Statistical Analysis of Texts of the Balkans Electronic Media Columnists

Nedim Smailović

Pan-European University "APEIRON", Banja Luka, Republika Srpska, Bosnia and Herzegovina

Contribution to the state of the art

DOI: **10.7251/JIT1901005S** UDC: **659.3/.4:316.776]:004.738.5**

Abstract: This paper presents results of statistical analysis of some segments in texts of the four columnists in the Balkans electronic media: Bosnia and Herzegovina – *Dnevni avaz* (Muhamed Filipović), Serbia – *Politika* (Aleksandar Apostolovski), Croatia – *Jutarnji list* (Miljenko Jergović) and Montenegro – *Vijesti* (Miodrag Lekić). They write about different themes, in different language styles, but statistical analysis clearly points to large similarities in certain segments, such as number of particular alphabet letters, most common combinations of two or three words, etc. These results leave space to conclude that it is one polycentric language, which is not a rare phenomenon in the modern world. Naturally, the final judgement about this should be given by the linguists.

Keywords: linguistics, language, electronic media, text analysis, visualization of data.

Introduction

Researching the speech and writing systems is a field that receives large attention nowadays. It is understandable, since quality transmission of the verbal signal is impossible without good knowledge of properties of speech. The significance of linguistic works is best confirmed by the fact that highly expensive projects of researching the speaking is financed by the military. Such activity received a large impulse during the WW2. A lot is being done on studying the psychology of listening, understanding the speech and its resilience to interference, synthesis and analysis of speech, translation of encoded messages and foreign language.

What are the touching points, and where do the linguists and the technicians get apart? The linguists are interested in the correctness of content, while the technicians take care of the optimal signal transfer. A linguist wants highest possible intelligibility, while a technician will settle to a minimum intelligibility to satisfy the economical principle. Language terms, such as: rhythm of words, vowels, consonants, stress, length, gender, number and case of each noun and adverb, agreement of sounds in words, words in a sentence, etc. – all these serve the language system to increase intelligibility, secure undisturbed transfer of information and enhance expressivity. Means that improve resilience to interference in language are called redundancy. It is believed that languages contain up to 50% of redundancy, i.e. one half of language means contains the necessary information, while the role of the other half is to enable more secure transfer and receipt of information.

How much redundancy is there in a language is shown in the following example, where, with a little effort, it is possible to understand the meaning of the following text, in which most of the vowels are left out:

I WLD NT B MPSSBL T NDRSTND VN A TXT LK THS, WHCH S WTHT VWELS AND WTHT SPC N TH PLC F TH VWELS.

IT WLD B TH SMPLST ND MST PRCTICL IF W WLD LVE VWELS NLY T TH STRT F WRDS, N IMPRTNT PRFXS ND SFXS. W CLD SVE A LT F TM ND SPC. THAT IS HW IT S MAINLY DNE IN STNOGRAPHY, NT TO MNTION ARAB OR HEBREW WRITING.

Though at first glance it looks like some kind of game, solving such linguistic problems receives a great atten-

tion in the world today, since the roots of contemporary machine and automated translations, and even certain segments of artificial intelligence can be found there.

FREQUENCY OF LETTERS IN EUROPEAN LANGUAGES

Language can be described as a system of symbols that serves for communication among people. [9]. It is a form that mankind expresses itself as a thinking being, uncovering its essence and its difference compared to other living beings. [9].

Linguistics deals with internal order among units of language. Different languages usually have different sets of voices and different sets of letters to write them down. Frequency of use of particular letters is different in every language, as well as in comparison between languages.

In the given languages, there are 84 different letters used, and those are: a, á, à, â, ä, ã, å, ą, æ, b, c, ĉ, ç, č, ć, d, ď, ð, e, ě, é, è, ê, ë, ę, f, g, ğ, ĝ, h, ĥ, i, ı, í, ì, î, ï, j, ĵ, k, l, ł, m, n, ň, ń, ñ, o, ó, ò, ô, ö, õ, ø, œ, p, q, r, ř, s, ś, ŝ, ş, ß, š, t, ť, þ, u, ŭ, ú, ù, û, ü, ů, v, w, x, y, ý, z, ź, ż, ž.

Table 1. *Relative frequences of use of letters in some European languages (descending towards English)*

Letter	English	French	German	Spanish	Esperanto	Italian	Turkish	Swedish	Polish	Danish	Czech
e	12.702%	14.715%	16.396%	12.181%	8.995%	11.792%	9.912%	10.149%	7.352%	15.453%	7.562%
t	9.056%	7.244%	6.154%	4.632%	5.276%	5.623%	3.314%	7.691%	2.475%	6.862%	5.727%
a	8.167%	7.636%	6.516%	11.525%	12.117%	11.745%	12.920%	9.383%	10.503%	6.025%	8.421%
o	7.507%	5.796%	2.594%	8.683%	8.779%	9.832%	2.976%	4.482%	6.667%	4.636%	6.695%
i	6.966%	7.529%	6.550%	6.247%	10.012%	10.143%	9.600%*	5.817%	8.328%	6.000%	6.073%
n	6.749%	7.095%	9.776%	6.712%	7.955%	6.883%	7.987%	8.542%	6.237%	7.240%	6.468%
s	6.327%	7.948%	7.270%	7.977%	6.092%	4.981%	3.014%	6.590%	5.224%	5.805%	5.212%
h	6.094%	0.737%	4.577%	0.703%	0.384%	0.636%	1.212%	2.090%	1.015%	1.621%	1.356%
r	5.987%	6.693%	7.003%	6.871%	5.914%	6.367%	7.722%	8.431%	5.243%	8.956%	4.799%
d	4.253%	3.669%	5.076%	5.010%	3.044%	3.736%	5.206%	4.702%	3.725%	5.858%	3.475%
l	4.025%	5.456%	3.437%	4.967%	6.104%	6.510%	5.922%	5.275%	2.564%	5.229%	3.802%
c	2.782%	3.260%	2.732%	4.019%	0.776%	4.501%	1.463%	1.486%	3.895%	0.565%	0.740%
u	2.758%	6.311%	4.166%	2.927%	3.183%	3.011%	3.235%	1.919%	2.062%	1.979%	2.160%
m	2.406%	2.968%	2.534%	3.157%	2.994%	2.512%	3.752%	3.471%	2.515%	3.237%	2.446%
w	2.360%	0.049%	1.921%	0.017%	0	0.033%	0	0.142%	5.813%	0.069%	0.016%
f	2.228%	1.066%	1.656%	0.692%	1.037%	1.153%	0.461%	2.027%	0.143%	2.406%	0.084%
g	2.015%	0.866%	3.009%	1.768%	1.171%	1.644%	1.253%	2.862%	1.731%	4.077%	0.092%
y	1.974%	0.128%	0.039%	1.008%	0	0.020%	3.336%	0.708%	3.206%	0.698%	1.043%
p	1.929%	2.521%	0.670%	2.510%	2.755%	3.056%	0.886%	1.839%	2.445%	1.756%	1.906%
b	1.492%	0.901%	1.886%	2.215%	0.980%	0.927%	2.844%	1.535%	1.740%	2.000%	0.822%
v	0.978%	1.838%	0.846%	1.138%	1.904%	2.097%	0.959%	2.415%	0.012%	2.332%	5.344%
k	0.772%	0.074%	1.417%	0.011%	4.163%	0.009%	5.683%	3.140%	2.753%	3.395%	2.894%
j	0.153%	0.613%	0.268%	0.493%	3.501%	0.011%	0.034%	0.614%	1.836%	0.730%	1.433%
x	0.150%	0.427%	0.034%	0.215%	0	0.003%	0	0.159%	0.004%	0.028%	0.027%
q	0.095%	1.362%	0.018%	0.877%	0	0.505%	0	0.020%	0	0.007%	0.001%
z	0.074%	0.326%	1.134%	0.467%	0.494%	1.181%	1.500%	0.070%	4.852%	0.034%	1.503%

Source table at the address: *https://en.wikipediam.org/wiki/Letter_frequency [4] also contains data on frequency of use of other letters being used in each of the given 11 languages*

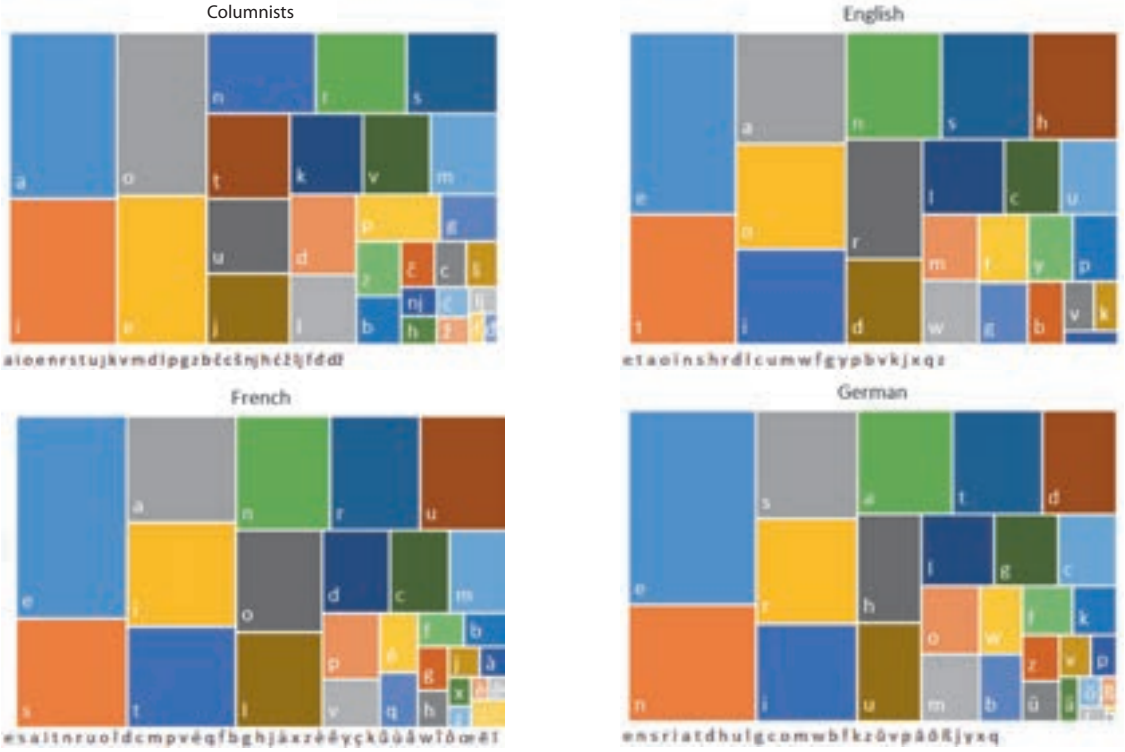
Table 2. *Eleven languages and sorted set of letters used in them*

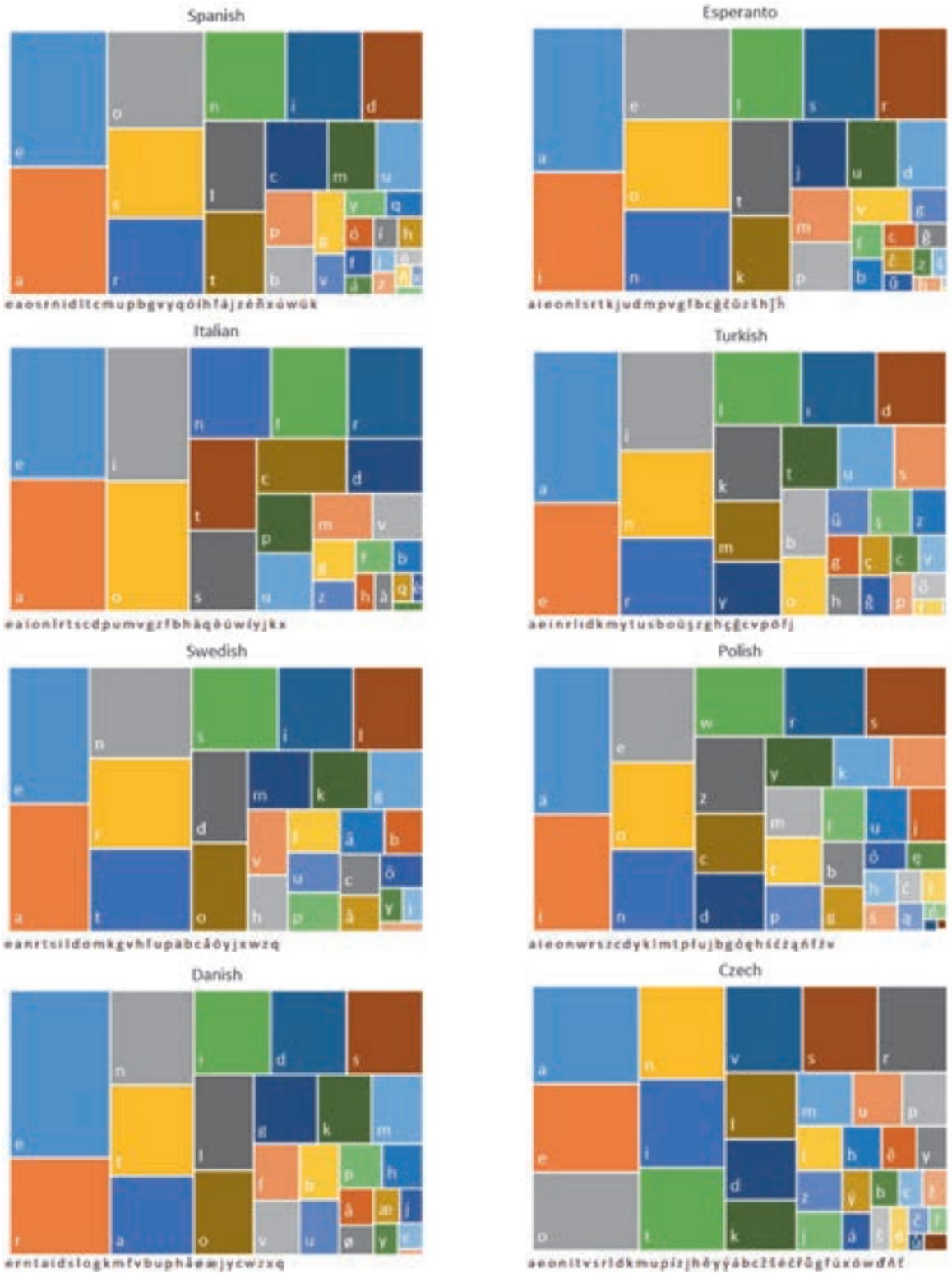
Letter	English	Letter	French	Letter	German	Letter	Spanish	Letter	Esperanto	Letter	Italian	Letter	Turkish	Letter	Swedish	Letter	Polish	Letter	Danish	Letter	Czech
e	12,70%	e	14,72%	e	16,40%	e	12,18%	a	12,12%	e	11,79%	a	11,92%	e	10,15%	a	10,50%	e	15,45%	a	8,42%
t	9,06%	s	7,95%	n	9,78%	a	11,53%	i	10,01%	a	11,75%	e	8,91%	a	9,38%	i	8,33%	r	8,96%	e	7,56%
a	8,17%	a	7,64%	s	7,27%	o	8,68%	e	9,00%	i	10,14%	i	8,60%	n	8,54%	e	7,35%	n	7,24%	o	6,70%
o	7,51%	i	7,53%	r	7,00%	s	7,98%	o	8,78%	o	9,83%	n	7,49%	r	8,43%	o	6,67%	t	6,86%	n	6,47%
i	6,97%	t	7,24%	i	6,55%	r	6,87%	n	7,96%	n	6,88%	r	6,72%	t	7,69%	n	6,24%	a	6,03%	i	6,07%
n	6,75%	n	7,10%	a	6,52%	n	6,71%	l	6,10%	l	6,51%	l	5,92%	s	6,59%	w	5,81%	i	6,00%	t	5,73%
s	6,33%	r	6,69%	t	6,15%	i	6,25%	s	6,09%	r	6,37%	ı	5,11%	i	5,82%	r	5,24%	d	5,86%	v	5,34%
h	6,09%	u	6,31%	d	5,08%	d	5,01%	r	5,91%	t	5,62%	d	4,71%	l	5,28%	s	5,22%	s	5,81%	s	5,21%
r	5,99%	o	5,80%	h	4,58%	l	4,97%	t	5,28%	s	4,98%	k	4,68%	d	4,70%	z	4,85%	l	5,23%	r	4,80%
d	4,25%	l	5,46%	u	4,17%	t	4,63%	k	4,16%	c	4,50%	m	3,75%	o	4,48%	c	3,90%	o	4,64%	l	3,80%
l	4,03%	d	3,67%	l	3,44%	c	4,02%	j	3,50%	d	3,74%	y	3,34%	m	3,47%	d	3,73%	g	4,08%	d	3,48%
c	2,78%	c	3,26%	g	3,01%	m	3,16%	u	3,18%	p	3,06%	t	3,31%	k	3,14%	y	3,21%	k	3,40%	k	2,89%
u	2,76%	m	2,97%	c	2,73%	u	2,93%	d	3,04%	u	3,01%	u	3,24%	g	2,86%	k	2,75%	m	3,24%	m	2,45%
m	2,41%	p	2,52%	o	2,59%	p	2,51%	m	2,99%	m	2,51%	s	3,01%	v	2,42%	l	2,56%	f	2,41%	u	2,16%
w	2,36%	v	1,84%	m	2,53%	b	2,22%	p	2,76%	v	2,10%	b	2,84%	h	2,09%	m	2,52%	v	2,33%	p	1,91%
f	2,23%	é	1,50%	w	1,92%	g	1,77%	v	1,90%	g	1,64%	o	2,48%	f	2,03%	t	2,48%	b	2,00%	í	1,64%
g	2,02%	q	1,36%	b	1,89%	v	1,14%	g	1,17%	z	1,18%	ü	1,85%	u	1,92%	p	2,45%	u	1,98%	z	1,50%
y	1,97%	f	1,07%	f	1,66%	y	1,01%	f	1,04%	f	1,15%	ş	1,78%	p	1,84%	ł	2,11%	p	1,76%	j	1,43%
p	1,93%	b	0,90%	k	1,42%	q	0,88%	b	0,98%	b	0,93%	z	1,50%	ä	1,80%	u	2,06%	h	1,62%	h	1,36%
b	1,49%	g	0,87%	z	1,13%	ó	0,83%	c	0,78%	h	0,64%	g	1,25%	b	1,54%	j	1,84%	å	1,19%	ě	1,22%
v	0,98%	h	0,74%	ü	1,00%	í	0,73%	ĝ	0,69%	à	0,64%	h	1,21%	c	1,49%	b	1,74%	ø	0,94%	y	1,04%
k	0,77%	j	0,61%	v	0,85%	h	0,70%	ĉ	0,66%	q	0,51%	ç	1,16%	å	1,34%	g	1,73%	æ	0,87%	ý	1,00%
j	0,15%	à	0,49%	p	0,67%	f	0,69%	ŭ	0,52%	è	0,26%	ğ	1,13%	ö	1,31%	ó	1,14%	j	0,73%	á	0,87%
x	0,15%	x	0,43%	ä	0,58%	á	0,50%	z	0,49%	ú	0,17%	c	0,96%	y	0,71%	ę	1,04%	y	0,70%	b	0,82%
q	0,10%	z	0,33%	ö	0,44%	j	0,49%	š	0,39%	w	0,03%	v	0,96%	j	0,61%	h	1,02%	c	0,57%	c	0,74%
z	0,07%	è	0,27%	ß	0,31%	z	0,47%	h	0,38%	í	0,03%	p	0,89%	x	0,16%	ś	0,81%	w	0,07%	ž	0,72%
		ë	0,22%	j	0,27%	é	0,43%	ĵ	0,06%	y	0,02%	ö	0,78%	w	0,14%	ć	0,74%	z	0,03%	š	0,69%
		y	0,13%	y	0,04%	ñ	0,31%	ĥ	0,02%	j	0,01%	f	0,46%	z	0,07%	ż	0,71%	x	0,03%	é	0,63%
		ç	0,09%	x	0,03%	x	0,22%			k	0,01%	j	0,03%	q	0,02%	ą	0,70%	q	0,01%	č	0,46%
		k	0,07%	q	0,02%	ú	0,17%			x	0,00%					ń	0,36%			ř	0,38%
		û	0,06%			w	0,02%									f	0,14%			ů	0,20%
		ù	0,06%			ü	0,01%									ź	0,08%			g	0,09%
		â	0,05%			k	0,01%									v	0,01%			f	0,08%
		w	0,05%																	ú	0,05%
		î	0,05%																	x	0,03%
		ô	0,02%																	ó	0,02%
		œ	0,02%																	w	0,02%
		ë	0,01%																	ď	0,02%
		ï	0,01%																	ň	0,01%
																				ť	0,01%

If we present the usage frequency of those letters for each language in the descending order, we get the following table:

The following figures present the data from the table above, visualized by Treemap type of charts. Letter usage frequency is proportional to the corresponding surface on the chart. This kind of presentation clearly shows dominance (higher frequency) of particular letters compared to other letters. Below each chart there is a descending order of letters for that language, left to right.

Table 3. *Graphical presentation of letter usage frequency in particular languages*





For easier remembering of the order of letter usage frequency, the bibliography offers the first twelve most frequently used letters given in two (non-existing) six-letter words. For our eleven languages, those words are:

English	etaoin shrdlc
French	esaitn ruoldc
German	ensria tdhulg
Spanish	eaosrn idltcm
Esperanto	aieonl srtkju
Italian	eaionl rtscdp
Turkish	aeinrl ıdkmyt

Swedish	eanrts ildomk
Polish	aieonw rszcdy
Danish	erntai dslogk
Czech	aeonit vsrldk

World statistics confirms that there are over 7100 spoken languages today. It is not easy to establish the exact number of languages as there are no clear boundaries between certain languages and dialects. Besides, some languages are disappearing. In the course of the 20th century, 110 languages were proclaimed extinct, while 12 languages have disappeared in this century.

Published scientific linguistic works, printed and electronic, analyze language and its characteristics in many places. Naturally, such analyses usually refer to dominant world languages: Chinese, Spanish, English, Hindi, Arab, Bengali, Portuguese, Russian, Japanese… [1]

This list of the largest (by number of people speaking) world languages is different, if sorted according to their impact in trade and industry, social, political and economy circles. English language is on top of that list, followed by French, which is an official language in 25 countries. There are 2303 languages spoken in Asia, while in Europe there are 285.

Language is a living organism. It is constantly changing. Some words emerge, some get into a language from other languages, some words change their meaning, and some become archaic over the time and disappear. Therefore, new analyses should be done in continuity, while that job is significantly alleviated with new IT tools.

Statistical Analysis of Text in Selected Columns

A smaller or larger sample is necessary for any analysis. The samples for this analysis were taken from texts by four columnists in the Balkans electronic media: Bosnia and Herzegovina – *Dnevni avaz* (Muhamed Filipović), Serbia – *Politika* (Aleksandar Apostolovski), Croatia – *Jutarnji list* (Miljenko Jergović) and Montenegro – *Vijesti* (Miodrag Lekić). [2] [3] [7] [8] All of them still write for the mentioned media, so we can say that this analysis deals with modern language.

In order to be able to make a comparative analysis of language of aforementioned columnists, it is necessary to compare samples of equal length, so the texts were selected and merged to be equal in number of characters (without blanks). In all examples, the merged set of columns for each columnist had exactly 148.232 characters (without blanks). Among these characters there are also letters not used in the columnists' alphabets (such as w, q and similar). For the said number of characters, it took: 23 columns by Miljenko Jergović, 32 columns by Filipović, 34 columns by Apostolovski and 25 columns by Lekić. In some tables, we analyzed a summary-merged text of all columnists, which gives a sample of over half a million characters.

Table 4. *Framework statistics of text samples*

	ALL	JERGOVIĆ	FILIPOVIĆ	APOSTOLOVSKI	LEKIĆ
No. of pages	163	40	46	41	36
No. of words	109.707	27.976	28.857	27.509	25.365
No. of characters without blanks	592.928	148.232	148.232	148.232	148.232
No. of characters with blanks	701.032	175.893	176.607	175.417	173.115
No. of paragraphs	1.936	323	565	425	623
No. of lines	7.997	1.953	1.946	2.161	1.937
No. of sentences	5.519	1.301	1.140	1.503	1.575

It is important to say that in this paper we did not analyze standpoints or opinions of columnists or their editors, but exclusively statistical analysis of the text produced.

The authors have their style of writing, and some combinations occur more often than others. In the following table, we gave three-words combinations that occur with each author over five times.

In a joined text of all columnists, "ono što je" (which is) was the most often repeated combination of three words.

Individually, it was as follows:

Jergović: "ono što je" (13 times), Filipović: "Bosne i Hercegovine" (24 times), Apostolovski: "da li je" (12 times) and Lekić: "u Crnoj Gori" (25 times).

Table 5. *The most often repeated combination of three words*

3 words together	ALL	3 words together	JERGOVIĆ	3 words together	FILIPOVIĆ	3 words together	APOSTOLOVSKI	3 words together	LEKIĆ
ono što je	33	ono što je	13	Bosne i Hercegovine	24	da li je	12	u Crnoj Gori	25
i da se	30	koji su se	12	Bosni i Hercegovini	23	se da je	10	s druge strane	10
s druge strane	29	i to je	11	a to je	19	kako bi se	10	o Crnoj Gori	9
u Crnoj Gori	27	a onda i	11	u Bosni i	18	da li će	9	u isti mah	9
Bosne i Hercegovine	26	ne samo da	10	ono što je	18	kao da je	8	Crnoj Gori i	8
ono što se	25	u vrijeme kada	8	s druge strane	18	ne može da	7	i da se	6
Bosni i Hercegovini	25	je riječ o	8	i da se	15	kao što je	6	SAD i Rusije	6
kao što je	23	prije nego što	8	na taj način	13	i da se	6	radi se o	6
tako da je	23	ono što se	7	kao što je	12	je da je	6	na međunarodnom planu	6
a to je	22	tako da se	7	da je to	12	je reč o	6	u svakom slučaju	6
koji su se	22	tako da je	7	tako da je	12	saveza za s Srbiju	6	i da je	5
kao što su	21	kao što su	6	Bosna i Hercegovina	12	u kojoj je	6	u vezi sa	5
da je to	20	a zatim i	6	da se ne	11	pravoslavne nove godine	6	u kojem je	5

Table 6. *The most commonly repeated combination of two words*

2 words together	ALL	2 words together	JERGOVIĆ	2 words together	FILIPOVIĆ	2 words together	APOSTOLOVSKI	2 words together	LEKIĆ
da je	418	da je	86	da se	171	da je	127	da je	63
da se	395	što je	78	da je	140	da se	100	da se	59
što je	210	da se	63	i da	90	da su	48	pa i	44
koji je	174	je u	60	što je	86	koji je	43	crnoj gori	39
je u	167	i u	47	koji je	73	su se	36	je u	36
i da	157	su se	39	koji su	55	je u	36	koji su	28
koji su	134	kao i	37	to je	54	da će	36	i u	27
i u	127	što se	37	ono što	45	koji se	36	koji je	27
su se	127	koji su	32	koja je	43	da li	32	u crnoj	26
da su	126	ono što	32	je bio	42	je to	30	je bio	26
to je	113	koji je	31	tako da	41	je da	30	i to	23
je to	107	bi se	30	zbog toga	41	se u	30	je i	21
što se	102	je to	29	da će	40	se da	29	su se	20

In a joined text of all columnists, "da je" (that is) was the most often repeated combination of two words – 418 times. With authors individually, two most commonly words used together were: Jergović: "da je" (86 times), Filipović: "da se " (171 times), Apostolovski: "da je" (127 times) and Lekić: "da je" (63 times).

Looking at individual words, it is noticeable that among the most frequently used ones in a joined text there were no nouns, verbs, adjectives, numbers… The dominating were connections, auxiliary verbs, prepositions (non-lexical, function words).

It is interesting that: **i**, **je**, **u**, **da**, **se**, **na** with all authors were among the first six most frequent words.

In that sense, we could perform a separate analysis of lexical density in the columns, measuring how informative the text is.

Remark:

Lexical density is defined as a number of lexical words (or content words) divided by a total number of words. Lexical words give meaning to a text. Those are nouns, adjectives, verbs and adverbs. Other types of words (functional words), such as auxiliary verbs, prepositions or conjunctions, are more of grammatical nature and they give little or no information about the subject matter.

With individual authors, the most frequent nouns were: Jergović: "vrijeme" (time) and "godina" (year) (47 times), Filipović: "države" (state) (67 times) and "ljudi" (people) (65 times), Apostolovski: "godine" (year) (66 times) and Lekić: "istorije" (history) (62 times) and "rata" (war) (52 times).

Table 7. *The most frequent words*

the most frequent word	ALL	the most frequent word	JERGOVIĆ	the most frequent word	FILIPOVIĆ	the most frequent word	APOSTOLOVSKI	the most frequent word	LEKIĆ
i	4603	i	1389	i	1153	je	1033	i	1137
je	3889	je	1058	je	1098	i	924	u	873
u	3390	u	827	da	1088	da	847	je	700
da	2796	se	520	u	859	u	831	da	407
se	1986	da	454	se	580	se	539	se	347
na	1551	na	416	na	389	na	412	na	334
su	1223	su	323	sam	319	su	365	su	243
za	768	što	297	su	292	za	235	sa	184
što	692	ne	224	to	262	kao	200	za	174
to	684	a	196	a	241	od	176	o	144
ne	678	kao	196	koji	232	koji	165	od	132
a	673	s	195	što	230	a	160	iz	116
koji	663	od	190	ne	182	ne	156	ne	116
kao	652	nije	188	za	176	ali	147	koji	115
od	622	za	183	s	165	će	145	to	111
o	562	to	181	kao	161	s	138	kao	95
s	528	bi	171	o	158	bi	137	nije	81
nije	515	o	164	mi	141	to	130	a	76
bi	452	koji	151	bio	130	nije	128	treba	72
sam	440	ali	136	od	124	iz	114	dakle	69
iz	439	ili	134	koja	119	što	111	do	67
će	421	će	117	nije	118	o	96	pa	65
ali	401	iz	115	on	108	kako	91	bio	63
ili	353	tako	96	tako	108	ili	79	istorije	62
sa	337	samo	95	odnosno	102	kada	76	već	61
bio	330	ni	94	će	101	–	74	koje	60
koja	277	bilo	94	bi	96	bio	72	će	58
kako	273	nego	88	koje	95	već	70	bez	57

koje	272	po	83	iz	94	sa	66	sve	57
tako	268	sve	82	bilo	92	godine	66	ili	56
bilo	257	te	79	smo	92	ga	65	što	54
samo	257	kada	73	ja	88	po	65	rata	52
kada	245	kako	68	bih	86	li	62	ali	51
sve	242	koja	68	jer	86	mu	60	uz	49
do	236	biti	66	ili	84	jer	60	koja	49
po	228	koje	65	kako	79	ako	58	zemlje	49
pa	223	bio	65	nego	78	samo	58	bi	48
ni	218	pa	63	do	75	sve	56	još	48
mi	214	ga	62	toga	75	posle	55	po	45
jer	212	do	57	tome	71	koje	52	kada	45
on	201	jer	54	kad	71	ni	50	poslije	41
nego	193	onda	51	rekao	70	dok	50	sam	41
već	190	bila	51	ali	67	pa	49	dvije	40
ga	189	ono	49	države	67	još	49	između	40
te	180	on	48	samo	65	jedan	47	crnoj	40
biti	167	vrijeme	47	ljudi	65	tako	44	gori	39
smo	167	godina	47	šta	62	više	43	samo	39
ako	165	može	46	vrlo	62	koja	41	godine	39
li	158	ona	46	način	61	on	40	kako	35
bila	158	ništa	45	prema	59	sam	39	države	35

Digrams, or digraphs, from Greek δίς dís, "twice, two times" and γράφω gráphō, "to write", are combinations of at least two written units, letters (graphemes) to mark a single phoneme in a language. A digraph is not the same as two characters pronounced consecutively. Digraphs are often present in foreign languages, for example: qu, ch, ph, ee, cs, dz, dzs, gy, ly, ny, sz, ty, zs, dh, gj, ll, nj, rr, sh, th, xh, zh. When a digraph is capitalized, both letters are in capital character.

In Bosnian, Serbian, Croatian and Montenegrin language, which were treated as single Serbo-Croat language prior to the fall of SFR Yugoslavia, the digraphs were dž, lj and nj.

Each author has his own style of writing, characterized by different properties. In the analyzed

Table 8. *Frequency of words according to number of characters*

svi		Jergović		Filipović		Apostolovski		Lekić	
Broj karaktera u riječima	Frekvencija (%)	Broj karaktera u riječima	Frekvencija (%)	Broj karaktera u riječima	Frekvencija (%)	Broj karaktera u riječima	Frekvencija (%)	Broj karaktera u riječima	Frekvencija (%)
1	8,9	1	9,9	1	8,9	1	7,8	1	9,0
2	17,1	2	16,8	2	19,5	2	18,0	2	13,7
3	6,6	3	6,8	3	7,6	3	6,7	3	5,0
4	10,9	4	11,8	4	11,5	4	10,1	4	10,1
5	10,4	5	10,5	5	11,0	5	10,0	5	10,0
6	10,1	6	10,1	6	8,8	6	11,1	6	10,8
7	9,3	7	8,6	7	9,5	7	9,5	7	9,7
8	7,8	8	7,5	8	7,3	8	8,0	8	8,6
9	6,2	9	6,1	9	5,5	9	5,7	9	7,5
10	4,8	10	4,5	10	4,0	10	4,8	10	6,0
11	3,5	11	3,2	11	3,1	11	3,4	11	4,2
12	1,8	12	1,7	12	1,6	12	1,8	12	2,3
13	1,1	13	1,1	13	0,8	13	1,0	13	1,4
14	0,7	14	0,7	14	0,4	14	0,8	14	0,9

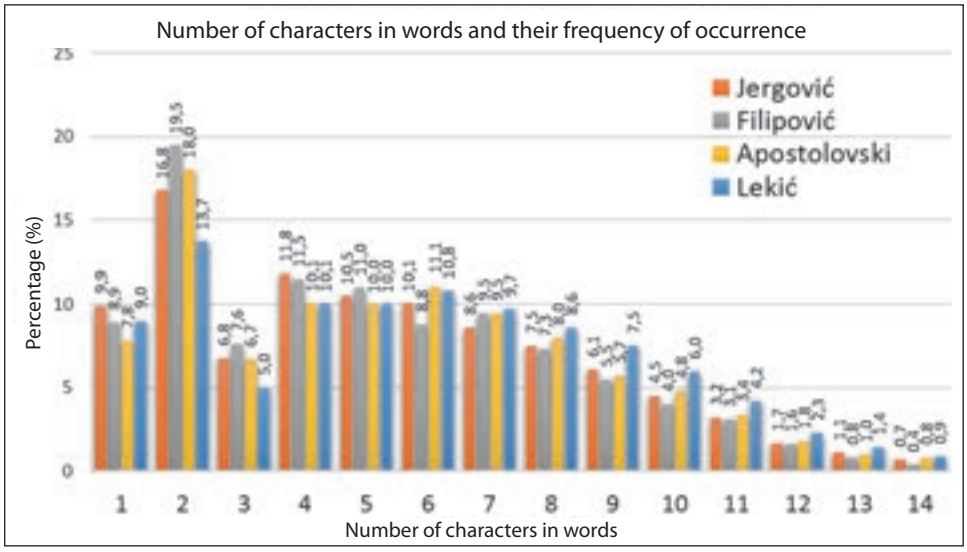


Figure 1. *Frequency of words according to number of characters*

Table 9. *The most frequent combinations of two consecutive letters in analyzed columns*

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
je	na	ra	ko	ni	st	ij	no	an	li	ti	da	po	ta	ka	re	ov	ne	ja	to	pr	ma	en	va	ri	im	oj	al	nj	vi

texts of all authors, words of up to 14 characters make over 99% of text, and their frequency per number of characters is given in the following table and chart.

There were more and less frequent combinations of two consecutive letters in the texts. Some of them were even words, such as: iz, na, li, on, to, mi and similar. In the sample texts there were even combinations of the same letters one next to another, but mostly because the authors were citing original transcription of names, or were using foreign words.

In Bosnian, Serbian, Croatian and Montenegrin language, there are rare examples of two consecutive letters which are the same. They appear in compounds, such as: najjači, najjasniji, narodnooslobodilački, prekookeanski, kooperativan etc.

Table 10. *An overview of frequency of combinations of two consecutive letters*

	a	b	c	č	ć	d	dž	đ	e	f	g	h	i	j	k	l	lj	m	n	nj	o	p	r	s	š	t	u	v	z	ž		rank
a	5	1114	755	743	591	5644	33	314	168	225	1626	261	72	4495	4816	3603	613	4430	8525	882	74	1576	7689	2787	562	5079	178	4378	3019	746	a	65.003	1
b	290	0	5	0	0	164	10	0	610	2	2	2	110	55	0	36	1	154	10	0	1533	0	314	10	0	55	378	0	402	75	b	4.218	22
c	1248	0	6	2	0	1	0	0	287	0	0	0	1467	29	168	32	2	27	460	3	277	37	172	128	0	10	167	81	0	0	c	4.604	19
č	1083	0	0	0	0	0	0	0	375	0	0	0	1707	16	13	2	0	6	45	0	571	8	34	1	29	2	632	11	0	0	č	4.535	20
ć	196	0	0	0	0	0	0	0	963	0	0	0	947	0	8	0	0	8	2	0	203	57	13	0	122	0	598	1	0	0	ć	3.118	23
d	2560	15	1	0	0	1	0	0	3034	0	144	5	607	58	0	48	19	40	456	0	3682	1	210	15	0	21	827	79	173	103	d	12.099	15
dž	30	0	0	0	0	0	0	0	6	0	0	0	16	17	0	0	0	0	2	0	5	0	6	0	0	0	17	0	0	0	dž	99	30
đ	378	0	0	0	0	0	0	0	542	0	0	0	25	0	0	0	0	0	0	0	216	0	75	0	0	0	82	0	0	2	đ	1.320	28
e	50	860	790	938	1143	1728	39	450	7	231	389	201	51	14848	1517	1983	868	2621	4725	1862	37	639	4759	2975	771	2689	71	3062	577	932	e	51.813	2
f	156	0	0	0	0	0	0	0	153	3	0	0	100	4	1	49	0	13	109	0	119	0	16	39	0	6	16	0	0	0	f	784	29
g	605	0	9	0	0	174	0	0	1460	3	2	0	494	10	2	23	0	0	161	0	3592	0	219	1	0	0	795	13	202	0	g	7.765	16
h	229	3	59	0	0	0	0	0	84	0	1	0	2054	5	1	1	0	0	5	0	95	19	95	87	0	0	137	0	0	0	h	2.875	25
i	166	2831	2652	1368	829	2570	40	91	50	480	530	230	2	2755	1750	5927	331	1743	6820	700	103	934	4253	1141	512	5702	27	3845	1127	846	i	50.355	3
j	2522	211	60	30	0	433	0	0	205	0	0	0	6186	4	2	2677	0	353	3873	0	4077	194	88	268	1	92	516	616	49	21	j	22.478	7
k	3007	2	36	1213	23	0	0	0	1868	1	0	5	2309	126	0	62	21	8	343	1	1212	13	272	3040	357	327	557	50	0	0	k	14.853	13
l	4054	401	10	84	0	212	0	0	2184	32	684	25	2797	20	591	24	0	328	4	0	2092	388	237	1751	251	13	688	752	331	12	l	17.965	10
lj	302	59	1	0	0	21	0	0	375	0	2	0	227	2	108	0	0	230	2	0	273	76	25	132	80	36	20	389	1	12	lj	2.373	26
m	2625	1	0	2	0	101	0	25	2575	0	45	18	4179	72	15	102	0	4	0	0	3802	2	277	680	12	25	650	3	319	9	m	15.543	12
n	5939	193	0	640	116	1970	0	0	4386	2	62	25	3535	478	253	715	122	346	21	0	2822	58	881	982	364	695	673	1241	817	247	n	27.583	5
nj	1267	3	0	13	4	165	0	0	490	0	10	0	268	12	195	0	1	38	0	0	26	7	20	0	276	28	20	4	0	37	nj	2.884	24
o	1975	1173	70	199	48	1966	21	53	488	309	2770	446	1876	614	7181	2702	23	2740	6007	93	39	5292	2959	568	55	4445	130	3454	302	44	o	48.042	4
p	772	0	0	0	0	6	0	0	458	0	0	0	337	47	4	13	0	239	15	0	1008	1	248	894	18	112	706	0	0	0	p	4.878	18
r	2950	797	366	0	0	1538	0	0	2578	207	1101	309	1251	8	1076	0	0	175	30	0	3392	4432	13	574	2	2129	613	1237	130	34	r	24.942	6
s	1697	3	0	0	0	406	0	0	1732	20	2	1	2802	448	306	84	57	110	974	28	3703	383	500	30	0	79	1189	109	0	0	s	14.663	14
š	1022	1	0	0	0	1	0	0	554	0	0	0	812	15	7	3	6	11	9	0	721	104	202	0	0	1	837	88	0	0	š	4.394	21
t	3037	0	21	0	0	2	0	0	2334	6	11	63	2926	47	549	267	2	15	1245	0	1351	80	436	6661	1782	22	837	5	1	0	t	21.700	8
u	289	433	316	304	218	778	7	479	140	104	536	44	0	2687	1592	897	523	871	1113	298	76	568	1670	2002	110	1456	5	553	382	89	u	18.540	9
v	3605	14	18	9	0	424	0	0	1274	0	17	160	1708	116	442	11	0	1	23	0	4728	0	562	1766	27	1203	293	0	450	6	v	16.857	11
z	1275	35	0	0	0	12	0	0	806	2	20	0	2213	47	1	2	0	1	101	0	662	0	100	0	0	24	433	0	7	0	z	5.741	17
ž	528	0	0	0	0	0	0	0	236	0	0	0	188	0	0	2	0	0	12	0	514	0	541	0	0	0	305	0	0	0	ž	2.326	27
	a	b	c	č	ć	d	dž	đ	e	f	g	h	i	j	k	l	lj	m	n	nj	o	p	r	s	š	t	u	v	z	ž		
	43.862	8.149	5.175	5.545	2.972	18.317	150	1.412	30.422	1.627	7.954	1.795	41.266	27.035	20.598	19.265	2.589	14.512	35.092	3.867	41.005	14.869	26.886	26.532	5.331	24.251	12.397	19.971	8.289	3.215		474.350
rank	1	18	22	20	25	13	30	29	5	28	19	27	2	6	10	12	26	15	4	23	3	14	7	8	21	9	16	11	17	24		

The following table gives frequency of all combinations of neighboring letters in a unified text of all authors and columns. The table is read by first looking at the column, and then at the row. The cross section gives the number of those combinations in the text. For example, the combination "do" was found 1966 times. Fields with number of noticed combinations were conditionally colored in a way that in each column the least frequent are red, and as the frequency increases, the color proceeds into white, and finally into green. In the table, it is easy to distinguish, by red color, combinations of two letters that do not appear at all, or very rarely, and those in green, which are the most frequent ones (in a column).

It may be noticed that two consecutive letters most often have letter "a" as the first (43.862 times), followed by i, o, n, e, j, etc., while least frequent were letters: dž (150 times), đ, f, h, lj, etc.

In combination of two consecutive letters, the most frequent second letter is "a" (65.003 times), followed by e, i, o, n, r, etc., and least frequent were letters: dž (99 times), f, đ, ž, lj, h, etc.

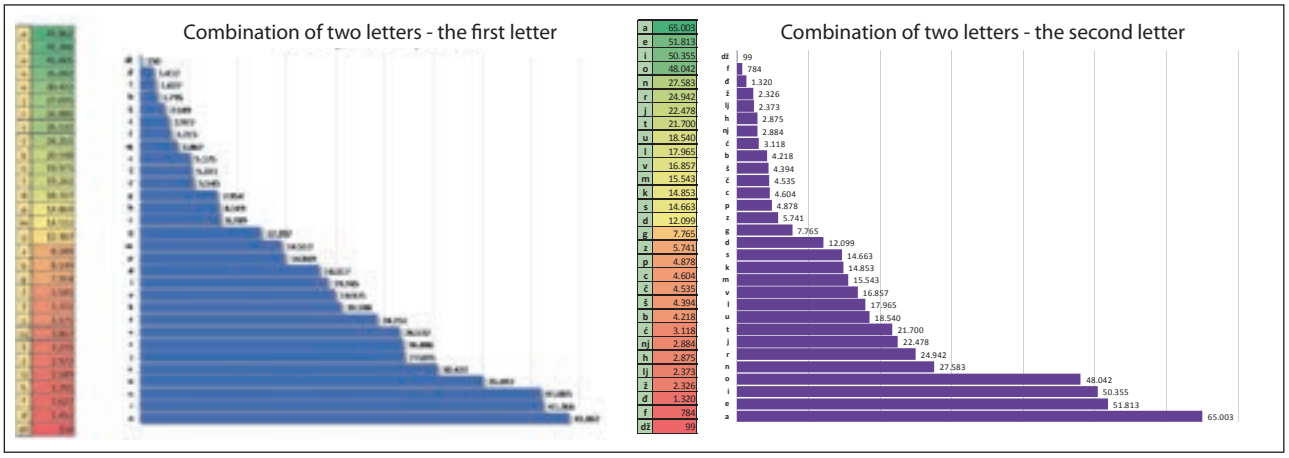


Figure 2. *Order of combinations of two consecutive letters*

Table 11. *Frequency of letters of individual columnists – alphabetically and in descending order*

			Filipović	Jergović	Apostolovski	Lekić	SVI
1	a	A	17.213	15.795	17.140	16.067	66.215
2	b	B	2.168	1.952	2.266	1.748	8.134
3	c	C	1.132	1.248	1.445	1.549	5.374
4	č	Č	1.269	1.610	1.419	1.355	5.653
5	ć	Ć	961	809	1.227	630	3.627
6	d	D	5.506	4.359	5.271	4.616	19.752
7	dž	Dž	53	21	56	28	158
8	đ	Đ	371	293	404	345	1.413
9	e	E	11.923	12.619	12.539	12.807	49.888
10	f	F	422	341	372	545	1.680
11	g	G	2.426	2.412	2.504	2.553	9.895
12	h	H	1.025	1.129	800	823	3.777
13	i	I	14.615	15.252	14.051	14.503	58.421
14	j	J	6.000	5.960	4.600	5.451	22.011
15	k	K	5.062	5.471	5.646	5.239	21.418
16	l	L	3.913	4.070	4.624	4.241	16.848
17	lj	Lj	612	804	584	701	2.701
18	m	M	5.333	5.036	4.469	5.047	19.885
19	n	N	8.048	8.497	8.019	8.605	33.169
20	nj	Nj	949	1.027	841	1.056	3.873
21	o	O	14.294	13.591	13.509	12.784	54.178
22	p	P	3.465	3.432	4.075	3.911	14.883
23	r	R	6.426	6.356	7.173	7.896	27.851
24	s	S	6.973	6.818	6.822	6.853	27.466
25	š	Š	1.412	1.528	1.399	971	5.310
26	t	T	6.651	6.965	5.974	6.615	26.205
27	u	U	5.806	5.564	6.141	5.843	23.354
28	v	V	4.974	5.102	5.233	4.748	20.057
29	z	Z	2.238	2.203	2.291	2.386	9.118
30	ž	Ž	913	885	752	642	3.192
			142.153	141.149	141.646	140.558	**565.506**

sort

		Filipović
a	A	17.213
i	I	14.615
o	O	14.294
e	E	11.923
n	N	8.048
s	S	6.973
t	T	6.651
r	R	6.426
j	J	6.000
u	U	5.806
d	D	5.506
m	M	5.333
k	K	5.062
v	V	4.974
l	L	3.913
p	P	3.465
g	G	2.426
z	Z	2.238
b	B	2.168
š	Š	1.412
č	Č	1.269
c	C	1.132
h	H	1.025
ć	Ć	961
nj	Nj	949
ž	Ž	913
lj	Lj	612
f	F	422
đ	Đ	371
dž	Dž	53
		142.153

sort

		Jergović
a	A	15.795
i	I	15.252
o	O	13.591
e	E	12.619
n	N	8.497
t	T	6.965
s	S	6.818
r	R	6.356
j	J	5.960
u	U	5.564
k	K	5.471
v	V	5.102
m	M	5.036
d	D	4.359
l	L	4.070
p	P	3.432
g	G	2.412
z	Z	2.203
b	B	1.952
č	Č	1.610
š	Š	1.528
c	C	1.248
h	H	1.129
nj	Nj	1.027
ž	Ž	885
ć	Ć	809
lj	Lj	804
f	F	341
đ	Đ	293
dž	Dž	21
		141.149

sort

		Apostolovski
a	A	17.140
i	I	14.051
o	O	13.509
e	E	12.539
n	N	8.019
r	R	7.173
s	S	6.822
u	U	6.141
t	T	5.974
k	K	5.646
d	D	5.271
v	V	5.233
l	L	4.624
j	J	4.600
m	M	4.469
p	P	4.075
g	G	2.504
z	Z	2.291
b	B	2.266
c	C	1.445
ć	Č	1.419
š	Š	1.399
ć	Ć	1.227
nj	Nj	841
h	H	800
ž	Ž	752
lj	Lj	584
đ	Đ	404
f	F	372
dž	Dž	56
		141.646

sort

		Lekić
a	A	16.067
i	I	14.503
e	E	12.807
o	O	12.784
n	N	8.605
r	R	7.896
s	S	6.853
t	T	6.615
u	U	5.843
j	J	5.451
k	K	5.239
m	M	5.047
v	V	4.748
d	D	4.616
l	L	4.241
p	P	3.911
g	G	2.553
z	Z	2.386
b	B	1.748
c	C	1.549
č	Č	1.355
c	C	1.056
nj	Nj	971
š	Š	823
h	H	701
lj	Lj	642
ž	Ž	630
ć	Ć	545
f	F	345
đ	Đ	28
dž	Dž	140.558

sort

		SVI
a	A	66.215
i	I	58.421
o	O	54.178
e	E	49.888
n	N	33.169
r	R	27.851
s	S	27.466
t	T	26.205
u	U	23.354
j	J	22.011
k	K	21.418
v	V	20.057
m	M	19.885
d	D	19.752
l	L	16.848
p	P	14.883
g	G	9.895
z	Z	9.118
b	B	8.134
č	Č	5.653
c	C	5.374
š	Š	5.310
nj	Nj	3.873
h	H	3.777
ć	Ć	3.627
ž	Ž	3.192
lj	Lj	2.701
f	F	1.680
đ	Đ	1.413
dž	Dž	158
		565.506

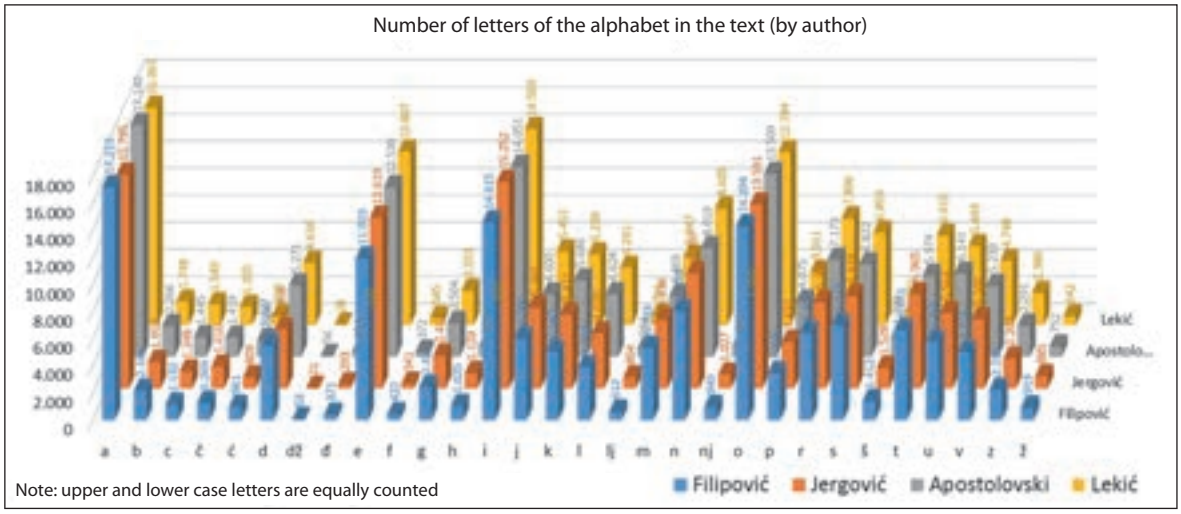


Figure 3. *Number of particular alphabet letters in texts of the column authors*

Analysis of Frequency of Certain Letters in a Text

Frequency of particular letters in a text may depend on several factors. Among them are: language of use, type of text (prose, poetry, technical text…), theme of the text, author's style and the usual fund of words the author uses, etc. In the Internet, one can find pages [6] and programs to get the number of particular letters in the given text, but a great deal of those tools does not recognize our characters č, ć, dž, đ, lj and nj. In that case, possibilities offered by text processors may be used.

Here is the analysis of frequency of particular letters for each of the columnists separately.

Order of letters per individual author:

Filipović:	aioenstrjudmkvlpgzbščchćnjžljfđdž
Jergović	aioentsrjukvmdlpgzbčšchnjžćljfđdž
Apostolovski	aioenrsutkdvljmpgzbcčšćnjhžljđfdž
Lekić	aieonrstujkmvdlpgzbcčnjšhljžćfđdž
ALL	aioenrstujkvmdlpgzbčcšnjhćžljfđdž

It is indicative to compare these orders with the orders given in Table 3: Graphic presentation of frequency of letters in particular languages.

Conclusion

Linguistics deals with language (French *linguistique,* from Latin *lingua* – language), which may be subdivided to: phonetics (science of sounds), phonology (science of function of sounds), morphology (science about forms of language units), syntax (science about organization of sentence), semantics (science about meanings in language), etc. Linguistics is multidisciplinary science, and therefore its specialized branches emerged in the 20th, such as: mathematical linguistics, psycholinguistics, sociolinguistics, neurolinguistics. In this paper we statistically analyzed the language of four distinguished columnists in electronic media. They wrote about different topics, with different language styles, but there were noticeable large similarities in some segments. Such results give space to conclude that it is one, single, polycentric language, which is not a rare phenomenon in the modern world. Naturally, the final word about this should be given by linguists. Polycentric languages are English (Great Britain and the USA), German (Germany, Austria, Switzerland), French (France, Canada (Québec), Belgium), Spanish (Spain, Argentina, Mexico), Persian (Iran, Afghanistan, Tajikistan), Portuguese (Brazil, Portugal), Arabic (Saudi Arabia, Iran, Iraq, Tunisia, Egypt…), etc. Each of the variants of any polycentric language has its standard national variation, grammar and orthography, distinguishable according to some differences. [5]

There are also monocentric languages, such as Japanese or Russian, which do not have more standardized variations.

Modern electronics enables language analyses that could not be imagined until recently, which will probably impact the creation of new branches of linguistics and new findings about the language.

REFERENCES

[1] Ethnologue: Jezici svijeta, sedamnaesto izdanje. Dallas, Teksas: SIL International. Online verzija: http://www.ethnologue.com.
[2] http://www.politika.rs/scc/authors/texts/901
[3] https://avaz.ba/tag/4975/muhamed-filipovic
[4] https://en.wikipediam.org/wiki/Letter_frequency
[5] https://hr.wikipedia.org/wiki/Policentri%C4%8Dni_standardni_jezik
[6] https://norvig.com/mayzner.html
[7] https://www.jutarnji.hr/autori/miljenko-jergovic
[8] https://www.vijesti.me/autor/miodrag-lekic
[9] Jahić, Dž. (1999). Trilogija o bosanskom jeziku. Knj. 3, Školski rječnik bosanskog jezika Sarajevo: Ljiljan biblioteka Linguos.

Submitted: May 20, 2019
Accepted: May 24, 2019

ABOUT THE AUTHORS

Nedim Smailovic was born in Tuzla. He has been living in Banja Luka since 1973. He graduated from the Faculty of Electrical Engineering, department of Telecommunications. Since 1982 he has worked in RO PTT traffic of Bosnia and Herzegovina, and a series of organizational transformation it is now called Mtel doo Banja Luka. His first work experience was in designing and maintaining the PTT capacities. He obtained his Master's degree from Pan European University 'Aperion' Banja Luka, in 2005. There he also defended his doctoral thesis titled: Computer information graphics in presenting Bosnia and Herzegovina on the road to accessing the European Union. He was elected Associate Professor in 2013 and he has been teaching since in three universities in Bosnia and Herzegovina subjects relating to computer technology. He is an author and co-author of several books from the field of information technology and mathematics. He is married, father of two daughters.

FOR CITATION

Nedim Smailović, Statistical Analysis of Texts of the Balkans Electronic Media Columnists, *JITA – Journal of Information Technology and Applications Banja Luka*, PanEuropien University APEIRON, Banja Luka, Republika Srpska, Bosna i Hercegovina, JITA 9(2019) 1:5-16, (UDC: 659.3/.4:316.776]:004.738.5), (DOI: 10.7251/JIT1901005S), Volume 9, Number 1, Banja Luka, june 2019 (1-48), ISSN 2232-9625 (print), ISSN 2233-0194 (online), UDC 004