PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 88 DECEMBER 2007 73–90

How Can Typological Distances between Latin and Some Indo-European Language Taxa Improve Its Classification?

Yuri Tambovtsev

Abstract

The article deals with an Antique language—Latin. A new method of phonostatistics is proposed here. It is based on the structure of the frequency of occurrence of consonants in the speech sound chain. It is a good clue for defining the typological closeness of languages. It allows a linguist to find the typological distances between Latin and the other languages of different genetic groups of the Indo-European language family. This method can put any language in a language taxon, i.e. a sub-group, a group or a family. The minimum distance may be a good clue for placing Latin in this or that language taxon. The method of calculating Euclidean distances is used. It adds new information for classifying languages.

Key words: consonants, phonological, distance, typology, frequency of occurrence, speech sound chain, statistics, Euclidean distances, closeness, language taxon, taxa of languages, classification.

The aim of the article is to analyse an Antique language—Latin in order to put it in this or that language taxon. The new method of phonostatistics developed by the author is proposed here (**Tambovtsev, 1977; 2002-c; 2002-d; 2003-b; 2004-a; 2004-b**). It allows a linguist to find the typological distances between the languages under study (**Tambovtsev, 1994-b; 2001-d; 2002-a**). The obtained distances indicate to which language taxon a language belongs. In fact, the received language distances show similarity between the languages in question, the less the distance—the more similar the languages (**Tambovtsev, 2001-e; 2002-b; 2004-a**).

Now Latin is classified into the Italic group of the Indo-European language family (**Crystal, 1992: 199; JaDM, 1982: 19**). However, not so long ago Latin was placed into one group with the Romance languages (**Chikobava, 1953: 207–208**). May be, it is more logical, when the parent language is in the same group with its offsprings. It would be very strange if we put Old Slavonic in some separate group, but not in the Slavonic group. Our method shows the typological distances which may lit light on the closeness of Latin to the Romance languages since it is not

Please cite this article as: Yuri Tambovtsev, How Can Typological Distances between Latin and Some Indo-European Language Taxa Improve Its Classification?. The Prague Bulletin of Mathematical Linguistics No. 88, 2007, 73–90.

possible to find enough long and reliable texts in the true Italic languages: Faliscan, Oscan, Umbrian and Venetic which are dead by now. Therefore, Latin may have been placed in this language group for the lack of information. Though the number of texts in the Italic languages is limited and they are short, there are some linguists who claim that Latin belongs to the group of Italic languages. Rex E. Wallace goes even further than that. He claims without much evidence that Latin enters the Latino-Faliscan group of the Italic branch of the Indo-European language family (**Wallace, 2001: 412**). One must pay attention to the fact that he opens a new group and a new branch. More logically it is to call his new group the Latino-Faliscan subgroup. While his new branch is nothing else but the commonly accepted Italic group within the Indo-European language family. Though the information on the other Italic languages is scarce and unreliable, Rex E. Wallace insists that Oscan, Umbrian, South Picene, Vestinian, Marrucinian, Paelignian, Marsian, Volscian, Aequian and Hernican are more distant from Latin that Faliscan (**Wallace, 2001: 412**). However, it is quite possible that all the Italic languages mentioned above are just the sub-dialects and dialects of Latin. Though usually Latin is a term for the Classical Latin language, which was used only by the educated classes of Rome. Rex E. Wallace correctly points out that there were numerous different sub-dialects and dialects of Latin. He is also right to state that there were different variants of Latin for different social levels, e.g. Vulgar Latin as the speech of the common folk (**Wallace, 2001: 412**).

It is possible to agree that meanwhile it is advisable to place Latin into the Italic group of the Indo-European language family until more solid and reliable information is received. At the same time one cannot agree to the fact that this group is called a language family. A fair representative of the linguists who believe that there could be a family inside a family is David Crystal (**Crystal, 1992: 199**). Unfortunately, he is not the only one who makes a logical mistake like this. April McMahon and Robert McMahon also speak about the Germanic family, which is embraced into the Indo-European language family (**McMahon et al., 2005: 3-4**). However, if one takes into consideration all the reasoning of their book, one may realise that the abundance of data leads them to the conclusion that Indo-European family looks like a sort of a super-family, called here **a language unity**, i.e. the next level of classification. Usually, the languages as the objects at this higher level are not so similar as at the lower levels. If a classification is correct, i.e. natural, then the languages at the lower levels are more similar (**Tambovtsev, 2004-a: 201–210; Tambovtsev, 2004-b: 147–151).**

It is high time to reconsider all the established language families and other language taxa. If it is done so, then it may be discovered that Italic and Romance groups must be merged together into one group called Romano-Italic with two subgroups: Romance and Italic. There are some arguments, which allow us to do it. One of the arguments may be the distance between Latin and the Romance languages (**Tambovtsev, 2001-a**). If Latin is closer to the languages of the Romance group of languages, then it surely belongs to them, rather than to any other set of languages. Our results show the shortest mean distance of Latin to the languages of the Romance group, than to the other languages (c.f. Tab 1–13).

It is good to see that the logical mistake of classification described above is not made by other classifiers. Thus, Kenneth Katzner calls Italic a subgroup of the Indo-European language family (**Katzner, 1986:2**). However, strictly speaking he also makes a sort of a logical mistake,

since his subgroup does not enter a group, but a family. Thus, he omits one classification step. A logical classification of languages must incorporate subgroups into a group, groups into a family, families into a unity, unities into a phylum, phyla into a union, unions into a language community (**Tambovtsev, 2004: 145**).

It is high time to establish a universal and strict logical hierarchy of language taxa. All the linguists in the world should keep to one and the same order of language taxa (**Tambovtsev, 2003-a: 3**). The ordered series of the taxa of the world languages should include old and dead languages like Latin, Old Greek, Old Russian, Old Turkic, etc (**Tambovtsev, 2001-b; Tambovtsev, 2001-b; Tambovtsev, 2001-c**). While reconsidering and building new language taxa linguists should take into account the special rules. First of all it is the idea that they must separate all world languages into sets in such a way that the distances between languages in a language taxon must be less than the distances of these languages to the other world languages (**Tambovtsev, 2003-a**). The structure of a taxon is more dense (tight), that is compact, if the languages selected for it are more similar (**Tambovtsev, 2002-b**). In our studies it is usually the total of the distances between the ideal language in this or that set of language, which is expressed by the mean of a set (**Tambovtsev, 2001-e**). In a compact set the distances between the mean and the other values are minimal. First we developed this idea of compact and sparse sets of languages on the data of the frequency of occurrence of phonemes in the speech chain (**Tambovtsev, 1977**). Then, we went on applying the idea of the measure of compactness on the basis of the consonantal coefficient, which is the ratio of the frequency of occurrence (**Tambovtsev, 1986**)

We have nothing against placing Latin into the group of Italic languages of the Indo-European language family. Nevertheless, it is necessary to point out that in physics, chemistry, biology and other natural sciences old classifications are often reconsidered (**Kuhn, 1977; Rozova, 1986**). We must also point out that basing on the same known Indo-European isoglosses, Tomas V. Gamkrelidze and Vjacheslav Vs. Ivanov do not construct the group of Romance languages and the Italic group of the Indo-European language family. Instead, they define only one group of languages, i.e. the Italic group. Presumably, their Italic group embraces both Italic and Romance languages, since they do not provide a separate Romance group (**Gamkrelidze et al.,1984: 415**). It is fruitful that they also include not only the phonetical but the lexical and grammatical isoglosses, which allows them to obtain a more complete and reliable scheme. We have analysed this scheme in detail elsewhere and came to conclusion that their scheme is different from the usual traditional one in this aspect (**Tambovtsev, 1989, 134–137**).

Comparing the distances between Latin and Old Greek or Modern Greek one must bear in mind that Old Greek and Modern Greek are considered genetically isolated languages (**Crystal, 1992: 11; JaDM, 1982: 23**). There are some other languages, which have not been placed into any language family: Basque, Japanese, Korean, Ainu, Nivhi, Yukaghir and Ket (Yug). However, for the latter, a new language family—Yenissey has been invented. So, now Ket with all its dialects is the only memeber of the Yenissey family. Nevertheless, it is not a solution of the problem. If we follow this way, then we must also establish separate language families for Ainu, Basque, Japanese and the other isolated languages.

The new data, which we received for Latin may allow it to enter this or that group of lan-

guages. It is the first attempt to establish the phonostatistical measures for the typological closeness of Latin with the language groups, to which it may be supposed to enter. Usually, genetically close languages are also typologically close. However, the typologically close languages may be or may not be genetically close. Nevertheless, in the majority of cases typologically close languages are genetically close. We can find the phonostatistical closeness, which can give a good clue for the genetic relatedness. It was found for some Finno-Ugric, Turkic, Mongolic, Tungus-Manchurian and Paleo-Asiatic languages (**Tambovtsev, 2001-d; 2001-e; 2002-a; 2002-b; 2002-c; 2002-d; 2003-a; 2003-b; 2004**). Therefore, it is a good reason to believe this method should also work for Latin or any other language.

Why should one use quantitative methods in studying languages? A great philosopher and scientist Immanuil Kant (1724–1804) in his well-known works explaining the structure of the world stated that everything in this world possesses quantity and quality. Quantitative data characterise an object sometimes better, especially when the objects are very similar. Languages are similar in their qualitative characteristics. This is why, one should rely on the quantitative characteristics more. Actually, quantity may go over into quality when it is great enough (**FS, 1980: 144**). In this case, English is a fair example. Must it be considered a Germanic or a Romance language? Many words of its stock are of Romance origin as the result of the Norman Conquest in 1066. It is believed that quantitative characteristics work better in the cases when qualitative characteristics fail to distinguish two linguistic objects.

Long ago, in 1935, George Kingsley Zipf stated that it was necessary to introduce the so-called "Dynamic Philology" to achieve fruitful results in studying the structure and entity of Language (**Zipf. 1935:XII**). As George A. Miller correctly put in the introduction to Zipf's book, one who wishes to study a rose should count its petals, not just enjoy it. G. K. Zipf believed that it is necessary to study the massive statistical regularity of every linguistic unit or phenomenon (**Zipf, 1935:V–VI**).

Quantitative research needs the use of mathematical statistics. One can't help agreeing with Christopher Butler, who requires a quantitative treatment in any linguistic research because it is difficult otherwise to understand and evaluate how relevant are the linguistic results (**Butler, 1998: 255–264**).

Establishing genetic language families linguists compare every language with some other language or a group of languages. Jiri Kramsky is correct to remark that one can establish a typology of languages basing on the quantitative data received after comparing languages. The quantitative data gives a clearer vision of the differences and similarities between languages. The quantitative load of particular language phenomena is different in different languages. Kramsky is quite right to observe that in linguistics there is a very close relation between quality and quantity, even if the conditions of the transition of quantity into quality are not established so safely as they are in natural sciences. Nevertheless J. Kramsky assumes that in linguistics qualitative changes are asserted with the help of quantitative factors (**Kramsky, 1972: 15**).

Our method measures distances between languages on the phonological level. It gives a vivid picture of the typological similarity of the sound pictures of the languages under investigation. It allows us to find out the archetype of this or that language family. The mean values of the frequency of the consonantal groups

The use of quantitative data ensures that the languages are similar if the frequency of occurrence of certain linguistic units are similar. It takes into account both cases when the units are used very frequently or very seldomly. However, in classical linguistics, where the frequency is not taken into consideration, it is more often than not that the usual elements are compared with the rare elements. J. Kramsky is correct to point out that the language units which are in the centre of some language system should not be compared to those of the periphery (**Kramsky, 1972: 15**). The quantitative analysis shows us the units, which are in the centre of a language system and those which are at the periphery of it. Therefore, the typology of languages based on the quantitative data may add much to the established language families (**Tambovtsev, 2001-a; 2001-b; 2001-c; 2003**).

Latin, as any other human language, has a specific structure of the speech sound chain. It can be distinguished by its structure from any other language. Every language has a unique structure of distributions of speech sounds in its phonemic chain. The distribution of Latin vowels will not be considered till the second stage of the investigation. The frequency of occurrence will be considered if and only if the frequency of occurrence of different groups of consonants will not differentiate Latin from the other world languages. Let's point out that consonants bear the semantic load in the word, not vowels. Therefore, it is more possible to understand the meaning of the message by consonants, rather by vowels. Some linguists use consonants to consider statistical models in language taxonomy.

Let us consider the way one of statistical methods, namely, Chi-square is applied to place English and German in one group. On the basis of the frequency of fricative consonants [s] and [f] Alan Ross proved, and then April and Robert McMahon proved again that English and German are related, i.e. the use of these fricative consonants is not random (**McMahon et al., 2005: 59–61**). Actually, an outstanding American mathematician of Hungarian origin G. Polya used the same way of reasoning to establish the similarity of Hungarian to English, Swedish, Danish, Dutch, German, French, Spanish, Italian and Polish. He came to the conclusion on the sample of ten numerals that Hungarian is quite different from these languages (**Polya, 1975: 315–319**)

However, if we fail to recognise and distinguish two languages, then we resort to the structure of occurrence of vowels in the speech sound chain. While comparing languages, it is necessary to keep to the principle of commensurability. Having it in mind, it is not possible to compare languages on the basis of the frequency of occurrence of separate phonemes, because the sets of phonemes in languages are usually different. The articulartory features may serve as the basic features in phono-typological reasoning.

Before the computer measures the phonological distances, one has to choose the phonological features, which are necessary and sufficient. One has to select the system of the informative features. In pattern recognition such features are called basic (**Zagoruiko, 1972: 54–75**). Therefore, we have chosen all the features basic for the articulation of any speech sound. At the first stage we shall deal with consonants.

First of all, it is the classification of consonants according to the work of the active organ of speech or place of articulation (4 features). Secondly, it is the classification from the point of view of the manner of articulation or the type of the obstruction (3 features). Thirdly, it

is the classification according to the work of the vocal cords (1 feature). In this way, 8 basic features are obtained: 1) labial; 2) forelingual or front; 3) mediolingual or palatal; 4) guttural or back or velar; 5) sonorant; 6) occlusive non-sonorant; 7) fricative non-sonorant; and 8) voiced non-sonorant consonants. One should take the values of the frequency of occurrence of these 8 features in the speech chain of Latin and compare them to those of the other languages. On the basis of the "chi-square" test and Euclidean distance, we have developed our own method of measuring the phono-typological distances between languages (**Tambovtsev, 1994-a; 1994-b; 2004**). It takes into account the frequency of occurrence of the 8 consonantal groups mentioned above and builds up the overwhelming mosaic of the language sound picture.

It is very important to find some typological characteristics in order to endeavour to place it in some defined language family. Some linguists consider it impossible to put Latin in any of the known language families because it was unsufficiently studied before. Actually, it is considered here that it is possible to put Latin in a language family if its phonostatistical characteristics are studied better. Therefore, we undertook the study of the frequency of Latin phonemes on the vast sample of Latin texts. Fortunately, unlike the other Italic languages mentioned above, Latin has an abundance of reliable texts.

We fed into the computer the following Latin texts: 1) Latin proverbs and sayings from the book by V. N. Kuprejanova and N. M. Umnova and small texts by different Latin authors from the book by Ja.M. Borovskij and Bildyrev (Kuprejanova et al., 1975; Borovskij et al., 1949). 2) Aeneid by Vergilius.

After Aleksandr A. Derjugin, Larisa M. Lukjanova, Ja.M. Borovskij and A. V. Boldyrev, we define the following Latin phonemes:

Vowels: [i, u, e, o, a, i:, u:, e:, o:, a:, ae, oe, au, eu]

Consonants: [p, b, v, f, m, t, d, ts, s, z, n, l, r, j, k, g, h]

The classification of the Latin consonants by the work of the active of speech (i.e. place of articulation):

Labial: [p, b, v, f, m]

Forelingual (front): [t, d, ts, s, z, n, l, r]

Mediolingual (palatal): [j]

Guttural (velar or back): [k, g, h]

The classification by the manner of articulation (the character of the obstruction):

Sonorant: [m, n, l, r, j]

Occlusive non-sonorant: [p, b, t, d, ts, k, g]

Fricative non-sonorant: [v, f, s, z, h]

The classification by the work of the vocal cords:

Voiced non-sonorant consonants: [b, v, d, z, g]

After computing the Latin text by V. N. Kuprejanova, N. M. Umnova, Ja.M. Borovskij and A.V. Boldyrev, we received the following frequencies of the phonemic occurrence in the sound chain:

	Frequency	% to all ph.	% to cons.
Labial:	4561	13.82	24.12
Forelingual (front)	12248	37.12	64.77
Palatal (mediolingual)	140	0.42	0.73
Guttural (back)	1964	5.95	10.38
Sonorant	7463	22.62	39.47
Occlusive non-sonorant	7297	22.11	38.58
Fricative non-sonorant	4153	12.58	21.95
Voiced non-sonorant	2702	8.19	14.29

The total of consonants: 18913 phonemes — 57.31 %

The total of vowels: 14087 — 42.69 %

The value of the consonantal coefficient (i.e. the ration of consonants to vowels): 1.34

Sample volume of the Latin proverbs: 33000 phonemes.

Zip's data has the following frequency of the phonemic occurrence in the sound chain (Zipf et al., 1939):

	Frequency	% to all ph.	% to cons.
Labial:	560	11.20	20.86
Forelingual (front)	1705	34.10	63.50
Palatal (mediolingual)	25	0.50	0.93
Guttural (back)	395	7.90	14.71
Sonorant	1076	21.52	40.07
Occlusive non-sonorant	1149	22.98	42.79
Fricative non-sonorant	460	9.20	17.13
Voiced non-sonorant	260	5.20	9.68

The total of consonants: 2685 phonemes — 53.70 %

The total of vowels: 2315 — 46.30 %

The value of the consonantal coefficient (i.e. the ratio of consonants to vowels): 1.16

Sample volume of the Zipf's Latin text: 5000 phonemes.

The author has also computed the epic poem "Aeneidos" by Vergilius. It is long and consists of 12 chapters describing the legends dedicated to Rome. Publius Vergilius Maro received a good education in philosophy, poetry and rhetoric. He wrote his poem for some 11 years. It is considered to be a good sample of classical Latin. "Aeneid" has the following frequency of the phonemic occurrence in the sound chain:

	Frequency	% to all ph.	% to cons.
Labial:	43514	12.15	21.19
Forelingual (front)	135892	37.95	66.20
Palatal (mediolingual)	1504	0.41	0.72
Guttural (back)	24411	6.82	11.89
Sonorant	80515	22.48	39.21
Occlusive non-sonorant	82351	23.00	40.12
Fricative non-sonorant	42455	11.85	20.67
Voiced non-sonorant	25218	7.04	12.28

The total of consonants: 205321 phonemes — 57.33 %
The total of vowels: 152800 — 42.67 %
The value of the consonantal coefficient (i.e. the ratio of consonants to vowels — 1.34
Sample volume of the Latin text of Aeneid: 358121 phonemes.

The united data computed by the author consists of Latin proverbs and "Aeneid". It has the following frequency of the phonemic occurrence in the sound chain:

	Frequency	% to all ph.	% to cons.
Labial:	48075	12.29	20.97
Forelingual (front)	148140	37.88	64.63
Palatal (mediolingual)	1644	0.42	0.73
Guttural (back)	26375	6.74	11.76
Sonorant	87978	22.49	39.23
Occlusive non-sonorant	89648	22.92	39.98
Fricative non-sonorant	46608	11.92	20.79
Voiced non-sonorant	27920	7.14	12.45

The total of consonants: 224234 phonemes — 57.33 %
The total of vowels: 166887 — 42.67 %
The value of the consonantal coefficient (i.e. the ratio of consonants to vowels — 1.34
Sample volume of the Latin text of Aeneid: 358121 phonemes.

It is recommended to use in linguistics some exact measure to place the languages more objectively. In pattern recognition such exact measures of distances between two objects are used. Nikolai G. Zagoruiko recommends to use the Euclidean distances when the value of the features are equal (**Zagoruiko, 1999: 198–199**). We consider all our features to be equal since we cannot claim that the frequency of occurrence of labials is more important than the frequency of occurrence of sonorants, or the frequency of occurrence of palatals is more important than the frequency of occurrence of the fricatives and so on.

We measure here the distances by the well-known formula of measuring the distance between points in the Euclidean space:

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 + \text{etc.}}$$

where
D - distance
x_1 - the frequency of occurrence of labials in Latin
x_2 - the frequency of occurrence of labials in the second language
y_1 - the frequency of occurrence of front consonants in Latin
y_2 - the frequency of occurrence of front consonants in the second language
z_1 - the frequency of occurrence of palatals in Latin
z_2 - the frequency of palatals in the second language, etc.

The details of calculating Euclidean distances may be found elsewhere (**Tambovtsev, 2003-c: 122**). This method is good because it can use any number of features in any number of languages. Therefore, a linguist can take as many linguistic features as he wants. The number

of languages is not limited either. So, this method calculated the distance between Basque and Latin (10.54). Though the least distances were between Basque and Kazah (5.310) or Tofalar (5.96) and the other Turkic languages (**Tambovtsev, 2003-c: 125**).

It is necessary to introduce some system of references when dealing with the distances between Latin and the other languages. Such point may be the distance between two texts in some language. We calculated the distances between two texts in the Markiz language, one of the Austronesian languages. It is 0.505. Now let us take any other language as a point for the system of references. It can be any language, which is far away from Latin and the contacts with which is not probable. Such a language may be Ainu. The native speakers of Ainu live in Japan. So the influence of Latin on Ainu is not possible. For calculating the distances between Ainu and the other languages we used the same method. The language closest to Ainu is one of the Austronesian languages—Tagalog with the distance of 9.310. The closest language to Latin by this method is Moldavian (4.275), then comes Italian (5.242) and then Romanian (6.913). We can see that Latin is much closer to Moldavian, than Ainu to its closest language. In fact, it is by two times closer. We can see the other distances between Latin and Romance languages in Tab. 1.

The least distance between Latin and Moldavian means that they are the closest languages among the chosen Romance and other languages (c.f. Tab. 1–13). It is not surprising since Moldavian and Romanian are spoken by the descendants of Roman soldiers and settlers, who occupied the Roman province of Dacia (Carlton, 2001: 598). In my mind, Italian, Moldavian and Romanian preserved the articulation base of Latin and thus the frequency of occurrence of sounds in Latin and in these languages is more similar, than in the others. Actually, the smallest distance between Latin and Moldavian may speak for many more remnants in Moldavian rather than in Italian. It is always so that at the periphery there are more obsolete features than in the centre. These distances may also point out that the articulation base of these three languages is rather similar.

As a matter of fact, articulation base is the main factor in ruling the frequency of occurrence of speech sounds in any language. We can see it on the examples of other languages, e.g. Ainu. Let us remember the words of N. A. Nevskij that Ainu is close to Paleo-Asiatic languages (**Tambovtsev, 2001-b**). Indeed, one of the Paleo-Asiatic languages, i.e. the Chookchi language with the distance 10.954 is rather close. The next closest language is also a Paleo-Asiatic language—Koriak with the distance 12.781. Korean is a bit closer — 12.636. Japanese is more far away — 15.269. As we can see from the tables below the other languages are also rather far away. So, the closest Tungus-Manchurian language is Ul'ch with the distance 13.464.

However, the most close to Ainu turned the American Indian languages of the North and South America. So, Quechua has the distance of 5.451 and Inga 7.388. They both belong to the Quechua family of American Indian languages. Quechua and Inga Indians live in South America.

Let us take some other languages as reference points. Japanese is a good choice since it is an isolated language. Having compared Japanese to some languages, we received the following phono-typological distances: Japanese–Ujgur (6.77); Japanese–Nanaj (8.12); Japanese–Jakut (8.26); Japanese–See Dajak (8.86); Japanese–Kazah (9.02); Japanese–Turkish (9.05); Japanese–Ket

(9.52); Japanese–Baraba Tatar (9.76); Japanese–Uzbek (10.63); Japanese–Hausa (10.98); Japanese–Georgean (11.05); Japanese–Kazan Tatar (11.07) and so on. One can see, that Ujgur, Jakut, Kazah, Turkish, Baraba Tatar, Uzbek and Kazan Tatar are Turkic languages. Nanaj is a Tungus-Manchurian language. Therefore, one can notice that Japanese is closer to the so-called Altaic languages which include Turkic, Mongolian and Tungus-Manchurian languages. Many world languages were compared to Japanese. We can't show all the distances here for the lack of space. However, the maximum distances were found for Japanese–German (22,24); Japanese–English (19.83); Japanese–Rumanian (15,08) and Japanese–Swedish (17.03). As a conclusion, we can also state that speech sound picture of Japanese is rather far away from the languages, which are geographically close: Chinese, Nivh, Itelmen or Indonesian. It was a surprise to us. Our data state that the speech sound pattern of Japanese resembles that of Ujgur–one of the Turkic languages spoken in the Middle Asia. The Ujgur people are often linked to the Old Turkic tribes, who used to live in the stepps of Southern Russia before the Tatar-Mongols captured them in the 9th century A.D. We must point out that it is not a coincidence since the other native Altaic people have a very similar data of closeness to Japanese. Turkic and Tungus-Manchurian tribes may have had a sort of common origin with Japanese. It may verify the Altaic hypothesis of Japanese origin. It is especially vivid, when the Austro-Oceanic and other languages do not show such a great closeness.

Considering the mean distance between Latin and the other languages and sets of languages, one may notice a clear preference. The mean distance between Latin and the Romance languages is the least 6.706 (c.f. Tab.1). The Baltic languages (Latvian and Lithuanian) are also rather close (8.504) to Latin (c.f. Tab.5). Latin is closer in general to the Eastern Slavonic languages (Russian, Old Russian, Ukrainian and Belorussian), than to the other two Slavonic subgroups. The mean distance is less (9.259) than that of Latin to Southern Slavonic (9.810) or Western Slavonic (13.008). So, it speaks again for similarity between Eastern and Southern Slavonic subgroups (c.f. Tab 1–4).

The Iranian group is closer (10.673), than Germanic (11.160) or Indic (12.400) groups.

It is possible to see that Old Greek (8.482) and Modern Greek (8.653) are not so close to Latin. However, Armenian is a bit further (8.838). Albanian is not close enough either (9.325).

Nevertheless, the Indo-European languages are closer to Latin than the Samoyedic family (15.400) or the Ob-Ugrian subgroup of the Ugric group of the Finno-Ugric family(16.333). The Northern dialect of Mansi (19.017) or the Konda dialect of Mansi (18.261) may be the champions (c.f. Tab. 14).

In conclusion, it is possible to state a great typological closeness between Latin and some languages of the Romance group of the Indo-European family. We are far from stating that genetically Latin is closer to the languages of the Romance group than to the languages of the Italic group. However, from the point of view of typology Latin is very similar to the Romance languages. Having this typological clue, linguists may have a closer look at Latin from the genetic point of view. May be, it is advisable to reconsider both Italic and Romance groups and unite them into one group Romano-Italic with two sub-groups: Romance and Italic.

References

Borovskij et al., 1949 - Borovskij Ja. M. and Boldyrev A. V. Latinskij jazyk [Latin]. - Moskva: ILIJA, 1949.

BSE - Bol'shaja Sovetskaja Entsyklopedija [Great Soviet Encyclopaedia]. 30 Volumes. Moskva: Sovetskaja Entsiklopedija Publishing House, 1978.

Butler, 1998 - Butler, Christopher. Statistics. - In: Projects in Linguistics. A Practical Guide to Researching Language. Alison Wray, Kate Trott and Aileen Bloomer with Shirley Reay and Chris Butler. - London-New York: Arnold-Hodder, 1998.–303 p.

Carlton, 2001 - Carlton, Charles M. Romanian. - In: Facts About the World's Languages: An Encyclopaedia of the World's Major Languages, Past and Present (Edited by Jane Garry and Carl Rubino). - New York and Dublin: The H. W. Wilson Company- A New England Publishing Associates Book, 2001, p. 598–601.

Chikobava, 1953 - Chikobava A, S. Vvedenie v jazykoznanie [Introduction into Linguistics]. - Moskva: Prosveshchenie, 1953.–243 pages.

Crystal, 1992 - Crystal, David. An Ecyclopedic Dictionary of Language and Languages. - Oxford: Blackwell, 1992–428 p.

FS, 1980 - Filosofskij slovar' [Phylosophy Dictionary] (Editor I. T. Frolova). - Moskva: Politizdat, 1980–445 pages.

Derjugin et al., 1986 - Derjugin, Aleksandr A., Lukjanova Larisa M. Latinskij jzyk [Latin]. - Moskva: Vysshaja shola, 1986–295 p.

Gamkrelidze et al., 1984 - Tomas V. Gamkrelidze and Vjach. Vs. Ivanov. Indo-European and the Indo-Europeans. Volume 1 and 2. [in Russian] - Tbilisi: The Tbilisi University Press, 1984–1392 p.

JaDM, 1982 - Jazyki i dialekty mira. [Languages and Dialects of the World]. - Moskva: Nauka, 1982.

Katzner, 1986 - Katzner, Kenneth. The Languages of the World. - London: Routledge and Kegan Paul Publishers, 1986–376 pages.

Kuhn, 1977 - Kuhn T. S. The Structure of Scientific Revolutions [in Russian]. - Moskva: Progress Publishing House, 1977.–300 pages.

McMahon et al, 2005 - McMahon, April and McMahon, Robert. Language Classification by Numbers. - Oxford: Oxford University Press, 2005.–263 pages.

Polya, 1975 - Polya, George. Mathematics and Plausible Reasoning. - Moskva: Nauka, 1975. - 463 pages. [Russian translation of his book published in 1954 by Princeton University Press].

Rozova, 1986 - Rozova S. S. Klassifikatsionnaja problema v sovremennoj nauke [Classification Problem in Modern Science]. - Novosibirsk: Nauka, 1986.–223 pages.

Tambovtsev, 1977 - Tambovtsev, Yuri. Nekotorye harakteristiki raspredelenija fonem mansijskogo jazyka [Some Characteristics of the Distribution of the Mansi Language]. - In: Soviet Finno-Ugric Studies, Volume XIII, # 3, 1977 (Tallin), p. 195–198.

Tambovtsev, 1986 - Tambovtsev, Yuri. Konsonantnyj koeffitsient v jazykah raznyh semej [Consonant Coefficient in the Languages of Different Families]. - Novosibirsk: Novosibirsk

University, 1986. - 17 pages.

Tambovtsev, 1989 - Tambovtsev, Yuri. Review on the book by Tomas V. Gamkrelidze and Vjach. Vs. Ivanov. Indo-European and the Indo-Europeans. Tbilisi: The Tbilisi University Press, 1984. Volume 1 and 2. - In: Gengo Kenkyu, # 96, 1989 (Tokyo), p. 119–143.

Tambovtsev, 1994-a - Tambovtsev, Yuri. Dinamika funktsionorovanija fonem v zvukovyh tsepochkah jazykov razlichnogo stroja. - Novosibirsk : NGU, 1994-a. - 133 pages.

Tambovtsev, 1994-b - Tambovtsev, Yuri. Tipologija uporjadochennosti zvukovyh tsepej v jazyke. - Novosibirsk: NGU, 1994-b. - 199 pages.

Tambovtsev, 2001-a - Tambovtsev, Yuri. Kompendium osnovnyh statisticheskih harakteristik funktsionirovanija soglasnyh fonem v zvukovoj tsepochke anglijskogo, nemetskogo, frantsuskogo i drugih indoevropejskih jazykov. [Compendium of the basic statistical characteristics of functioning of consonants in the speech chain of English, German, French and other Indo-European languages]. - Novosibirsk: Novosibirskij klassicheskij institut, 2001-a. - 129 pages.

Tambovtsev, 2001-b - Tambovtsev, Yuri. Funktsionorovanie soglasnyh fonem v zvukovoj tsepochke uralo-altajskih jazykov. [Functioning of consonants in the sound chain of Ural-Altaic languages]. Novosibirsk: Novosibirskij klassicheskij institut, 2001-b. - 132 pages.

Tambovtsev, 2001-c - Tambovtsev, Yuri, Nekotorye teoreticheskie polozhenija tipologii upor'adochennosti fonem v zvukovoj tsepochke jazyka i kompendium statisticheskih harakteristik osnovnyh grupp soglasnyh fonem. [Some theoretical fundamentals of the typology of oderliness of phonemes in the sound chain of language and the compendium of statistical characteristics of the basic groups of consonants], - Novosibirsk: Novosibirskij klassicheskij institut, 2001-c. - 130 pages.

Tambovtsev, 2001-d - Tambovtsev, Yuri. The phonological distances between Mongolian and Turkic languages based on typological consonantal features. - In: Mongolian studies. Journal of the Mongolia Society (USA), Vol.24, 2001-d, p.41–84.

Tambovtsev, 2001-e - Tambovtsev, Yuri. Perednejazychnye soglasnye kak pokazatel' odnoj iz lingvisticheskih universalij jazykov mira [Front consonants as an indicator for one of the universals of world languages]. - In: Sibirskij Lingvisticheskij Seminar, # 2 (2), Novosibirsk, 2001-e, p.18–27.

Tambovtsev, 2002-a - Tambovtsev, Yuri. Comparative typological study of language distances based on the consonants in sound chains of various languages. - In: The 5th National Colloquium for Computational Linguistics in the UK. Proceedings of the Conference. (Edited by John Elliot). 8-9 January, 2002. University of Leeds, UK. - Leeds: University of Leeds, 2002-a, p.77–80.

Tambovtsev, 2002-b - Tambovtsev, Yuri. Is Kumandin a Turkic language? - In: Dilbilim Arashtirmalari (Istanbul), 2002-b, p. 63–104.

Tambovtsev, 2002-c - Tambovtsev, Yuri. Korean and Japanese as Members of the Altaic Language Family. - In: Abstracts. Permanent International Altaistic Conference 45th Meeting, Budapest June 23-28, 2002. Budapest: Research Group for Altaic Studies. Hungarian Academy of Sciences, 2002-c, p. 13–14.

Tambovtsev, 2002-d - Tambovtsev, Yuri. Structure of the frequency of occurrence of con-

sonants in the speech sound chain as an indicator of the phono-typological closeness of languages. - In: ALL - ACH 2002. New Directions in Humanities Computing. The 14th Joint International Conference, University of Tuebingen, 24–28 July, 2002. Conference Abstracts. - Tuebingen: Universitaet Tuebingen, 2002-d, p. 138–139.

Tambovtsev, 2003-a - Tambovtsev, Yuri. Tipologija Funktsionorovanija fonem v zvukovoj tsepochke indoevropejskih, paleoaziatskih, uralo-altajskih i drugih jazykov mira: kompaktnost' podgrupp, grupp, semej i drugih jazykovyh taksonov. [Typology of functioning of phonemes in sound chain of Indo-European, Paleo-Asiatic, Ural-Altaic and other world languages: compactness of subgroups, groups, families and other language taxins]. - Novosibirsk: Sibirskij nezavisimyj institut, 2003-a. - 143 pages. [in Russian].

Tambovtsev, 2003-b - Tambovtsev, Yuri. The phonological similarity between Turkic languages based on some phonological features of consonants. - In: Linguistic and Oriental Studies from Poznan', Vol. 5, Poznan', (Poland), 2003-b, p. 85–118.

Tambovtsev, 2003-c - Tambovtsev, Yuri. Phonological similarity between Basque and other world languages based on the frequency of occurrence of certain typological consonantal features. - In: The Mathematical Bulletin of Mathematical Linguistics # 79–80, 2003 (Praha), p. 121–126.

Tambovtsev, 2004-a - Tambovtsev, Yuri. Uralic Language Taxon: Natural or Artificial? (Typological Compactness of Uralic Languages and other Language Taxons: Branches, Subgroups, Groups, Families and Superfamilies). - In: Fenno-Ugristica, # 26, 2004 (Tartu), p. 200–246.

Tambovtsev, 2004-b - Tambovtsev, Yuri. Phonostatistical Distance Totals as an Indicator of the Compactness of Language Taxons: a Typologo-Metrical Approach. - In: Lingua Posnaniensis. Vol. XLVI, Poznan' (Poland), p. 145–172.

Wallace, 2001 - Wallace, Rex E. Latin. - In: Facts About the World's Languages: An Encyclopaedia of the World's Major Languages, Past and Present (Edited by Jane Garry and Carl Rubino). - New York and Dublin: The H. W. Wilson Company- A New England Publishing Associates Book, 2001, p.201–416.

Zagoruiko, 1972 - Zagoruiko, Nikolai G. Metody raspoznavanija i ih primenenie. [The Methods of Pattern Recognition and Their Application] (in Russian). - Moskva: Sovetskoe Radio, 1972.–206 pages.

Zagoruiko, 1999 - Prikladnye metody analiza dannyh i znanij. [Applied Methods of Data and Knowledge Analysis](in Russian). - Novosibirsk: Institute of Mathematics SORAN, 1999. - 269 pages.

Zipf, 1935 - Zipf, George Kingsley. The Psycho-Biology of Language. An Introduction to Psycho-Biology of Language. - Cambridge: Massachusetts Institute of Technology Prtess, 1935. - 336 pages.

EUCLIDEAN DISTANCES between Latin and other world languages, united in different genetic families and other language taxa

Tab. 1

Phonostatistical EUCLIDEAN DISTANCES between Latin and Romance language group of the Indo-European language family. The mean of the distances — 6.706.

Language	Distance
	Latin
0. Latin	0.000
1. Moldavian	4.275
2. Italian	5.242
3. Rumanian	6.913
4. Spanish	7.353
5. Portuguese	9.747

Tab. 2

Phonostatistical EUCLIDEAN DISTANCES between Latin and the Eastern Subgroup of the Slavonic language group of the Indo-European language family.

The mean of the distances – 9.259.

Language	Distance
	Latin
0. Latin	0.000
1. Russian	4.275
2. Old Russian	9.048
3. Belorussian	10.124
4. Ukrainian	10.169

Tab. 3

Phonostatistical EUCLIDEAN DISTANCES between Latin and the Southern Subgroup of the Slavonic language group of the Indo-European language family.

The mean of the distances – 9.810.

Language	Distance
	Latin
0. Latin	0.000
1. Macedonian	7.502
2. Slovenian	8.582
3. Serbian	9.579
4. Bulgarian	13.577

Tab. 4

Phonostatistical EUCLIDEAN DISTANCES between Latin and the Western Subgroup of the Slavonic language group of the Indo-European language family.

The mean of the distances – 13.008.

Language	Distance
	Latin
0. Latin	0.000
1. Slovak	11.653
2. Czech	11.743
3. Luzhits-Sorbian	11.789
4. Polish	16.848

Tab.5

Phonostatistical EUCLIDEAN DISTANCES between Latin and Baltic language group of the Indo-European language family.

The mean of the distances – 8.504.

Language	Distance
	Latin
0. Latin	0.000
1. Latvian	7.344
2. Lithuanian	9.664

Tab. 6

Phonostatistical EUCLIDEAN DISTANCES between Latin and Indic language group of the Indo-European language family.

The mean of the distances – 9.231.

Language	Distance
	Latin
0. Latin	0.000
1. Gypsy	6.939
2. Sanskrit	8.074
3. Marathi	8.097
4. Bengali	10.268
5. Hindi	12.779

Tab. 7

Phonostatistical EUCLIDEAN DISTANCES between Latin and Iranian language group of the Indo-European language family.

The mean of the distances – 10.673.

Language	Distance
	Latin
0. Latin	0.000
1. Persian (Iranian)	7.877
2. Osetian	9.804
3. Tadjik	14.338

Tab. 8

Phonostatistical EUCLIDEAN DISTANCES between Latin and Celtic language group of the Indo-European language family.

Language	Distance
	Latin
0. Latin	0.000
1. Irish	13.057

Tab. 9

Phonostatistical EUCLIDEAN DISTANCES between Latin and Germanic language group of the Indo-European language family.

The mean of the distances – 11.160.

Language	Distance
	Latin
0. Latin	0.000
1. Dutch	8.075
2. Norwegian	8.793
3. Old English	10.002
4. English	11.763
5. Gothic	12.258
6. German	16.067

Tab. 10

Phonostatistical EUCLIDEAN DISTANCES between Latin and Isolated languages of the Indo-European language family.

Language	Distance
	Latin
0. Latin	0.000
1. Old Greek	8.482
2. Modern Greek	8.653
3. Armenian	8.838
4. Albanian	9.325

Tab. 11

Phonostatistical EUCLIDEAN DISTANCES between Latin and Esperanto—an artificial language.

Language	Distance
	Latin
0. Latin	0.000
1. Esperanto	7.330

Tab. 12

Phonostatistical EUCLIDEAN DISTANCES between Latin and the Ob-Ugric Subgroup of the Ugric language group of the Finno-Ugric language family.

The mean of the distances – 16.333.

Language	Distance
	Latin
0. Latin	0.000
1. Eastern Hanty	11.823
2. Kazym Hanty	16.231
3. Konda Mansi	18.261
4. Northern Mansi	19.017

Tab. 13

Phonostatistical EUCLIDEAN DISTANCES between Latin and the Samoedic language family.

The mean of the distances – 15.400.

Language	Distance
	Latin
0. Latin	0.000
1. Nenets	14.375
2. Nganasan	15.572
3. Selkup	16.252

Tab. 14

The Ordered Series of the Mean Phonostatistical EUCLIDEAN DISTANCES between Latin and Some Subgroups and Groups of the Indo-European family. The mean of the distances inside every language taxon.

Language	Mean Distance
	Latin
0. Latin	0.000
1. Romance	6.706
2. Baltic	8.504
3. Eastern Slavonic	9.259
4. Southern Slavonic	9.810
5. Iranian	10.673
6. Germanic	11.160
7. Indic	12.400
8. Western Slavonic	13.008