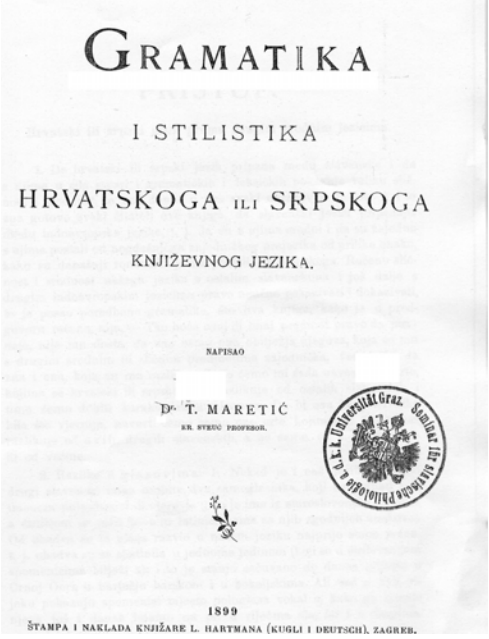
XXIV. Tomo Maretić's first Croatian and/or Serbian Sound Statistics (1899)

The first Croatian and/or Serbian sound statistic we know of was published by the renowned linguist Tomo Maretić in 1899.

Tomo Maretić is considered to be one of the greatest of Croatian linguists. He was born on October 13, 1854 in Virovitica, a small Croatian town in the region of Slavonia. Here, he attended primary school, and then attended high school at Varaždin, Slavonska Požega and Zagreb. In 1875, he enrolled to study Slavic languages, and simultaneously Latin and Greek, at the Philosophical Faculty of Zagreb University.

Having passed his exams in Classical Philology, intending to become a middle school teacher, he began his Ph.D. studies in Slavistics, and received his doctorate in 1883. After spending some time at well-known European universities, namely in the Neogrammarians' center of Leipzig (where Leskien was the leading Slavist), and Prague, Tomo Maretić first was appointed extraordinary professor for "Slavic philology with particular emphasis on Croatian and Serbian history of language and literature", in 1886, and ordinary professor three years later. Also in 1886, he became a corresponding member of the Yugoslav Academy of Science and Art, to which he was elected a full member four years later. Later, he twice became head of the philological-historical class of the academy, first from 1906-1913, then a second time from 1919-1928, after being president of the Academy in the years 1915-1918. Tomo Maretić died on January 15, 1938.

Maretić's linguistic oeuvre comprises more than a hundred important contributions to Slavic languages and literatures in general, and to Serbian and Croatian philology in detail. It is impossible to mention all his works here; only some major achievements can be cited by way of example, in order to demonstrate the broad spectrum of his interests. Maretić's interest in linguistics can be traced back to the late 1870s when he published his first work on accentology. Basically, this study was an addendum to one of his many translations from Greek; in his commentary, Maretić attempted to prove an earlier claim (by writer, translator and theoretician Ivan Trnski) that Croatian prosody is based on a particular type of accent (*štokavski*), rather than on the quantity of syllables. Later, Maretić continued this line of research, for example in his study *O nekim pojavama kvantitete i akcenta u jeziku hrvatskom ili srpskom* (1883). In the early 1880s, he published some important works on folkore, *O narodnoj zagonetci hrvatskoj* (1881) and *Studije iz pučkog vjerovanja i pričanja Hrvata i Srba* (1882). A third major field of interest, in addition to accentology and folklore, was lexicology; this is well documented in works like *O narodnim imenima i prezimenima u Hrvata i Srba* (1886), *Ruske i češke riječi u književnom hrvatskom jeziku* (1892) or *Imena rijeka i potoka u hrvatskim i srpskim zemljama* (1893). Maretić's lexicological and lexicographic competence is best proven, of course, in his engagement in the *Rječnik hrvatskoga ili srpskoga jezika*: Maretić edited six of the 23 volumes of this outstanding dictionary, the compilation of which had begun in 1880 on the initiative of Đuro Daničić. As a professor at the Philosophical Faculty of Zagreb University, Maretić was, however, obliged to work not only on linguistics, but on literature, as well; thus, his early interest in all aspects of philology in a broad understanding of this term, continued in his later works, as well. In his literary studies, however, Maretić rather concentrated on folk literature, as for example in his well-known *Metrika narodnih naših pjesama* (1909). Maretić's official academic carreer ended in 1914; yet, he returned to university



GRAMATIKA

I STILISTIKA

HRVATSKOGA ILI SRPSKOGA

KNJIŽEVNOG JEZIKA.

NAPISAO

Dr. T. MARETIĆ

KR. SVEUČ. PROFESOR.

1899

ŠTAMPA I NAKLADA KNJIŽARE L. HARTMANA (KUGLI I DEUTSCH), ZAGREB.

from 1919-24, teaching Indo-European studies Irrespective of the high value of all these studies, which are very much appreciated still today, Maretić's magnum opus remains his *Gramatika i stilistika hrvatskoga ili srpskoga jezika* (1899, ^{2}1932, ^{3}1963). This book served as the basis for linguistic education of several generations of Croatian and Serbian linguists. It was in this book that, as far as we know, the first Croatian or Serbian statistics of this kind were published.

In his introductory ruminations on the overall position of Croatian or Serbian in the context of other Slavic languages, Maretić mainly concentrates on differences with regard to sounds and cases.
The discussion of sounds is then intensified in the first chapter (p. 9ff). According to his presentation, Croatian or Serbian is characterized by 31 sounds, one of which has no alphabetic correspondence, neither in the Latin nor the Cyrillic variant. This particular sound [ç], according to Maretić, occurs quite rarely; we are concerned here with an assimilating consonant, which may precede the sound [c] – we would rather denote it as [ts] today –, just like [d] preceding [t] or [g] preceding [k]. Maretić conseqeuntly assumes the basic sound inventory to consist of the following 30 items, which are identical with the corresponding 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
а, б, в, г, д, ђ, е, ж, з, и, ј, к, л, љ, м, н, њ, о, п, р, с, т, ћ, ѕ, ф, х, ц, ч, џ, ш

Starting from the more or less basic observation that the frequency of individual sounds differs for any given language, Maretić presents a table of the frequency of Croatian or Serbian sounds. Maretić selected ten passages per 1,000 sounds (i.e., random samples) from Vuk Karadžić's translation of the *New Testament* and then counted the frequency of all sounds in the order of their occurence. The authenticity of this dana material is not unproblematic. After all, Vuk's translation, published in Vienna in 1847, was based on an earlier «Serbian» translation by Atanasije Stoiković, a Serbian writer who worked as a professor at the Russian university of Kharkov. Stojković's translation, however, which had been published by the Russian Bible Society in Saint Petersburg in 1824, was not written in the vernacular, but represented a mixture of Church Slavonic and Serbian (Slavenoserbian). In contrast to Maretić's problematic choice of his text basis, he displayed an enormous carefulness and a suprisingly high degree of methodological reflection in treating this material: according to his information, the total sum of 10,000 sounds was based on approximately 2,000 words; assuming that it would take about half an hour to pronounce this amount of material, Maretić assumed it to be sufficient for his purposes, referring to the similar approach by William Dwight Whitney in his analysis of Old Indian sounds.[1] In order to prevent particular biases, Maretić selected only passages in which one and the same words were repeated as rarely as possible, referring to the fact that in the specific text of the *New Testament*, the occurrence and repetition of specific proper nouns may significantly change the frequency structure; for

[1] Maretić referred to the German translation of Whitney's *Sanskrit Grammar: Including Both the Classical Language, and the Older Dialects, of Veda and Brahmana,* published under the title of *Indische Grammatik: umfassend die klassische Sprache und die älteren Dialecte* (Leipzig, 1879).

the same reason, he tried to avoid passages with foreign words. Further, in order to prevent ordinary errors, he counted all passages more than once. As to the definition of the items counted, it seems important to take into consideration that Maretić counted all sounds *as they are pronounced, not as they are written*. Therefore he interpreted *s njim* as [š njim], *bratski* and *gradski* as [bracki] or [gracki], respectively; furthermore, words like *donio* or *mislio* are interpreted as [donijo] and [mislijo]. The sound [r] is counted separately, depending on whether it fulfils a vowel or a consonant function *prst* vs *ruka*.

As a result, the sound frequencies indicated in Table 1 are obtained.

	I	II	III	IV	V	VI	VII	VIII	IX	X	total
a	99	127	105	103	114	101	106	104	120	100	1079
b	11	21	10	16	17	15	21	21	23	18	173
c	11	8	5	8	11	2	2	1	3	5	56
ć	24	9	18	12	12	11	16	5	9	4	120
d	2	9	11	9	18	6	8	8	7	13	91
dž (џ)	41	30	45	44	41	50	42	40	53	42	428
đ	0	0	0	0	0	0	1	0	0	0	1
e	104	108	120	113	116	115	89	116	107	111	1099
g	15	17	23	21	11	17	19	23	24	22	192
h	4	9	14	8	9	7	6	7	3	6	73
i	113	119	109	120	103	100	108	81	89	93	1035
j	62	56	42	48	47	52	37	58	51	50	503
k	30	41	28	24	38	34	28	28	33	34	318
l	15	25	16	21	11	18	15	31	23	19	194
lj (љ)	9	12	5	13	6	6	12	11	9	7	90
m	29	36	46	37	54	43	33	34	25	30	367
n	40	41	46	43	48	49	55	47	50	38	457
nj (њ)	3	5	3	10	2	11	10	5	5	9	63
o	104	79	87	91	99	100	100	101	84	115	960
p	30	12	18	26	10	26	30	22	18	34	226
r (vokal.)	5	2	7	5	2	7	5	1	3	8	45
r (kons.)	44	35	41	18	27	36	34	38	32	38	343
s	47	31	42	45	40	44	47	62	59	47	464
š	22	14	13	20	14	13	21	10	9	14	150
t	37	49	29	34	46	31	43	49	38	44	400
u	51	43	43	49	40	33	45	35	53	37	429
v	32	43	41	41	46	48	39	35	38	39	402
z	11	16	21	10	13	11	18	19	16	12	147
ž	2	2	6	5	3	9	8	7	12	10	64
	997	999	994	994	998	995	998	999	996	999	

Table 1 represents the frequency for each of the ten passages, as well as the total of the combined samples. The sounds [f] and [ç] are not listed, since they did not occur once in any of the samples.

Mentioning that the frequency of the sounds [i] would be slightly higher, and the frequency of [j] slightly lower, in case the *ekavian* or *ikavian* variant of the analyzed texts were taken, Maretić in conclusion concentrates on the vowel-consonant proportions. For the totality of all then texts, this proportion is 46.47% vowels as compared to 53.53% consonants. Maretic interprets this ratio in terms of a language's weakness or hardness, based on the assumption that a higher percentage of consonants renders a language "harder", whereas a higher percentage of vowels makes it "weeker" – meaning that it is more or less easy to pronounce a word. In this respect, Maretić then compared the results obtained with similar data he obtained for German (38.86), Polish (41.43), Russian (41.69), Lithuanian (43.00), Czech and Latin (43.09), French (43.36) Greek (46.01), Italian (47.73), and Old Church Slavonic (48.37), as well as for the Old Indian (43.52) data provided by Whitney (see above). According to Maretić's interpretation, the weakness of Croatian or Serbian is not very different from Italian, and even exceeds Greek, which usually is considered to be one of the most euphonic languages. Having to concede in this context that there may be hard-to-pronounce consonant clusters in Croatian or Serbian, such as *k bratu*, *k zdravome*, *kumstvo*, and others, Maretić refers to the rareness of these cases, on the one hand, and to equivalent German examples such as *Angst*, *herbstlich*, on the other.

Despite Maretić's enormous productivity in the field of lingiustics, his analysis of sound frequency reported here remained his only systematic study in this direction; comparable follow-up studies were conducted only decades later. It would lead too far, here, to deal with these later studies; rather, by way of a conclusion, a cautious first attempt to re-analyse Maretić's data shall be presented, showing some remaining problems and pointing out future tasks for systematic study.

Figure 1 represents the result of fitting the negative hypergeometric function to the whole corpus of texts. This distribution is chosen since it has repeatedly turned out to be the best model for Slavic letter, sound, and phoneme frequencies. As can easily be seen, the fit is far from being convincing, with $C = X^2 / N = 0.034$. To be sure, not any one of the otherwise discussed models (such as geometric, Zipf, Zipf-Mandelbrot, and others) yields more satisfying results. Figure 1 very clearly shows the crucial positions responsible for the largest deviations: positions 2, 3, and 4, being clearly underestimated, and positions 1, 5, and 6, clearly overestimated.

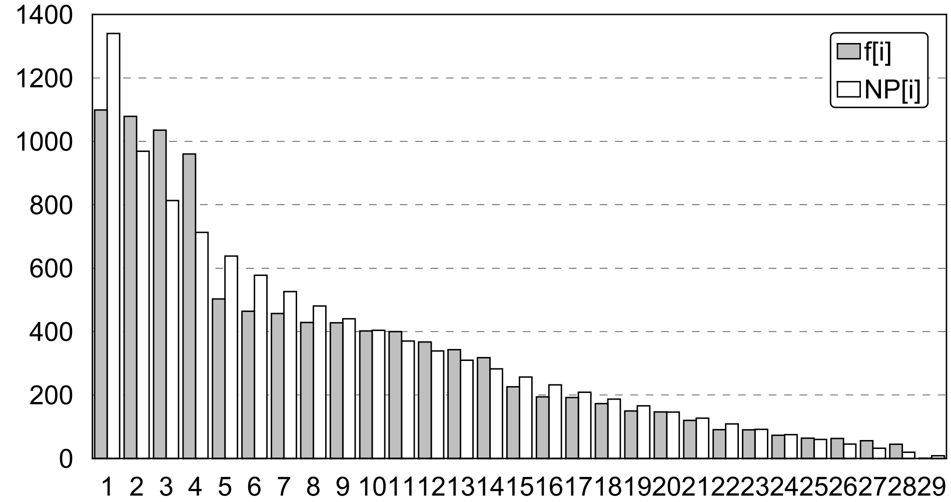


Fig. 1: Fitting the negative hypergeometric distribution to the combined corpus of 10 samples

Yet, analyzing all ten samples individually provides a surprising result: fitting the negative hypergeometric distribution under these circumstances, statistically satisfying results are obtained in all ten cases, though with extremely varying X^2 values (ranging from $0.07 \leq X^2 \leq 0.99$. These findings raise two important questions of rather general kind:

1. The question of data homogeneity comes into play, taking into consideration the good results for the individual texts, on the one hand, and the poor result for the combined corpus.
2. Taking into consideration the diverging X^2 values of the ten individual samples, the question of sample size might turn out to be important, perhaps being the reason for the varying goodness of fit.

With this perspective in mind, a closer look at sample #9 (cf Fig. 2), representing the relatively best statistical result, clearly shows that, despite the good value, it is still the same range of positions which displays the same kind of deviations characteristic of the combined corpus (cf. the deviations in either direction, with the significant overcrossing at position 5).

Thus, in addition to the obviously arising questions of data homogeneity and sample size, mentioned above, we seem to be concerned with some additional problems leading us back to the issues discussed above:

3. Is the chosen text basis authentic language material for Croation or Serbian sound frequencies?
4. Are the definitions of ‘sound’ valid, and, as a consequence, is the sound inventory as a whole adequately defined?

It seems that there are more questions than answers. In any case, it was Tomo Maretić who, in 1899, asked these important questions with regard to Croatian and / or Serbian, questions which deserve further attention.

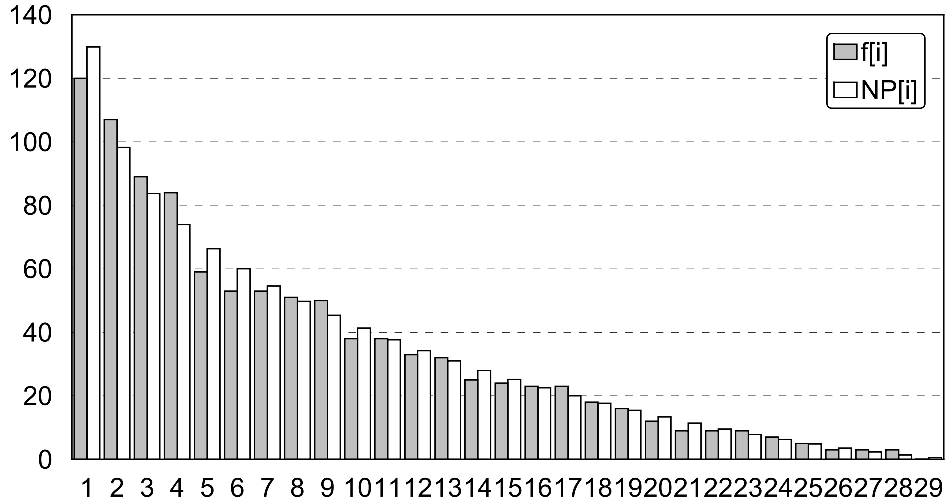


Fig. 2: Fitting the negative hypergeometric distribution to the sample #9

References

Babić, Stjepan (1993). Tomo Maretić 1854-1938. In: *Portreti hrvatskih jezikoslovaca*. Zagreb, 137-143.

Belić, Aleksandar (1939). Dr. Tomislav Maretić. *Naš jezik 6(3), 65-69.*

Belić, Aleksandar (1938/39). Dr. Tomislav Maretić. 13.X.1854 – 15.1.1938. *Južnoslovenski filolog 17, 218-221.*

Ivšić, Stjepan (1940). *Prof. dr. Tomislav Maretić*. Zagreb.

Silić, Josip (1984). Tomo Maretić – jedan od najvećih hrvatskih lingvista. In: *Virovitički zbornik 1234-1984.* Virovitica, *395-403.*

Skok, Petar (1949). Tomo Maretić 1854-1938. *Ljetopis jugoslavenske akademije znanosti i umjetnosti 54, 310-335.*

Skok, Petar (1952). Tomo Maretić (1854-1938). In: *Ljetopis jugoslavenske akademije znanosti i umjetnosti za godine 1949-1950, 56, 319-326.*

Peter Grzybek

Glottometrics 13

2006

RAM-Verlag

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitschrift für die quantitative Erforschung von Sprache und Text

Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden

Glottometrics kann aus dem **Internet** heruntergeladen, auf **CD-ROM** (in PDF Format) oder in **Buchform** bestellt werden

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals

Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors

Glottometrics can be downloaded from the **Internet**, obtained on **CD-ROM** (in PDF) or in form of **printed copies**

Herausgeber – Editors

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an
Orders for CD-ROM´s or printed copies to

RAM-Verlag RAM-Verlag@t-online.de

Herunterladen / Downloading: http://www.ram-verlag.de

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Glottometrics. –13 (2006) –. – Lüdenscheid: RAM-Verl., 2006
Erscheint unregelmäßig. – Auch im Internet als elektronische Ressource unter der Adresse http://www.ram-verlag.de verfügbar.-
Bibliographische Deskription nach 13 (2006)
ISSN 1617-8351