (Приложение 1, таблица 2) и вычислены, с помощью формулы Шеннона, парциальные энтропии всех комбинаций, получена энтропия диграммы из 8, 10 битов, сколько составляет сумма парциальных энтропий всех 1.371.095 комбинаций из двух слов. С учетом того, что энтропия диграммы - H_2 включает энтропию первого порядка - H_1 , количество информации, которое дает вторая буква диграммы, получается путем элиминации H_1 , на основании приведенного уравнения. Таким образом, энтропия языка Негоша, вычисленная на втором уровне, составляет $H_2 = 3,31$ бита/букву.

Поскольку энтропия второго порядка до сих пор оценивалась для $n = 31$ знака сербского печатного текста, сравнения могут проводиться с учетом только этих результатов. Так, например, у Ширбеговича $H_2 = 3,81$, у Милошевича - 3,34, а Йовичича - 3,45. Хотя, из-за большого числа знаков ($n = 32$), а также предполагалось, что из-за большого числа комбинаций диграмм, энтропия могла ожидаться немного больше. Случилось то, что условная энтропия H_2 языка Негоша реально немного меньше, что, без сомнения, является следствием специфики поэтического

высказывания Негоша, т.е. природы фонем языка Негоша, которые в создании диграмм бывают более селективными, чем в современном литературном стандарте.

Снижение энтропии после второй фонемы в языке Негоша по отношению к H_0 значительно, и составляет 1,59 бита, или 37,8%, а это значит, что для той же меры выросла упорядоченность совокупности.

2.3. Энтропия триграммы

На основании статистического анализа всего корпуса собраний сочинений Негоша, вычислена частотность появления триграммы (Приложение 1, таблицы 4А - 4-), которая для $n = 32$ имеет 32^3 , или вместе 1.048 576 комбинаций из трех слов.

Так же как и при вычислении энтропии диграммы, вычислена и парциальная энтропия всех комбинаций диграммы, которая для всего корпуса, после применения формулы $H_n = H_{n-gram} - H_{n-1gram}$, составляет $H_3 = 2,97$, а это значит, что энтропия после третьей фонемы (буквы) по отношению к H_{max} уменьшилась на 2,03 бита или 40,6%, т.е. на тот же процент увеличилась упорядоченность совокупности, т.е. определенность высказывания.

Так как нам не известны работы, изучающие вычисление энтропии третьего уровня сербского текста, мы не имеем данных, с которыми можно было бы сравнить наши результаты. Правда, С. Йович в своей книге [13] только указывает на данные для энтропии третьего порядка, $H_3 = 2,73$ для совокупности из $n = 31$ знака, так что такая разница ожидаема и понятна.

Сравнение энтропии всех трех уровней языка Негоша с энтропией некоторых европейских языков [5] позывает сходства и различия, проявляющиеся с увеличением порядка, для которого аппроксимируется энтропия.

Энтроп.	Язык Негоша. $n = 32$	Сербский язык. $n = 31$	Английский $n = 27$	Русский $n = 32$	Немецкий $n = 27$	Французск. $n = 27$	Испанский $n = 27$
H_0	5 (1)	4,95 (1)	4,76 (1)	5 (1)	4,76 (1)	4,76 (1)	4,76 (1)
H_1	4,79 (0,96)	4,18 (0,84)	4,03 (0,85)	4,35 (0,86)	4,10 (0,86)	3,95 (0,83)	3,98 (0,84)
H_2	3,41 (0,68)	3,45 (0,70)	3,32 (0,70)	3,52 (0,71)	3,40 (0,71)	3,17 (0,67)	3,20 (0,67)
H_3	2,97 (0,59)	2,73 (0,55)	3,10 (0,65)	3,01 (0,66)	3,12 (0,66)	2,83 (0,59)	2,90 (0,61)

Таблица 4. Максимальная энтропия, энтропия первого уровня и условные энтропии современного сербского языка Негоша и еще пяти европейских языков

Таким образом, бросается в глаза очень маленькое уменьшение энтропии H_1 по отношению к H_{max} , всего на 0,04 процента индекса (индекс =1), так что узнавание эвидентных разниц в частотности фонем не слишком увеличило упорядочность совокупности. Уже на втором уровне H_2 изменяется в рамках значений энтропии перечисленных языков (с примечанием, что только русский язык имеет совокупность из 32 элементов - знаков). По отношению к современному литературному языку,

уменьшение энтропии H_2 по сравнению с H_1 в языке Негоша вдвое больше (28 : 14 индексных процента) и указывает на свойство фонем Негоша, которые ведут себя более селективно, т.е. создают меньшее число комбинаций диграмм.

На третьем уровне номинальные значения языка Негоша имеют меньшие значения, чем энтропии большинства других сравниваемых языков, тогда как действительная энтропия (из-за большого числа знаков в графической системе) объективно еще меньше. Однако, падение энтропии на третьем уровне меньше (по отношению к энтропии диграммы), чем в исследованиях литературного языка, но и кроме того упорядочность совокупности, т.е. определенность после третьей известной фонемы у Негоша немного больше, что обусловлено статистическими закономерностями языка Негоша (Рис. 3), о чем будет больше говориться в следующих главах.

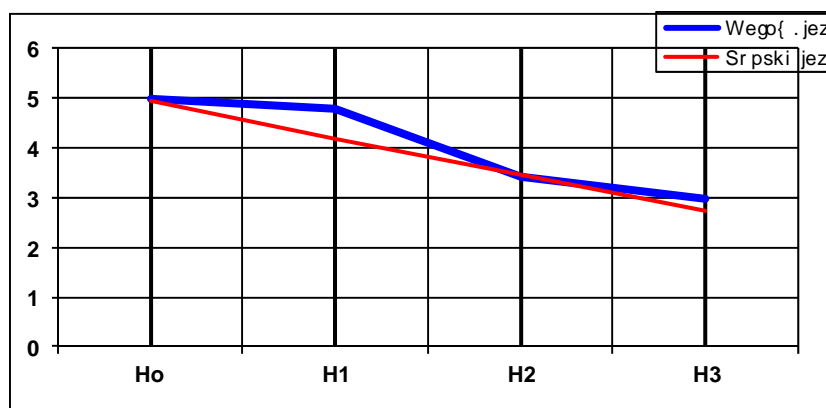


Рис. 3. Энтропия современного сербского языка Негоша

Энтропия первого порядка - H_1 , как и условные энтропии второго и третьего порядков H_2 и H_3 , сделаны на совокупном корпусе, а одновременно и на особых произведениях Негоша (Таблица 5).

Число	Произведение	N	H_0	H_1	H_2	H_3
1	GV	90.980	5	4,8031	3,3706	2,8727
2	LM	61.476	5	4,7684	3,3616	2,7653
3	LC[M	141.532	5	4,7736	3,3599	2,8619
4	Sv.	149.975	5	4,7956	3,3752	2,8000
5	GK	55.482	5	4,7625	3,3188	2,7254
6	OS - 1	51.520	5	4,8026	3,3730	2,7743
7	OS - 2	320.795	5	4,8031	3,3860	2,8571
8	Pj.	239.034	5	4,8012	3,4292	2,9837
9	PP	257.296	5	4,7482	3,3352	2,8382
10	Совок. Опус	1.368.098	5	4,7908	3,4087	2,9745

Таблица 5. Сравнительные данные H_0 , H_1 , H_2 и H_3 для всех частей опуса Негоша

Полученные данные показывают очень близкие значения энтропии для всех уровней аппроксимаций. Небольшие расхождения можно объяснить чувствительностью значения H к величине корпуса и типу исследуемого материала. В этом случае разница так незначительна, что на этом основании нельзя делать выводы о специфике выражения Негоша в особых его произведениях.

2.4. Энтропия для $n > 3$

Дальнейший анализ текста на уровне тетраграммы, пятиграммы ... n – граммы мог бы дать настоящее значение энтропии текста произведений Негоша. Однако, так как статистический анализ проведенный только на уровне тетраграммы, дал 32^4 комбинаций букв, то ясно, что аппроксимация порядка $n > 3$ практически невозможна, так что и дальнейшая энтропия вычисляется на основании экспериментального метода.

Метод основывается на возможности угадывания n – ного знака текста, если известны предшествующие ему $n - 1$ знаков. Экзаменуемый, обязательно знающий язык эксперимента, подряд угадывает знаки, т.е. буквы незнакомого текста. Экспериментатор записывает число попыток угадывания, причем среднее значение числа угадываний из определенного числа попыток представляет собой оценку энтропии для этого уровня (H_N). Угадывание, конечно, не проводится случайно, а лицо, на котором проводится эксперимент, использует свои знания, которые могли бы помочь ему в предикции. В нашем случае, это могло бы быть определение частотности появления фонем в языке Негоша, по эпической структуре стиха (цезура, рифма, звательная форма именительного падежа), определение коэффициента афинитета, т. е. меры взаимовлияния фонем, а также других, психологических фактов. Если известно, что все части слов не несут одинакового количества информации, а начальные фонемы несут наибольшее, и если нам известны именно эти части слов, то нет сомнений, что энтропия языка Негоша будет снижаться очень быстро, т. е. будет продолжаться тренд начальных уровней энтропии.

Что окончательно показывают представленные данные энтропии языка Негоша на первом, втором и третьем уровне упорядоченности высказывания? Во-первых, максимальные энтропии (H_0), т. е. энтропии, где все знаки арбитражны (одинаково вероятны), зависят только от числа элементов дискретного ряда, т. е. числа графических знаков текста. Во-вторых, энтропия первого порядка (H_1), где уже известна вероятность появления каждого из элементов, зависит от частотности появления каждого из них, что непременно снижает первоначальную энтропию.

В-третьих, энтропия второго и третьего порядка обусловлена числом комбинаций диграмм и триграмм, т.е. способностью фонем создавать такие комбинации. Если число комбинаций больше, уменьшение начальной комбинации будет меньше и наоборот.

Принимая во внимание то, что аппроксимации вычисления энтропии языка Негоша проводились на 32 символах, становится логичным то, что максимальная энтропия больше, чем в других исследованиях, на что указывают сравнительные данные энтропии, вычисленной на совокупности 30 символов (без расстояния и дефиса в полусложных словах). И энтропия первого порядка также на уровне принятых данных с

разницей, обусловленной специфическим распределением частотности звуков фонологической системы Негоша.

Энтропия второго и третьего порядков объективно меньше энтропии компаративных исследований, из-за того, что звуки фонологической системы Негоша более селективны в создании диграммных и триграммных секвенций. Кстати, прежние исследования проводились на образцах языка литературных произведений и языка массовой коммуникации, где существуют и слова иностранного происхождения (новые языковые влияния - германизмы, англицизмы и др.), в которых фонемы в создании слова менее селективны. Структура стиха Негоша и потребности версификации также должны были дать специфический признак комбинациям фонем (ассимиляция гласных, йотация зубных Д и Т, элизия гласных и т.п.). Больше внимания этим явлениям будет уделено в следующих работах.

ЛИТЕРАТУРА

- [1] Bertil Malmberg, *Moderna lingvistika*, Beograd, Slovo qubqe, 265 (1979).
- [2] Branko Kne`evi}, *Iformativna struktura Andri}evog kwi`evnog jezika uslovqena statisti~kim zakonomjernostima,-* Me}unarodni sastanak slavista u Vukove dane, Beograd, Novi Sad, Tr{i}, MSC, 417-427(1992).
- [3] Branko, Kne`evi}, *Informatika 1*, Beograd, Trebnik, 171 (1994).
- [4] Branko, Kne`evi}, *Semanti~ka vrednost pojma "neodre}enost" u teoriji informacija* - Beograd, Kwi`evnost i jezik, XXXXIX, 1 65-69 (1991).
- [5] Danilo Vu{ovi}, *Prilozi prou~avawu Wego{eva jezika*, Beograd, Jugoslovenski filolog, Kw. IX, 93-193 (1930).
- [6] Dejvid Kristal, *Kembri~ka enciklopedija jezika*, Beograd, Nolit, 474 (1996).
- [7] Du{an Jovi}, *Jezi~ki sistem i poetska gramatika*, Beograd, Bigz; Pri{tina, Jedinstvo, 62-78 (1985).
- [8] Georgije Lukatela, *Statisti~ka teorija telekomunikacija i teorija informacija I*, Beograd, Gra}evinska kwiga, 221-230 (1991).
- [9] James Bullock O., *Literacy in the Language of Mathematics*, Washington, The American Mathematical Monthly, Vol. 101/8, 735-743 (1994).
- [10] Milo}jko Kokovi}, *Matemati~ke metode i modeli u ra~unarskoj lingvistici*, iz : *Jezik uxbenika*, Beograd, Zavod za uxbenike i nastavna sredstva, 13-18 (1977).
- [11] Rodoqub Mi{i}, *Ocjene entropije jezika metodom [enona-Matemati~ka i ra~unarska lingvistika*, Beograd, Zbornik radova Dru}tva za primewenu lingvistiku Srbije (1987).
- [12] Slobodan Jovi~i} T., *Govorna komunikacija : filologija, psihologija, percepcija* - Beograd, Nauka, 558 (1999).
- [13] Soloman Markus, *Matemati~ka poetika*, Nolit, Beograd, 373 (1974).
- [14] Stani}a Milo{evi}, *Entropija {tampanog teksta srpskog jezika*, Beograd, Tehnika, 1378-1383, (1988/9), *Elektrotehnika*, 186-191 (1988/8).
- [15] Velimir Jovanovi}, Milan Š[ipka, *Optimalni kod srpskog štampnog teksta* - Beograd, Telekomunikacije, 12 , 7-12 (1964).

Приложение 1. Отдельные знаки корпуса – монограммы

Знак	Ранг		Частотность в абсолютной сумме	Относительная частотность $p(i)$		Парциальная энтропия $-p(i) \log_2 p(i)$		Количество информ. по феонеми
	n=30	n=32		n=30	n=32	n=30	n=32	
		1	242.645		0,1774		0,4426	2,2127
A	1	2	132.427	0,1177	0,0968	0,3634	0,3261	3,0864
B	18	19	20.219	0,0180	0,0148	0,1042	0,0899	5,7978
V	10	11	43.986	0,0391	0,0322	0,1829	0,1594	4,6765
G	17	18	22.512	0,0200	0,0165	0,1129	0,0975	5,6428
D	11	12	42.707	0,0380	0,0312	0,1792	0,1561	4,7190
\	28	29	3.931	0,0035	0,0029	0,0285	0,0243	8,1605
E	2	3	106.457	0,0946	0,0778	0,3219	0,2867	3,4013
@	27	28	5.859	0,0052	0,0043	0,0395	0,0337	7,5848
Z	20	21	15.291	0,0136	0,0112	0,0843	0,0725	6,2008
I	4	5	100.828	0,0896	0,0737	0,3119	0,2773	3,4797
J	14	15	39.523	0,0351	0,0289	0,1697	0,1477	4,8308
K	13	14	40.974	0,0364	0,0299	0,1741	0,1516	4,7788
L	16	17	29.999	0,0267	0,0219	0,1395	0,1208	5,2286
Q	26	27	6.137	0,0055	0,0045	0,0410	0,0350	7,5179
M	12	13	41.230	0,0367	0,0301	0,1748	0,1523	4,7698
N	6	7	56.247	0,0500	0,0411	0,2161	0,1893	4,3217
W	25	26	6.479	0,0058	0,0047	0,0429	0,0366	7,4397
O	3	4	103.413	0,0919	0,0756	0,3166	0,2816	3,4432
P	15	16	30.279	0,0269	0,0221	0,1404	0,1217	5,2152
R	5	6	56.891	0,0506	0,0416	0,2178	0,1908	4,3053
S	7	8	54.063	0,0481	0,0395	0,2105	0,1842	4,3789
T	9	10	50.125	0,0446	0,0366	0,2000	0,1748	4,4880
	22	23	10.372	0,0092	0,0076	0,0623	0,0534	6,7608
U	8	9	53.618	0,0477	0,0392	0,2093	0,1832	4,3908
F	29	30	1.057	0,0009	0,0008	0,0094	0,0080	10,0555
H	24	25	10.073	0,0090	0,0074	0,0609	0,0522	6,8030
C	23	24	10.258	0,0091	0,0075	0,0618	0,0529	6,7768
^	21	22	10.869	0,0097	0,0079	0,0647	0,0554	6,6933
X	30	32	517	0,0005	0,0004	0,0051	0,0043	11,0872
	19	20	18.453	0,0164	0,0135	0,0973	0,0838	5,9297
-		31	659		0,0005		0,0053	10,7371
		□	1.368.098	1,0000	1,0000	4,3429	4,7908	184,9145