Anton Perdih

LINGUISTIC ANALYSIS BASED ON THE FREQUENCY OF SOUND PAIRS AND TRIPLETS

Povzetek

JEZIKOVNE ANALIZE NA OSNOVI POGOSTOSTI GLASOV, DVOJČKOV IN TROJČKOV GLASOV

Na podlagi analize pogostosti glasov v 17 jezikih so ugotovljene meje, nad katerimi je velikost baze podatkov dovolj velika, da njena velikost ne vpliva več bistveno na rezultate izvedene iz pogostosti glasov, njihovih parov in trojčkov. Te meje so: več kot 700 posameznih glasov; več kot 8.000 parov glasov; več kot 30.000 trojčkov glasov.

Kriteriju za posamezne glasove ustrezajo vse uporabljene baze podatkov. Kriteriju za pare glasov ne ustrezajo baze podatkov za oskijski, starofrigijski, retijski in venetski jezik. Kriteriju za trojčke glasov ne ustrezajo tu uporabljene baze podatkov za etruščanski, hetitski, luvijski, mikenski, oskijski, starofrigijski, retijski, staroslovenski, umbrijski in venetski jezik. Zato so pri teh jezikih uporabni predvsem rezultati na podlagi pogostosti posameznih glasov. Selektivnost pristopa pa narašča v smeri: posamezni glasovi < pari glasov < trojčki glasov.

Na podlagi analize pogostosti glasov se kaže, da mikenska pisava Linear B in morebiti tudi luvijska pisava še nista dovolj dobro razvozlani in da bi bilo dobro pri njunem razvozlavanju upoštevati tudi slovanske pare glasov tipa soglasnik-soglasnik ter trojčke glasov tipa soglasnik-soglasnik-samoglasnik in soglasnik-soglasnik-soglasnik.

Introduction

Linguistic distance is a means to demonstrate the degree of similarity resp. dissimilarity of the languages in question. In principle, several language characteristics can be used for this purpose. For the comparison of some ancient languages with modern ones, only sound frequencies can be used since some ancient languages are known from a relatively small number of inscriptions, which are mostly short, broken or incomplete, making the composition of an extended and comprehensive linguistic Corpus difficult. In addition, a number of groups of inscriptions are written *in continuo*, i.e. without separation in words, and do not give any suitable clue about toponyms, verbs, and frequently used words that could be used for computational comparisons between these old languages and other better known languages.

For this reason, the average sum of absolute values of frequency differences based on few sets of data and on data for single sounds only was used [1, 2], resp. the normalized PCA [3]. Later on [4], the usefulness of six methods for estimating the linguistic distances

between 17 mostly ancient languages based on sound frequencies was demonstrated, not only on particular sounds, but also on sound pairs and triplets. The tested methods were: Principal Component Analysis (PCA), the sum of absolute values of frequency differences (SuD), the root-of-sum-of-square frequency differences (SuS), the correlation coefficient (R), the Fisher ratio (F), and the standard error of estimation (STE). This study [4] gave rise to a disturbing result, as well. Namely, the language distances estimated on the basis of frequency of sound pairs and especially on sound triplets gave different results than those based on frequency of single sounds. One obvious reason for this is the following. Among the languages, which are written *in continuo* and with no fixed word separation rules, there may be counted, depending on the choice of division of the continuous text into words, also too few or too many sound pairs resp. triplets. So the results based on counting sound pairs resp. triplets must be expected to be less plausible than those based on counting single signs.
In present paper, the validity of previous [4] results is tested from other points of view, including the dependence on the size of the database.

Data and methods

The sound frequency data of languages Bq, Cs, Es, Et, Fi, Gr, Hi, La, Lu, My, Os, Ph, Rt, Sl, Um, Ve, and Vz, are used as prepared for a previous study [3]. The meaning of these abbreviations is presented in Table 1 taken from ref. [3], where also the data about the number of characters, their pairs and triplets are presented. Some of these languages are studied in different reading variants, marked EtB, EtT, LaC, LaS, PhA, PhT, RtB, RtT, RtV, VeB, VeT, or VeV. The third character in these combinations indicates the following, cf. [3] for detailed references:

A in PhA – the reading according to A. Ambrozic is applied to all considered inscriptions by A. Perdih;

B in EtB, RtB, VeB – the reading according to M. Bor is applied to all considered inscriptions by A. Perdih;

C in LaC – classical reading of Latin;

S in LaS – semiclassical reading of Latin;

T in EtT, PhT, RtT, VeT – the reading according to western mainstream scholars is prepared by G. Tomezzoli;

V in RtV and VeV – the reading by V. Vodopivec.

These languages as such are marked as Et, La, Ph, Rt, or Ve.

As the regression quality indicator the correlation coefficient R is used.

For the purpose of this paper, there are considered all sound pairs and triplets, regardless whether they are syllables or not. They are divided into several groups by the number of vowels (v) and / or consonants (c). The sound pairs are divided into groups: vowel-vowel marked as (vv), vowel-consonant marked as (vc), consonant-vowel marked (cv), and conconant-consonant marked (cc). Marking of sound triplets is analogous.

Table 1: Language abbreviations, number of countable sounds, their pairs and triplets in the Language Databases in [4]

Language Database	Abbreviation	Number of countable sounds		
		single	pairs	triplets
Basque	Bq	160.177	130.866	101.577
Old Church Slavonic	Cs	458.319	362.444	278.990
Estonian	Es	90.742	76.108	61.485
Etruscan	EtB, EtT	30.421	24.227	18.445
Finnic	Fi	449.075	381.686	314.298
Greek	Gr	117.109	93.503	71.502
Hittite	Hi	14.001	11.509	9.025
Latin Classic	LaC	1.029.312	848.168	667.718
Latin Semiclassic	LaS	1.019.977	838.833	658.383
Luvian	Lu	32.626	27.254	21.942
Mycenean	My	26.330	22.474	18.618
Oscan	Os	3.057	2.418	1.841
Old Phrygian	PhA	2.290	1.698	1.172
Old Phrygian	PhT	2.242	1.834	1.459
Rhaetic	RtB	2.102	1.719	1.394
Rhaetic	RtT	1.948	1.572	1.265
Rhaetic	RtV	2.097	1.754	1.440
Old Slovene	Sl	19.834	15.428	11.301
Umbrian	Um	25.063	20.657	16.288
Venetic	VeB	7.651	6.083	4.965
Venetic	VeT	7.427	6.119	4.843
Venetic	VeV	7.113	4.855	2.993
Venezian	Vz	320.794	234.563	153.903

Results

Sounds, sound pairs and triplets are counted in two different ways. The first way is counting of all observed sounds, sound pairs and triplets. The second way is counting of all different sounds, sound pairs and triplets.

Whereas in the former case all of them are counted wherever they appeare, in the latter case, for example, each of the sound pairs aa, ea, uu is counted only once, regardless of how many times it appears in the database.

Number of all observed sounds, sound pairs and triplets

The number of all observed sounds, sound pairs and triplets is presented in Table 1. Them and their subgroups are presented in Tables 2-4.

In Table 2 can be seen that among all tested languages, except RtT, the number of all vowels resp. consonants exceeds one thousand.

In Table 3 can be seen, however, that the number of observed consonant-consonant sound pairs is quite small in Myceaenan, but also in other ancient languages some sound pair groups do not exceed the number 200.

Table 2: Number of observed particular sounds

Language	all	(v)	(c)
Bq	160177	81926	78251
Cs	458319	223434	234885
Es	90742	44009	46733
EtB	30421	14316	16105
EtT	30421	12944	17477
Fi	449075	220624	228451
Gr	117109	60064	57045
Hi	14001	6850	7151
LaC	1029312	485747	543565
LaS	1019977	476000	543977
Lu	32626	17598	15028
My	26330	16571	9759
Os	3057	1362	1695
PhA	2290	1221	1069
PhT	2242	1201	1041
RtB	2102	1084	1018
RtT	1948	1005	943
RtV	2097	1083	1014
Sl	19834	9870	9964
Um	25063	11930	13133
VeB	7651	3801	3850
VeT	7427	3540	3887
VeV	7113	3712	3401
Vz	320794	157117	163677

(v) – number of observed vowels
(c) – number of observed consonants

Table 3: Number of observed sound pairs

Language	all	(vv)	(vc)	(cv)	(cc)
Bq	130866	11982	55632	56409	6843
Cs	362444	46662	104473	152257	59052
Es	76108	8267	26554	33312	7975
EtB	24227	2730	8427	9874	3196
EtT	24227	1366	8694	10169	3998
Fi	381686	47581	131605	159495	43005
Gr	93503	12864	35897	36559	8183
Hi	11509	1448	4139	4602	1320
LaC	848168	82142	331138	339264	95624
LaS	838833	72395	331138	339264	96036
Lu	27254	4141	9803	11304	2006
My	22474	5383	7333	9742	16

Os	2418	266	949	899	304
PhA	1698	260	647	692	99
PhT	1834	276	714	724	120
RtB	1719	188	631	748	152
RtT	1572	158	585	689	140
RtV	1754	202	639	754	159
Sl	15428	1618	5056	7130	1624
Um	20657	1920	7247	8602	2888
VeB	6119	1047	1958	2271	843
VeT	6083	974	1941	2133	1035
VeV	4855	763	1516	2175	401
Vz	234563	13532	70448	129492	21091

Table 4: Number of observed sound triplets

Language	aaa	vvv	vvc	vcv	vcc	cvv	ccv	cvc	ccc
Bq	101577	648	9521	33977	6749	10113	6718	33808	43
Cs	278990	10409	24046	64426	24004	31044	44098	69100	11863
Es	61485	1200	5740	14230	7773	6178	6896	19277	191
EtB	18445	596	1405	4107	2089	1661	2136	5933	518
EtT	18445	165	763	3776	2629	1048	2532	6804	728
Fi	314298	5587	35714	63899	42014	37503	41758	86835	988
Gr	71502	2202	6381	18692	6387	8507	7505	21522	306
Hi	9025	494	802	1609	1302	731	1303	2767	17
LaC	667718	9833	50093	155930	76745	57043	77946	232658	7470
LaS	658383	7196	45020	155410	77265	51059	78362	236605	7466
Lu	21942	1752	1974	6092	1999	1786	2005	6333	1
My	18618	2620	1166	7320	12	2331	16	5153	0
Os	1841	47	160	356	223	174	192	659	30
PhA	1172	64	99	348	71	130	71	382	7
PhT	1459	84	124	454	98	148	99	446	6
RtB	1394	35	106	448	107	123	118	443	14
RtT	1265	26	89	414	101	100	109	413	13
RtV	1440	42	112	460	109	125	123	449	20
Sl	11301	309	820	3218	851	972	1402	3617	112
Um	16288	695	738	3665	2161	853	2349	5495	332
VeB	4843	329	413	1128	428	521	489	1305	230
VeT	4965	318	343	1176	519	569	598	1316	126
VeV	2993	165	146	726	214	340	315	1063	24
Vz	153903	199	6102	41950	15572	13093	19714	55898	1375

In Table 4 can be seen that among the sound triplets the situation is still worse, i.e. the number of some triplet groups e.g. (vvv), (ccv), and especially (ccc), is quite low in several languages.

In Table 5 is presented the ratio of the number of all observed sounds, sound pairs and triplets to the theoretically possible number of different sounds, sound pairs and triplets. Table 5 indicates that whereas the results using particular sounds may be valid, the results using sound pairs may not be valid among the languages Os, Ph, Rt and Ve. The results using sound triplets may not be valid among the majority of tested languages, except La, Fi, Cs and possibly Vz, Bq, Gr, and Es.

Table 5: Observed number to possible number ratio, sorted

Sounds / 24		Pairs / 576		Triplets / 13824	
LaC	42888	LaC	1473	LaC	48.30
LaS	42499	LaS	1456	LaS	47.63
Cs	19097	Fi	663	Fi	22.74
Fi	18711	Cs	629	Cs	20.18
Vz	13366	Vz	407	Vz	11.13
Bq	6674	Bq	227	Bq	7.35
Gr	4880	Gr	162	Gr	5.17
Es	3781	Es	132	Es	4.45
Lu	1359	Lu	47	Lu	1.59
EtB	1268	EtB	42	My	1.35
EtT	1268	EtT	42	EtB	1.33
My	1097	My	39	EtT	1.33
Um	1044	Um	36	Um	1.18
Sl	826	Sl	27	Sl	0.82
Hi	583	Hi	20	Hi	0.65
VeB	319	VeT	11	VeB	0.36
VeT	309	VeB	11	VeT	0.35
VeV	296	VeV	8	VeV	0.22
Os	127	Os	4	Os	0.13
PhA	95	PhT	3	PhT	0.11
PhT	93	RtV	3	RtV	0.10
RtB	88	RtB	3	RtB	0.10
RtV	87	PhA	3	RtT	0.09
RtT	81	RtT	3	PhA	0.08

Number of different sounds, sound pairs and sound triplets

The number of different sounds, sound pairs and triplets in the database is presented in Tables 6-8.

Table 6: How many different sounds are observed in the database, sorted

Sounds Possible	all 24		(v) 5		(c) 19
Language		**Language**		**Language**	
Sl	24	Bq	5	***Sl***	19
Cs	23	*Cs*	5	*Cs*	18
VeB	23	Es	5	VeB	18
VeV	23	EtB	5	VeV	18
Vz	23	EtT	5	Vz	18
EtB	22	Fi	5	EtB	17
EtT	22	Gr	5	EtT	17
RtV	22	Hi	5	RtV	17
Um	22	LaC	5	Um	17
VeT	22	LaS	5	VeT	17
Bq	21	**My**	5	Bq	16
Os	21	Os	5	Os	16
RtB	21	PhA	5	RtB	16
LaS	20	PhT	5	LaS	15
RtT	20	RtB	5	RtT	15
Es	19	RtT	5	Es	14
Gr	19	RtV	5	Gr	14
Hi	19	***Sl***	5	Hi	14
LaC	19	Um	5	LaC	14
PhA	19	VeB	5	PhA	14
Fi	18	VeT	5	Fi	13
PhT	18	VeV	5	Lu	13
Lu	17	Vz	5	PhT	13
My	16	Lu	4	**My**	11

The highest number of consonants is observed in Sl, whereas almost one half less in My.

Table 7: Number of different sound pairs in the languages in the database, sorted

Pairs Max. possible	(all) 576		(vv) 25		(vc) 95		(cv) 95		(cc) 361
Language		**Language**		**Language**		**Language**		**Language**	
Cs	461	Bq	25	***Sl***	94	***Sl***	94	*Cs*	256
EtB	358	*Cs*	25	*Cs*	90	*Cs*	90	EtB	173
Sl	344	EtB	25	Vz	85	Vz	87	EtT	167
VeT	322	Fi	25	EtB	82	VeT	83	VeT	137
EtT	312	Gr	25	VeB	78	VeB	80	***Sl***	133
VeB	309	**My**	25	VeT	78	VeV	79	LaS	130
LaS	300	VeB	24	LaS	76	EtB	78	VeB	127
LaC	279	Vz	24	Bq	74	Bq	77	LaC	118
Vz	271	Es	23	Um	72	Um	76	Es	103
Es	262	LaC	23	Es	70	LaS	72	Um	90
Um	260	LaS	23	Gr	70	Gr	70	Gr	89
Gr	254	VeT	23	VeV	70	LaC	69	VeV	79

Bq	252	***Sl***	22	LaC	69	Os	68	Bq	76
VeV	249	Um	22	EtT	64	Es	66	RtV	76
Fi	213	PhA	21	Os	62	EtT	65	Vz	75
RtV	212	VeV	21	RtB	61	PhA	63	Hi	70
RtB	200	PhT	20	PhA	59	Fi	62	Fi	69
Hi	198	Hi	17	PhT	59	RtV	61	RtB	68
Os	194	Os	17	RtV	59	PhT	60	RtT	66
RtT	194	EtT	16	RtT	58	RtB	58	Lu	62
PhT	191	RtV	16	Fi	57	RtT	56	PhT	52
PhA	189	Lu	14	Hi	56	Hi	55	Os	47
Lu	164	RtT	14	**My**	47	**My**	45	PhA	46
My	122	RtB	13	Lu	45	Lu	43	**My**	6

In the Mycenaenan database is observed the by far lowest number of consonant-consonant pairs.

Table 8: Number of different sound triplets in the languages in the database, sorted
Abb.: trpl.: Triplets; poss.: Maximum possible; Lg.: Language

trpl. poss.	**(all) 13824**		**(vvv) 125**	.	**(vvc) 475**		**(vcv) 475**		**(vcc) 1805**	.	**(cvv) 475**	.	**(ccv) 1805**	.	**(cvc) 1805**	.	**(ccc) 6859**
Lg		**Lg.**		**Lg**		**Lg.**		**Lg.**		**Lg**		**Lg**		**Lg**		**Lg**	
Cs	3654	**My**	88	Fi	238	*Cs*	408	*Cs*	575	LaS	263	*Cs*	718	*Cs*	1012	*Cs*	501
LaS	2746	Fi	80	LaS	208	Vz	356	LaS	390	Fi	246	LaS	454	LaS	895	EtT	259
LaC	2445	EtB	76	Es	206	LaS	350	EtB	388	LaC	243	LaC	444	Vz	887	EtB	222
EtB	2438	LaS	73	EtB	206	Gr	330	EtT	383	Gr	241	EtB	382	LaC	719	LaS	113
EtT	2179	LaC	69	LaC	187	LaC	326	LaC	350	EtB	224	EtT	367	Gr	690	LaC	107
Gr	2177	Gr	65	Gr	179	***Sl***	326	Gr	307	*Cs*	221	Gr	333	Es	664	VeB	96
Vz	2087	***Sl***	48	*Cs*	177	Bq	305	Es	272	Vz	209	Fi	260	EtB	655	VeT	78
Es	1952	Es	44	Bq	152	EtB	285	Fi	262	Bq	197	Es	258	EtT	627	Es	60
Fi	1944	*Cs*	42	**My**	141	Es	279	Vz	204	Es	169	***Sl***	258	Bq	624	***Sl***	46
Sl	1700	Um	41	EtT	129	Fi	272	VeT	198	**My**	161	Vz	258	***Sl***	578	Um	46
Bq	1693	Bq	40	Vz	125	Um	241	***Sl***	195	EtT	152	VeT	203	Fi	570	Gr	32
Um	1322	VeB	35	***Sl***	121	EtT	229	Bq	175	***Sl***	128	Um	194	Um	450	Vz	24
VeT	1289	EtT	33	VeB	103	VeB	218	Um	168	Um	107	Bq	181	VeB	372	Bq	19
VeB	1277	VeT	31	VeT	102	VeT	218	VeB	168	VeB	104	VeB	181	VeT	371	Fi	16
My	969	Lu	29	Um	75	**My**	215	Hi	123	Hi	97	Hi	138	**My**	351	RtV	16
Hi	955	PhT	28	Hi	69	PhT	157	Lu	120	VeT	97	Lu	122	Hi	339	VeV	15
Lu	830	Hi	26	RtV	57	Hi	154	RtV	78	Lu	90	VeV	104	Lu	295	Os	14
VeV	728	PhA	26	Lu	55	RtV	147	RtB	76	VeV	79	RtV	89	VeV	256	RtB	11
RtV	704	Vz	24	RtB	54	VeV	147	RtT	71	PhT	66	RtB	81	RtB	237	Hi	9
RtB	680	VeV	23	PhT	51	RtB	146	VeV	66	PhA	65	Os	79	RtV	237	RtT	9
PhT	658	RtV	17	RtT	48	PhA	143	PhT	65	RtV	63	RtT	76	RtT	231	PhA	7
RtT	643	RtB	15	Os	46	RtT	140	Os	63	RtB	60	PhT	72	PhT	213	PhT	6
PhA	587	Os	12	PhA	45	Lu	118	PhA	48	Os	56	PhA	54	Os	199	Lu	1
Os	586	RtT	12	VeV	38	Os	117	**My**	7	RtT	56	**My**	6	PhA	199	**My**	0

Mycenaenan is characterized by the far the lowest number of different triplets of the (vcc), (ccv), and (ccc) type.

Ratio of sounds pairs and triplets to all possible ones

Ratio of the number of diferent sounds pairs and triplets are presented in tables 9 and 10.

Table 9: Ratio of the number of observed different sound pairs to all possible ones

	all		**(vv)**		**(vc)**		**(cv)**		**(cc)**
Cs	0.800	Bq	1	***Sl***	0.989	***Sl***	0.989	*Cs*	0.444
EtB	0.622	*Cs*	1	*Cs*	0.947	*Cs*	0.947	EtB	0.300
Sl	0.597	EtB	1	Vz	0.895	Vz	0.916	EtT	0.290
VeT	0.559	Fi	1	EtB	0.863	VeT	0.874	VeT	0.238
EtT	0.542	Gr	1	VeB	0.821	VeB	0.842	***Sl***	0.231
VeB	0.536	**My**	1	VeT	0.821	VeV	0.832	LaS	0.226
LaS	0.521	VeB	0.960	LaS	0.800	EtB	0.821	VeB	0.220
LaC	0.484	Vz	0.960	Bq	0.779	Bq	0.811	LaC	0.205
Vz	0.470	Es	0.920	Um	0.758	Um	0.800	Es	0.179
Es	0.455	LaC	0.920	Es	0.737	LaS	0.758	Um	0.156
Um	0.451	LaS	0.920	Gr	0.737	Gr	0.737	Gr	0.155
Gr	0.441	VeT	0.920	VeV	0.737	LaC	0.726	VeV	0.137
Bq	0.438	***Sl***	0.880	LaC	0.726	Os	0.716	Bq	0.132
VeV	0.432	Um	0.880	EtT	0.674	Es	0.695	RtV	0.132
Fi	0.370	PhA	0.840	Os	0.653	EtT	0.684	Vz	0.130
RtV	0.368	VeV	0.840	RtB	0.642	PhA	0.663	Hi	0.122
RtB	0.347	PhT	0.800	PhA	0.621	Fi	0.653	Fi	0.120
Hi	0.344	Hi	0.680	PhT	0.621	RtV	0.642	RtB	0.118
Os	0.337	Os	0.680	RtV	0.621	PhT	0.632	RtT	0.115
RtT	0.337	EtT	0.640	RtT	0.611	RtB	0.611	Lu	0.108
PhT	0.332	RtV	0.640	Fi	0.600	RtT	0.589	PhT	0.090
PhA	0.328	Lu	0.560	Hi	0.589	Hi	0.579	Os	0.082
Lu	0.285	RtT	0.560	**My**	0.495	**My**	0.474	PhA	0.080
My	0.212	RtB	0.520	Lu	0.474	Lu	0.453	**My**	0.010

More than half of the possible number of different sound pairs have Old Church Slavonic, Old Slovene, Etruscan and Venetic. The highest are the ratios at the sound pairs of the type (vv), followed by (vc) and (cv). The lowest are those at (cc).

Table 10: Ratio of the number of observed different sound triplets to all possible ones

	all		**(vvv)**		**(vvc)**		**(vcv)**		**(vcc)**		**(cvv)**		**(ccv)**		**(cvc)**		**(ccc)**
Cs	0.264	**My**	0.704	Fi	0.501	*Cs*	0.859	*Cs*	0.319	LaS	0.554	*Cs*	0.398	*Cs*	0.561	*Cs*	0.073
LaS	0.199	Fi	0.640	LaS	0.438	Vz	0.749	LaS	0.216	Fi	0.518	LaS	0.252	LaS	0.496	EtT	0.038
LaC	0.177	EtB	0.608	Es	0.434	LaS	0.737	EtB	0.215	LaC	0.512	LaC	0.246	Vz	0.491	EtB	0.032
EtB	0.176	LaS	0.584	EtB	0.434	Gr	0.695	EtT	0.212	Gr	0.507	EtB	0.212	LaC	0.398	LaS	0.016
EtT	0.158	LaC	0.552	LaC	0.394	LaC	0.686	LaC	0.194	EtB	0.472	EtT	0.203	Gr	0.382	LaC	0.016
Gr	0.157	Gr	0.520	Gr	0.377	***Sl***	0.686	Gr	0.170	*Cs*	0.465	Gr	0.184	Es	0.368	VeB	0.014
Vz	0.151	***Sl***	0.384	*Cs*	0.373	Bq	0.642	Es	0.151	Vz	0.440	Fi	0.144	EtB	0.363	VeT	0.011
Es	0.141	Es	0.352	Bq	0.320	EtB	0.600	Fi	0.145	Bq	0.415	Es	0.143	EtT	0.347	Es	0.009
Fi	0.141	*Cs*	0.336	**My**	0.297	Es	0.587	Vz	0.113	Es	0.356	***Sl***	0.143	Bq	0.346	***Sl***	0.007
Sl	0.123	Um	0.328	EtT	0.272	Fi	0.573	VeT	0.110	**My**	0.339	Vz	0.143	***Sl***	0.320	Um	0.007
Bq	0.122	Bq	0.320	Vz	0.263	Um	0.507	***Sl***	0.108	EtT	0.320	VeT	0.112	Fi	0.316	Gr	0.005
Um	0.096	VeB	0.280	***Sl***	0.255	EtT	0.482	Bq	0.097	***Sl***	0.269	Um	0.107	Um	0.249	Vz	0.003
VeT	0.093	EtT	0.264	VeB	0.217	VeB	0.459	Um	0.093	Um	0.225	Bq	0.100	VeB	0.206	Bq	0.003
VeB	0.092	VeT	0.248	VeT	0.215	VeT	0.459	VeB	0.093	VeB	0.219	VeB	0.100	VeT	0.206	Fi	0.002
My	0.070	Lu	0.232	Um	0.158	**My**	0.453	Hi	0.068	Hi	0.204	Hi	0.076	**My**	0.194	RtV	0.002
Hi	0.069	PhT	0.224	Hi	0.145	PhT	0.331	Lu	0.066	VeT	0.204	Lu	0.067	Hi	0.188	VeV	0.002
Lu	0.060	Hi	0.208	RtV	0.120	Hi	0.324	RtV	0.043	Lu	0.189	VeV	0.058	Lu	0.163	Os	0.002
VeV	0.053	PhA	0.208	Lu	0.116	RtV	0.309	RtB	0.042	VeV	0.166	RtV	0.049	VeV	0.142	RtB	0.002
RtV	0.051	Vz	0.192	RtB	0.114	VeV	0.309	RtT	0.039	PhT	0.139	RtB	0.045	RtB	0.131	Hi	0.001
RtB	0.049	VeV	0.184	PhT	0.107	RtB	0.307	VeV	0.037	PhA	0.137	Os	0.044	RtV	0.131	RtT	0.001
PhT	0.048	RtV	0.136	RtT	0.101	PhA	0.301	PhT	0.036	RtV	0.133	RtT	0.042	RtT	0.128	PhA	0.001
RtT	0.047	RtB	0.120	Os	0.097	RtT	0.295	Os	0.035	RtB	0.126	PhT	0.040	PhT	0.118	PhT	0.001
PhA	0.042	Os	0.096	PhA	0.095	Lu	0.248	PhA	0.027	Os	0.118	PhA	0.030	Os	0.110	Lu	0.000
Os	0.042	RtT	0.096	VeV	0.080	Os	0.246	**My**	0.004	RtT	0.118	**My**	0.003	PhA	0.110	**My**	0.000

Among sound triplets, the highest share of all of them have Old Church Slavonic, Latin, Etruscan, Greek and Venezian. The highest ratio is among the sound triplets of the (vcv) type.

Dependence on the size of database

Sound triplets

The relation between the size of the databases and the number of observed different sound groups was expected to be expressed in the present study the most in the case of sound triplets. Therefore these results are presented in Figures 1-5.

From Figures 1 to 5 follows that there is a nonlinear relation between the database size expressed as the number of all sound triplets, and the number of different sound triplets. Figure 2 resp. 5 present that that the log(y),log(x) plot resp. the 1/y,1/x plot indicate no other dependence on the database size, but appreciable spread of data due to differences in languages. This is supported by Figure 3 resp. 4 presenting the log,linear and power,linear dependence. Evident is also that Mycenaean and Luvian are outliers in this respect.

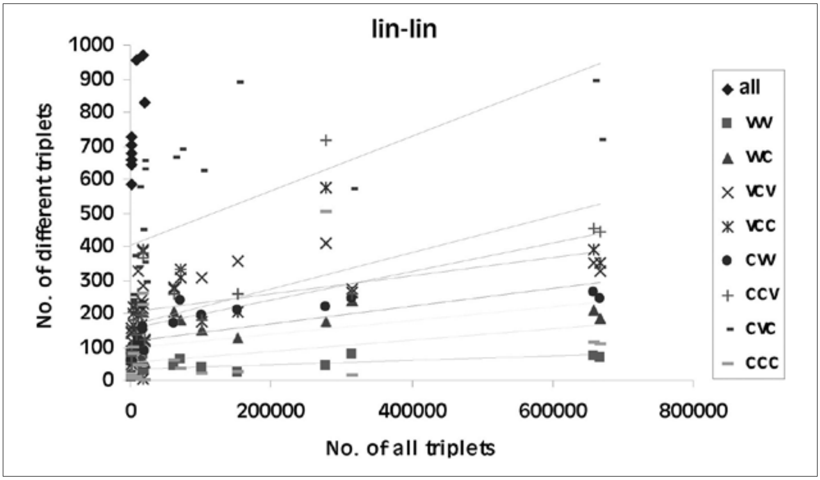

Figure 1: The linear-linear, lin(y,x), dependence between the size of the databases and the number of observed different sound triplets. The regression line for all triplets is above y = 1000

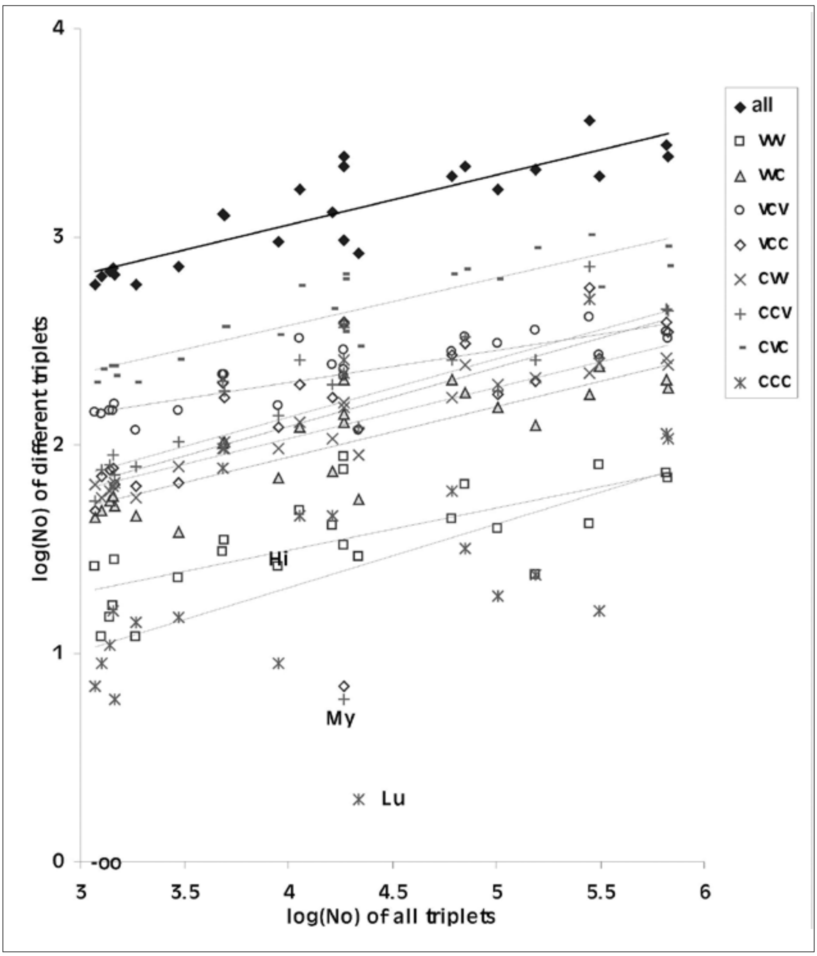

Figure 2: The log-log (log(y), log(x)) dependence between the size of the databases and the number of observed different sound triplets. Here it is the most obvious that Luvian and Mycenaean are outliers

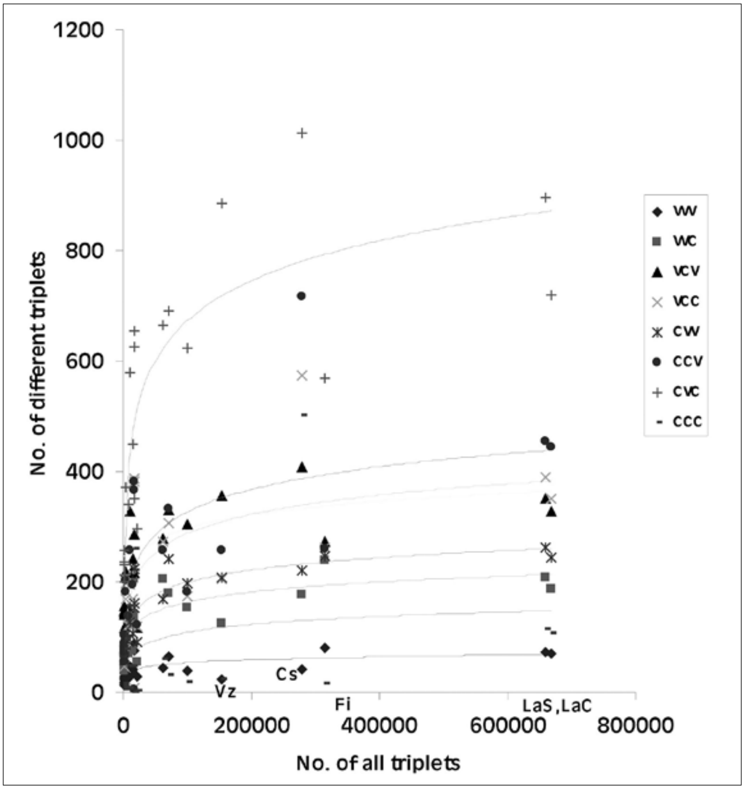

Figure 3: The log-linear dependence between the size of the databases and the number of observed different sound triplets. The "all" data are omitted for better visibility of other ones

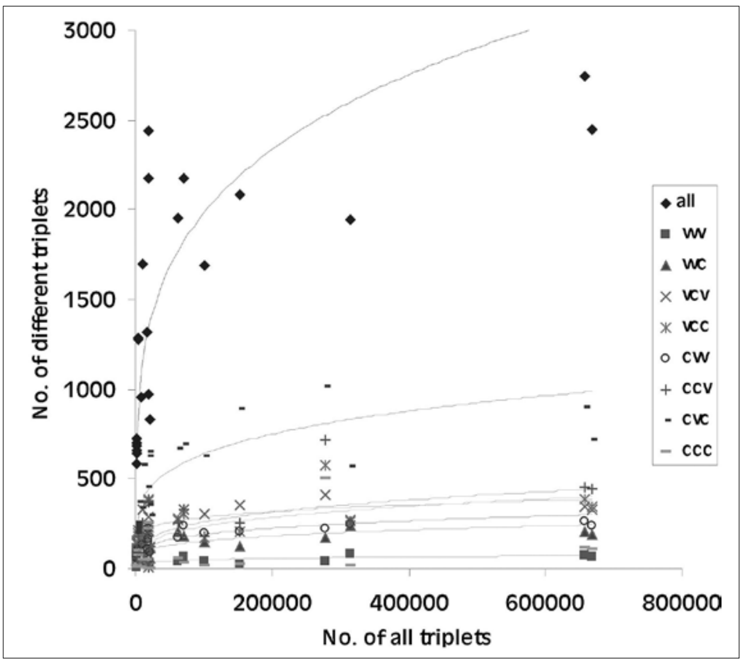

Figure 4: The power-linear (y = xn) dependence between the size of the databases and the number of observed different sound triplets

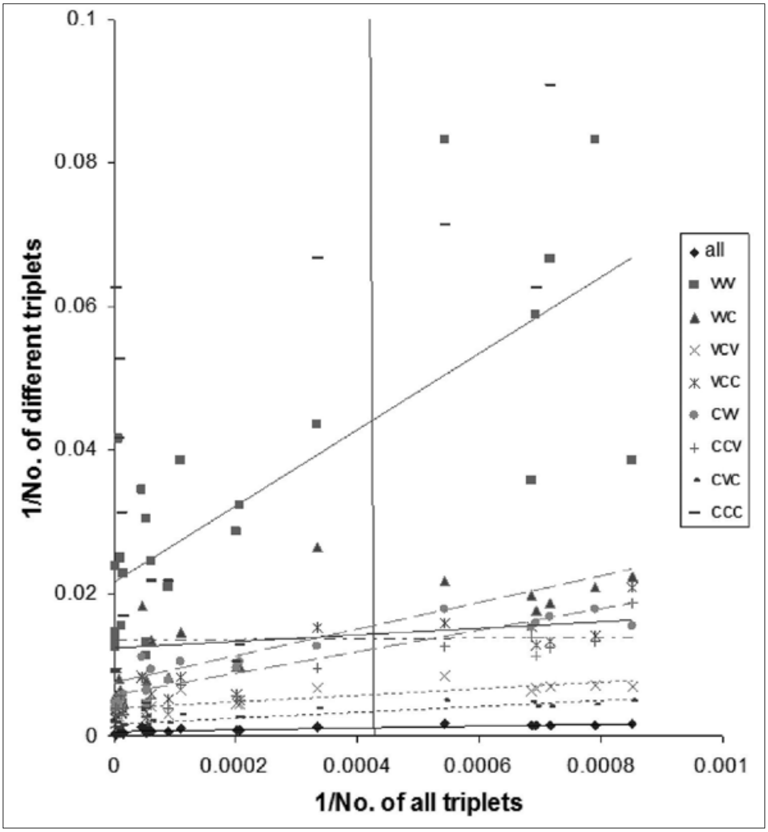

Figure 5: The Lineweaver-Burk plot of the dependence between the size of the databases and the number of observed different sound triplets. Also here i t is obvious that Luvian and Mycenaean are outliers

Table 11: Correlation (R) between the size of the databases and the number of observed different sound triplets taking into account data of all languages

triplets	lin(y,x)	log(x), log(y)	y = log(x)	y = x^n	1/y, 1/x
all	0.637	0.869	0.842	0.869	0.880
ccc	0.278	0.357	0.384	**np**	0.126
ccv	0.632	0.589	0.746	0.589	**0.003**
cvc	0.634	0.896	0.884	0.896	0.896
cvv	0.707	0.931	0.932	0.931	0.914
vcc	0.576	0.610	0.736	0.610	0.049
vcv	0.593	0.820	0.838	0.820	0.723
vvc	0.615	0.846	0.847	0.846	0.800
vvv	0.525	0.721	0.675	0.721	0.766

np – not possible

Table 12: Correlation (R) between the size of the databases and the number of observed different sound triplets taking into account data of all languages except the outliers Mycenaean and Luvian

triplets	lin(y,x)	log(x), log(y)	y = log(x)	$y = x^n$	1/y, 1/x
all	0.629	0.895	0.872	0.895	0.935
ccc	0.260	0.549	0.396	0.549	0.767
ccv	0.629	0.859	0.792	0.859	**0.951**
cvc	0.626	0.915	0.909	0.915	0.938
cvv	0.714	0.945	0.948	0.945	0.934
vcc	0.571	0.840	0.782	0.840	0.918
vcv	0.593	0.881	0.884	0.881	0.857
vvc	0.624	0.876	0.873	0.876	0.846
vvv	0.631	0.766	0.755	0.766	0.770

For this reason, the correlation coefficients of relationships observed in Figures 1 to 5 were obtained. They are presented in Table 11 to 14. Let us have first a look to the situation at the sound triplets, Tables 11 and 12.

Data in Table 11 and 12 present clearly that by omitting the data of Mycenaean and Luvian the correlations get improved, in two cases drastically. For this reason are given in following Tables correlation coefficients obtained without data of Mycenaean and Luvian.

The best correlation coefficients are observed using the 1/y,1/x plot, i.e. the Lineweaver-Burk form of the Michaelis-Menten equation $y = Y_{max}*x/(K+x)$, often used in biochemistry [5]. Next best ones are obseved at the log(x),log(y) ≈ $y = x^n$ function. Besides the better correlation coefficients, the Michaelis-Menten equation has another priority over the second best functions. It is namely a hyperbolic function having an upper limit. Also the maximum possible number of sound triplets in a database has a theoretical upper limit and this upper limit is far from being reached by actual data, cf. Table 10. Thusly, the Michaelis-Menten equation is to be considered the most appropriate one in present situation: R in 1/y,1/x > log(x),log(y) ≥ $y = x^n$ > y = log(x) >> lin(y,x). Regardless the function used, the correlation between the size of the databases and the number of observed different sound triplets is the highest among the triplets of the (ccv) and (cvc) type, whereas it is the lowest among the triplets of the (ccc) and (vvv) type.

Now let us look at the situation among the sound pairs and single sounds. The situation among the sound pairs is presented in Table 13.

Table 13: Correlation (R) between the size of the databases and the number of observed different sound pairs taking into account data of all languages except the outliers Mycenaean and Luvian

pairs	lin(y,x)	log(x), log(y)	y = log(x)	$y = x^n$	1/y, 1/x
all	0.261	0.517	0.481	0.517	0.717
vv	0.378	0.657	0.674	0.657	0.707
vc	0.177	0.441	0.428	0.441	0.605
cv	0.101	0.372	0.354	0.372	0.575
cc	0.268	0.483	0.432	0.483	0.675

Among the sound pairs, the correlation between the size of the databases and the number of observed different sound pairs is much lower than among the sound triplets. This indicates that the number of observed different pairs is not that dependent on the size of the database as in the case of the sound triplets. This means that the size of the database doesn't influence appreciably the results of the sound pairs and that the main contribution have the differences in sound pair frequency between the languages.

The situation among single sounds is presented in Table 14. Here, only the data using the Lineweaver-Burk form of the Michaelis-Menten equation are presented, since other correlation coefficients are still lower.

Table 14: Correlation (R) between the size of the databases and the number of observed different sounds taking into account data of all languages except the outliers Mycenaean and Luvian

single	R
all	0.149
v	0.000
c	0.144

The correlation coefficients are in this case very low, indicating that the size of the database in the case of single sounds has little if any influence on the results, as well as that the almost only contribution have the differences in sound frequency between the languages.

The situation using the Michaelis-Menten plot is illustrated in Figures 6-8. They are presented in two versions. Above is the situation where all languages are included. Below is the enlarged left hand part of it to see better the situation among languages for which smaller databases are available.

Sound triplets

In Figure 6 can be seen that the Michaelis Menten function is better than the log,log function not only due to higher correlation coefficients but also by its shape indicating an upper limit of the possible number of different sound triplets. The spread of the numbers of different sound triplets characteristic for the languages in question is clearly seen to be superimposed to their dependence on the size of the database. Thus, among a number of tested languages, especially those ancient languages for which too small databases could be prepared, the size of known texts is too small for a serious comparison based on the frequency of sound triplets observed in them. In Figure 6 we can see also that if we take the obtained Michaelis Menten function as an average of all data, then the languages placed below and to the right of the Michaelis Menten regression line have a subaverage number of different sound triplets. These languages are, e.g., Basque, Umbrian, Mycenaean, Luvian, Hittite, Oscan. The languages placed above and to the left of the Michaelis Menten regression line have an over-average number of different sound triplets. These languages are, e.g., Latin, Old Church Slavonic, Venezian, Greek, Etruscan, Old Slovene. However, as presented in Table 9 and 10, these values are, with few exceptions, well below the theoretically possible ones.

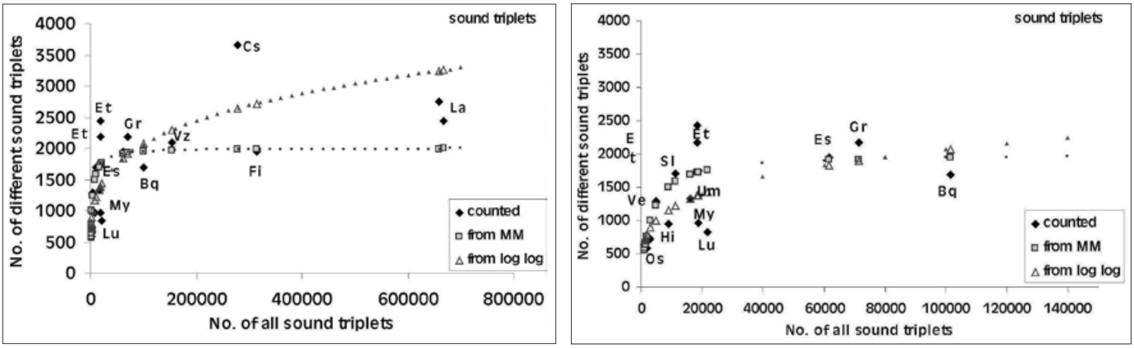

Figure 6: Comparison of data of the dependence between the size of the databases and the number of observed different sound triplets and those reconstructed using the Michaelis-Menten, MM, resp. the log,log function

Sound pairs

In Figure 7 can be seen that the spread of the numbers of different sound pairs is superimposed to their dependance on the size of the database. However, the dependence on the size of the database is in the case of sound pairs not as expressed as in the case of sound triplets. In spite of that, among a number of tested languages, especially those ancient languages for which too small databases could be prepared, the size of known texts is so small that a serious comparison based on the frequency of sound pairs observed in them is questionable.

In Figure 7 we can see that in the case of sound pairs, the languages placed below and to the right of the Michaelis Menten regression line having an subaverage number different of sound pairs are, e.g., Finnic, Greek, Basque, Estonian, Umbrian, Mycenaean, Luvian, Hittite, Oscan. The Latin and Venezian language are placed close to the regression line. The languages placed above and to the left of the Michaelis Menten regression line have an over-average number of different sound triplets. These languages are, e.g., Old Church Slavonic, Etruscan, Old Slovene, Venetic.

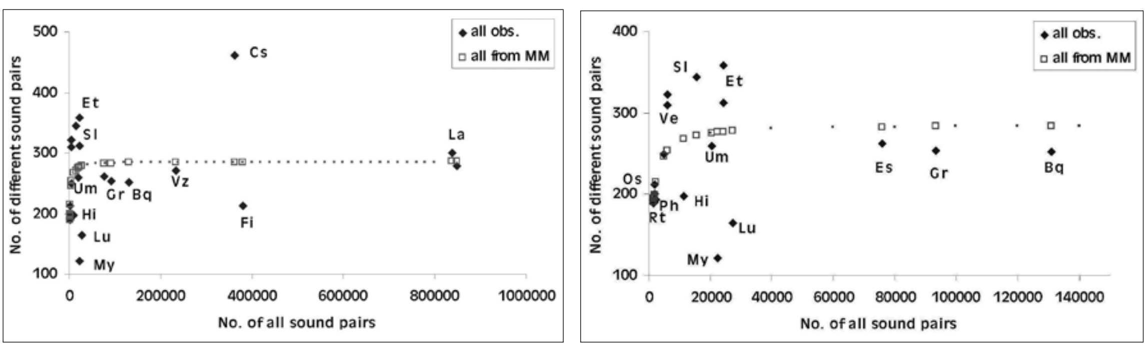

Figure 7: Comparison of data of the dependence between the size of the databases and the number of observed different sound pairs and those reconstructed using the Michaelis-Menten function

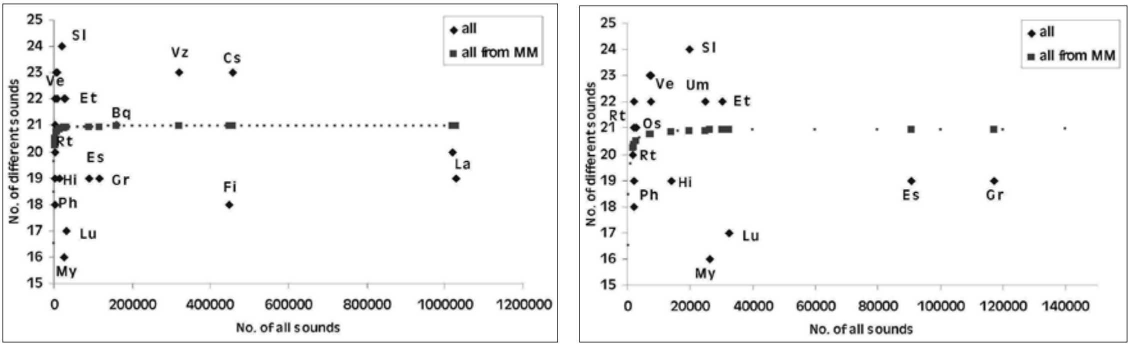

Figure 8: Comparison of data of the dependence between the size of the databases and the number of observed different sounds and those reconstructed using the Michaelis-Menten function

Single sounds

In Figure 8 can be seen that the spread of the numbers of different sounds is hardly dependent on the size of the database.

Discussion

Vowel-to-consonant ratio

The vowel-to-consonant ratio in tested languages is as follows [3]:

1.70 = My >> 1.20 > Lu > PhT > PhA > 1.10 > VeV > RtV > RtT > RtB > Gr > Bq > 1.00 > Sl > VeB > Fi > Vz > Hi > Cs > Es > VeT > Um > 0.90 > LaC > EtB > LaS > Os > 0.80 > EtT = 0.74

Obvious outliers are Mycenaean, where the vowels seem to prevail by far, followed by Luvian, whereas in Etruscan as read by the mainstream linguists the consonants prevail more than in any other tested language. Interestingly, by Bor's [6–7] way of reading, Etruscan falls between the two reading variants of Latin, thus it normalizes its position in present respect.

Importance of the size of the database

From the vowel-to-consonant ratio, data in Table 11 and 12, as well as Figure 2 and 5 can be concluded that Mycenaean and Luvian are outliers, drastically influencing some of the tested functions.

Tables 11-13 indicate that the Lineweaver-Burk form of the Michaelis-Menten equation gives the best correlations between the size of the database expressed as the number of all observed sound (singlets, pairs, triplets) and the number of different sound (singlets, pairs, triplets). Since this is a hyperbolic function, it is also theoretically the most appropriate one, since the number of different sounds and their combinations has an upper limit. Next to it, the log,log function and the power,linear (in fact root,linear) give good correlation.

All of them indicate that the number of observed different sounds and their combinations is a nonlinear function of the size of the database.

In Figure 8 can be seen that the number of observed different sounds reconstructed from the Michaelis Menten function falls close to or on the upper limit of this function derived from observed data. Only Luvian and Mycenaean deviate more than the others. From this observation follows that the spread of data in the case of single sounds is not the function of the size of the database but only of the differences between the languages. Thus, language distances based on frequencies of single sounds used in previous [1-4] and present work are credible.

That is reflected also in the values of the Michaelis-Menten constant K, Table 15, from which we can conclude that taking x = 10*K as a limit beyond which the size of the database has less than 10 % probability to influence the results, is appropriate. Among the single sounds a database containing over 700 signs would be of a sufficient size by this criterion. For the sound pairs to be taken into account, a database should contain more than 8000 sound pairs. Among the triplets, such a limit would be over 30.000 sound triplets. If x = 20*K would be taken as a criterion predicting less than 5 % probability that the size of the database would influence the results, then the respective values would be 1380, 15780, and 60500, respectively

Table 15: The values of the Michaelis-Menten constant K for the dependence of the number of different sound combinations on the number of all observed sound combinations as well as the necessary number of all sound combinations in the database in order that the influence of the size of the database is less than the given percentage

all observed different	K	probability of influence	
		<10 %	<5 %
single sounds	69	690	1380
sound pairs	789	7890	15780
sound triplets	3025	30250	60500

If we look now at the Table 1, we can see that all tested language databases exceed the 700 sounds limit as well as the 1380 sounds limit for single sounds. Thus the results obtained from the single sound frequencies [1-4] are valid. For the sound pairs the situation is different. The databases of languages Os, Ph, Rt, and Ve are smaller than the 8000 pairs limit and in addition, the databases of languages Hi and Sl are smaller than the 15780 pairs limit. The results obtained for these languages are thus questionable, at least. Still worse is the situation among the sound triplets. In this case, the databases of languages Et, Hi, Lu, My, Os, Ph, Rt, Sl, Um, and Ve are smaller than 30.000 triplets.

Thusly, exactly for the ancient languages Et, Ph, Rt, and Ve, for which the studies [1-4] were started, only the data on frequency of single sounds are useful, but not the data on sound pairs and triplets. For this reason, of the Tables 4-6 in ref [4] only the part presented in Table 16 is applicable.

Table 16: Applicable part of the language distances derived from the sound frequencies in [4]

Sounds	**Method**	**Etruscan**	**Old Phrygian**	**Rhaetic**	**Venetic**
Single	PCA	**Et**<Rt<Sl<La<Gr	**Ph**<Sl<La<Gr	**Rt**<Et<Sl<La<Gr	**Ve**<Cs<Gr<Sl<La
	F,R,STE	**Et**<Rt<Sl<La<Gr	**Ph**<Sl<Gr<La	**Rt**<Sl<Et <La<Gr	**Ve**<Cs<Sl<Gr<La
	SuD(S)	**Et**<Rt<Sl<La<Gr	**Ph**<Sl<Gr<La	**Rt**<Sl<Et<La<Gr	**Ve**<Cs<Sl<Gr<La

Sounds	**Method**	**Etruscan**	**Rhaetic**
Single	PCA	**EtT**<RtT<LaC<Sl<Gr	**RtT**<EtT<LaC<Sl<Gr
	F, R	**EtT**<RtT<Sl<LaC<Gr	**RtT**<EtT<Sl<LaC<Gr
	STE	**EtT**<RtT<Sl<Gr<LaC	**RtT**<EtT<Sl<LaC<Gr
	SuD	**EtT**<RtT<Sl<LaC<Gr	**RtT**<Sl<EtT<LaC<Gr
	SuS	**EtT**<RtT<Sl<LaC<Gr	**RtT**<EtT<Sl<LaC<Gr

Sounds	**Method**	**Old Phrygian**	**Venetic**
Single	PCA	**PhT**<Sl<LaC<Gr	**VeT**<Cs<Gr<Sl<LaC
	F, R	**PhT**<Sl<Gr<LaC	**VeT**<Gr<Cs<Sl<LaC
	STE	**PhT**<Sl<Gr<LaC	**VeT**<Cs<Gr<Sl<LaC
	SuD	**PhT**<Sl<LaC<Gr	**VeT**<Gr<Sl<Cs<LaC
	SuS	**PhT**<Sl<Gr<LaC	**VeT**<Cs<Sl<Gr<LaC

Sounds	**Method**	**Etruscan**
Pairs	R, STE	**Et**<Rt<Sl<La<Gr
	SuD(S)	**Et**<Rt<Sl<La<Gr
Pairs	R	**EtT**<RtT<Sl<LaC<Gr
	STE	**EtT**<RtT<Sl<LaC<Gr
	SuD	**EtT**<RtT<LaC<Sl<Gr
	SuS	**EtT**<RtT<Sl<LaC<Gr

Correlation between different languages

Correlation between tested languages is illustrated in Table 17 with two languages as examples; Latin read in the classical way as an example of a big database and Rhaetic read in the Bor's way as an example of a small database. Table 17 demonstrates that using the frequency of sound pairs and triplets, the selectivity of the method increases in this direction.

Table 17: Correlation of tested languages based on sound frequencies to Latin and Rhaetic

Single	R> 0.90	0.90>R>0.80	0.80>R>0.50	R>0.50
LaC	LaS>Os>Sl>Um	Fi>Gr>Es>My>RtT>RtV>Bq>RtB>PhT>PhA>VeB>VeT>VeV	Vz>EtB>Cs>EtT>Hi>Lu	
RtB	RtT>RtV	Sl>EtT>Fi>Os>Es>EtB>LaC>PhA>PhT>LaS>Lu	Bq>VeV>Hi>VeB>Cs>Um>VeT>My>Gr>Vz	

Pairs	R> 0.90	0.90>R>0.80	0.80>R>0.50	R>0.50
LaC	LaS		Os>Um>Bq>RtT>RtB>Sl>Es>RtV>Vz>My>Gr>Fi>PhA>PhT>EtT>VeT>VeV>VeB>EtB	Lu>Hi>Cs
RtB	RtT>RtV		EtT>Es>EtB>Sl>Fi>LaC>Lu>LaS>PhA>Os>PhT>Hi>Bq>Um>VeB>Cs>VeT>Vz	VeV>My>Gr

Triplets	R> 0.90	0.90>R>0.80	0.80>R>0.50	R>0.50
LaC	LaS			Vz>Gr>Um>Bq>Os>Es>RtV>My>RtB>RtT>EtT>Sl>Fi >PhT>PhA>EtB>Lu>VeT>VeB>Hi>VeV>Cs
RtB	RtT>RtV			EtT>EtB>Es>LaC>Sl>Lu>LaS>PhT>Fi>Um>PhA>VeB> Bq>My>Cs>VeT>Os>Vz>Hi>Gr>VeV

Thus, the use of sound pairs and triplets is advisable, if the available databases are sufficiently large.

Number of different sound pairs and triplets

Regardless of the function tested, the best correlation between the size of the database expressed as the number of all observed triplets, and the number of different sound triplets is observed among the triplets groups (cvv) and (cvc), whereas the worst is among (ccc) and (vvv). Among the latter ones as well as the rest of the other triplet subgroups the spread of data overwhelms the dependence on the size of the database.

The fact that Luvian and Mycenaean appeared as ouliers indicates that the decipherment of the signs with which they are written is not yet sufficiently solved. This is apparent already from the Younger's [8] table of Linear B signs, where there are ascribed to 5 signs the vowel (v) sound values, to 56 of them the sound values of the type (cv), to 2 of them of the type (vv), to 7 of them of the type (ccv), whereas at 8 signs the sound value is doubtful and at 9 signs it is unknown. In Table 18 are presented sound pairs of the type (cc) and sound triplets of the type (ccv) and (ccc) observed in Mycenaean. Table 19 presents them for Luvian.

Table 18: Some types of sound pairs resp. triplets observed in Mycenaean

(cc)	pt>tr>kr>ks>pr
(ccv)	pte>kri>kso>pri>tre>tri
(ccc)	(none)

Table 19: Some types of sound pairs resp. triplets observed in Luvian

(cc)	nz>nt>nd>st>rs>lh>sd>sh>rp>rn>sp>rh>rt>hr>rm>lz>mn>mp>rl>tn
(ccv)	nza>nzi>nta>nti>ndu>sta>nda>lha>rsa>sdu>sti>sha>spa>hra>rpa>rna>lza>mna>rta>rma
(ccc)	snz

From the above analysis of the situation among the other languages (Tables 6-10) it seems probable that in Linear B (and still more in Linear A) there are present additional signs having the sound value of the (ccv) and possibly even of (cc) or (ccc) type.

Since for the decipherment of the Linear B script there had been based on Latin and especially on Greek [9], then basing on the above analysis there should be allowed also for the possibility that some Slavic characteristics may be applicable, since Old Church Slavonic and Old Slovene have the highest numbers of different sound pairs of the type (cc) and sound triplets of the type (ccv) and (ccc). In Table 20 are presented twenty most frequent ones in these Slavic languages as an impetus for additional study.

Table 20: Twenty most frequent consonantal sound pairs resp. triplets in Old Church Slavonic and Old Slovene

(cc)	
Cs	st>pr>št>tv>sl>sv>vs>tr>sp>gl>žd>sk>dn>br>vr>vš>bl>mn>dr>vz
Sl	st>pr>sv>dn>br>sp>tr>rn>rt>vs>gr>pš>kr>bl>lž>št>zl>vr>sk>dv
(ccv)	
Cs	pri>šte>sti>tvo>sta>gla>šti>šta>pro>svo>sto>bla>žde>spo>svi>mno>bra>vsi>slo>sla
Sl	sta>stu>sti>bra>sve>dnu>spo>rni>pri>pše>gre>pra>pre>tri>svo>lža>rtu>vsa>tra>pro
(ccc)	
Cs	stv>str>mrt>smr>čst>vst>tvr>skr>drž>vsk>rtv>stn>rst>vzd>tgd>crk>prv>zdr>vzm>rkv
Sl	dvr>vrn>str>rst>lsk>rtr>stn>stv>štr>črn>lžn>mrt>rtv>slz>vst>bhr>brg>brn>drž>dst

Conclusions

Mycenaean and Luvian are obvious outliers in present study. In them vowels prevail more than in other tested languages. In Etruscan, as read by the mainstream linguists, the consonants prevail more than in any other tested language.

To obtain reliable results on studying the language distance based on sound frequency, the size of the database is important. Taking the frequency of single sounds as the basis for the approach, a database containing over 700 signs would be of a sufficient size. For the sound pairs to be taken into account, a database should contain more than 8000 sound pairs. For the sound triplets, such a limit would be over 30.000 sound triplets. Thus, in previous studies [1-4] only results based on the frequency of single sounds are reliable for the ancient languages like Etruscan, Old Phrygian, Rhaetic and Venetic.

The selectivity of the method, however, increases in the direction single sounds < sound pairs < sound triplets. For this reason, the use of sound pairs and triplets is advisable, if the available databases are sufficiently large.

In Luvian and especially in Mycenaean there seems that several additional sound pairs of the type consonant-consonant resp. sound triplets of the type consonant-consonant-vowel or even consonant-consonant-consonant should be taken into account. Since the Latin and Greek ones were already considered, the Slavic ones should be tested as well.

References

1. M. Silvestri, G. Tomezzoli, *Linguistic Computational Analysis to measure the distances between ancient Venetic, Latin and Slovenian Languages*, Proceedings of the Third International Topical Conference, Ancient Settlers of Europe, Založništvo Jutro, Ljubljana **2005**, 77-85, http://www.korenine.si/zborniki/zbornik05/tomezzoli_venslolat.htm.
2. M. Silvestri, G. Tomezzoli, *Linguistic distances between Rhaetian, Venetic, Latin and Slovenian languages*, Proceedings of the Fifth International Topical Conference, Origin of Europeans, Založništvo Jutro, Ljubljana **2007**, 184-190, http://www.korenine.si/zborniki/zbornik07/tomezzoli_dist07.pdf.
3. A. Perdih, G. Tomezzoli, V. Vodopivec, *Comparison of contemporary and ancient languages*. Zbornik šeste mednarodne konference Izvor Evropejcev (Proceedings of the Sixth International Topical Conference, Origin of Europeans), Jutro, Ljubljana **2008**, 40-87, http://www.korenine.si/zborniki/zbornik08/comparison_languages.pdf.

4. A. Perdih, *Comparison of some methods of estimation of linguistic distances*, Zbornik osme mednarodne konference Izvor Evropejcev, Jutro, Ljubljana **2010**, 78-86.
5. R. F. Boyer, *Concepts in biochemistry*, 3rd ed., J. Wiley & Sons, Hoboken **2006**.
6. M. Bor, J. Šavli, I. Tomažič, *Veneti naši davni predniki*, Editiones Veneti, Vienna: German Ed**. 1988**, Slovene Ed. **1989**, Italian Ed. **1991**, English Ed. (*Veneti. First Builders of European Community*) **1996**, Russian Ed. Part I **2002**.
7. M. Bor, *Etruščani in Veneti* (Tomažič I., Ed.), Editiones Veneti, Wien **1995**; Russian Ed.: M. Bor, I. Tomažič, *Veneti i Etruski*, Aleteiya, Sankt-Peterburg **2008**.
8. J. Younger, *Linear A & B Grids*, **2005-2010**, http://people.ku.edu/~jyounger/LinearA/ABgrids.html.
9. E. Doblhofer, *Voices in stone*, Paladin, London **1973**, 259-260.

Abstract

Based on the analysis of sound frequency in 17 languages there are found the limits above which the size of the database is sufficiently large, so that its size does not influence the results any more. These limits are: more than 700 single sounds, more than 8000 sound pairs, more than 30.000 sound triplets.

The criterion for single sounds fulfill all the used databases. The criterion for sound pairs is not fulfilled in the database of Old Phrygian, Oscan, Rhaetic and Venetic. The criterion for sound triplets is not fulfilled in the database of several additional languages. For this reason, there are of use first of all the results based on the frequency of single sounds. The selectivity of the approach, however, increases in direction single sounds < sound pairs < sound triplets.

Luvian and Mycenaean appeared to be outliers, having more vowels than the other tested languages, which may be the consequence of not having recognized several sound triplets of the type consonant-consonant-vowel or even consonant-consonant-consonant during their decipherment. Slavic sound groups of this type may be the remedy in this case.