Phonological similarity between Basque and other world languages based on the frequency of occurrence of certain typological consonantal features.

Yuri Tambovtsev.
Novosibirsk Pedagogical University, Novosibirsk

The aim of the present typological study of the Basque language based on the frequency of occurrence of consonants is to compare its sound chain to sound chains of various world languages. Then, one can construct typological distances between Basque and other world languages.

The structure of the frequency of occurrence of consonants in the speech sound chain is a good clue of understanding the typological closeness of languages. Typological investigation of human languages involves the task of the analysis of their speech chains by a human being or computer. A human being can realise that this or that language sounds closer to his own native language without understanding the meaning. At the stage when it is hard to teach computer to understand a human language, it is quite possible to make it recognise the sound closeness of a language to this or that language on the basis of the analysis of its sound speech chain. We have computed the frequency of phonemic occurrence of 119 world languages as a teaching sample for the computer. Then we took Basque as a token language. Actually, Basque was chosen for only one reason. It was taken up mainly because Basque as well as Japanese, Korean, Ainu, Burushaski, Nivh (Gilyak), Yukaghir are considered to be isolated languages, i.e. languages that do not belong to the known language families. In fact, they may be relics of the former family of languages (Crystal, 1992: 425). Basque is a fair example of an isolate language. Efforts have been made to show a relationship with Caucasian languages, North African languages, and Iberian, but none has been convincing (Crystal, 1992: 40). Merritt Ruhlen considers John Bengtson to be correct to include Basque, Burushaski and Shumerian into the new Dene-Caucasian family (Ruhlen, 1994: 25). R. L. Trask believes Basque to be unquestionably the last surviving pre-Indo-European language in Europe. He links Basque with the dead Aquitanian language (Trask, 2001: 72).

We tried to find out the distances between world languages with the help of computer in order to achieve the optimal and unbiased results. The computer had to measure the distances between a chosen language and the rest of the languages in the set. Then, the computer had to put it in a matrix: closer to some languages and far away from the others.We have chosen Basque and the other isolated languages (e.g. Basque, Yukaghir, Nivh, Korean and Japanese) because they are not categorically assigned to any language family. It is interesting to compare how the computer and how different linguists place them in different language families and super-families: Indo-European, Turkic, Finno-Ugric, Tungus-Manchurian, Uralic, Ural-Altaic, Paleo-Asiatic; Sino-Tibetan; Austro-Asiatic; etc. We measure the distances between Basque and the other world languages in the same way we did it with Japanese (Tambovtsev, 1988: 19). We believe that under the circumstances even some traits and hints from the typological point may help to find the languages genetically related to Basque (Tambovtsev, 2001: 83-85).

In this study we have used the procedures that are usually used in pattern recognition First of all, one must choose the features which should be necessary and sufficient (Zagoruiko, 1972). We believe the chosen features to be the most informative from the phonetic point of view.

Basque, 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 Basque vowels will not be considered until the second stage of the investigation, that is, if the data on the frequency of occurrence of consonants will not allow us to distinguish between the languages under investigation. 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 writing systems (Hebrew, Arabic, etc.) are a fair example of that since they denote only consonants. However, if we fail to recognise and distinguish two languages by consonantal groups, then we resort to the second stage of investigation, in which the frequency structure of occurrence of vowels in the speech sound chain is taken into consideration. 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 articulatory features may serve as the basic features in phono-typological reasoning.

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: labial, front, palatal, velar).

Secondly, it is the classification from the point of view of the manner of articulation or the type of the obstruction (3 features: sonorant, occlusive, fricative).

Thirdly, it is the classification according to the work of the vocal cords (1 feature: voiced).

In this way, 8 basic features are obtained: 1) labial; 2) front; 3) mediolingual or palatal; 4) back or velar; 5) sonorant; 6) occlusive; 7) fricative; and 8) voiced consonants.

It is necessary to mention that the total of these 8 features embrace all the main space of features which serve as the fundamental for the characteristics of consonants from the point of view of its production in any human language. Every language is being characterised as a point in the 8 dimensional space. This is why, it is important to take the features, which don't cross, i.e. which can be derived from one another. Our 8 features comprise the complete articulation system. One can hardly find any other basic articulatory feature. This is why, one can be sure that our typological classification is not just one more arbitrary typological classification of languages since it involves all the basic articulatory features of consonants.

One should take the values of the frequency of occurrence of these 8 features in the speech chain of Japanese 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, 2001). However, here our measurements use the well-known formula of Euclidean distance:

D = Ù x1 -x2)? + (y1 - y2)? + (z1 - z2)? + etc.

where x1 - is the frequency of occurrence of the first feature in the first language; in this case the frequency of labial consonants in Basque;

x2 - is the frequency of occurrence of the first feature (labials) in the second language (any other world language);

y1 - is the frequency of occurrence of the second feature (dentals and alveolars) in the first language;

y2 - is the frequency of occurrence of the second feature (dentals and alveolars) in the second language;

z1 - is the frequency of occurrence of the third feature (palatals) in the first language;

z2 - is the frequency of occurrence of the third feature (palatals) in the second language; etc.

In this way Basque is compared to all the languages of the world taken for this study. Therefore, one can see that this formula allows us to take any number of features and any number of languages. It is very important that the number of features and languages is not limited. However, now one should bear in mind that it takes into account the frequency of occurrence of the 8 basic consonantal groups mentioned above and builds up the overwhelming mosaic of the language sound picture.

We could not find the data on the frequency of occurrence of Basque language, that is why, we think that it is calculated for the first time.

We computed several Basque texts from two books. The first of them is "Emakume biboteduna" - The woman with a moustache (Xabier Montoia, 1992) with 3 tales in it and the second one is "Sei Ipuin Amodiozko"- Six stories of love (Xabier Mendiguren Elizegi, 1992) with 6 tales. We thank two native speakers of Basque Kepa Sarasola and Maite Oronoz for their help and useful advice. We defined Basque phonological segment inventory the same as Ian Maddieson (Maddieson, 1981: 168) and R. L. Trask (Trask, 2001: 73 - 74). The first text that we computed was "Amusa Margarita" by Xabier Mendiguren Elizegi. After computing the first Basque text, we received the following results:

	Frequency	*% to all phonemes*	*% to consonants*
Labial:	1021	5.09	9.69
Front:	7496	37.36	71.15
Palatal:	100	0.50	0.95
Velar:	1919	9.56	18.21
Sonorant:	4458	22.22	42.32
Occlusive:	4233	21.10	40.18
Fricative:	1845	9.19	17.50
Voiced:	2480	12.36	23.54

Consonant total - 10536 phonemes or 52.51%.
Vowel total - 9530 phonemes or 47.49%.
Consonantal coefficient, i.e. ratio of consonants to vowels - 1.11
Basque first linguistic sample - 20 066 phonemes.
The frequency of occurrence of Basque phonemes in the greater sample (130 550 phonemes) of 9 different stories showed similar results.

After computing the greater Basque sample, we received the following results:

	Frequency	*% to all phonemes*	*% to consonants*
Labial:	7470	5.72	11.13
Front:	46857	35.89	69.87
Palatal:	602	0.47	0.92
Velar:	12130	9.29	18.08
Sonorant:	26410	20.23	39.38
Occlusive:	28613	21.92	42.67
Fricative:	12036	9.22	17.95
Voiced:	17046	13.06	25.42

Consonant total - 67059 phonemes or 51.37%.
Vowel total - 63491 phonemes or 48.63%.
Consonantal coefficient, i.e. ratio of consonants to vowels - 1.06
Basque linguistic sample - 130 550 phonemes.

First of all, we measured the phonostatistical distances between Basque and 40 Indo-European languages. One can see the details in Tab. 1. The typologically closest language is Gypsy (7.89), the farthest - Vedic (21.09). The closest Finno-Ugric language is Ludikov dialect of Karel (6.90), the farthest - the Northern dialect of Mansi (23.30), see Tab. 2. Turkic languages seem to be the closest to Basque: Kazah (5.31), Tofalar (5.96), Tuvin (6.83), Altai- Chalkan (7.00), Shor (7.04), Kirgiz (7.09), Uzbek (7.42), Hakas (7.70), Tatar-Kazan (7.73), Turkish (7.84). Ten Turkic languages are closer to Basque, then the closest Indo-European language - Gypsy (Tab. 3). Mongolian languages are also close enough: Mongolian (7.28), Buriat (8.48) and Kalmyk (8.81). Tungus-Manchurian languages seem to be rather far away (Tab. 4): the nearest Nanai (12.51). Paleo-Asiatic languages are also far away (Tab. 5): the closest Koriak (18.07). The isolated languages have the following order: Korean (9.64), Ket (13.05), Yukaghir (20.43) and Nivh (24.29). In order to see if the distances given above are small enough, let us compare them to the distances between some other languages.

It is interesting to find out that Korean value of the consonantal coefficient obtained here (1.16) is close to those of Japanese (1.08), Oroch (1.00), Nanaj (1.02) and Ulch (1.10), which all belong to the Tungus-Manchurian language family (Tambovtsev, 1985).

Having compared Korean and Japanese to some languages by the formula of Euclidean distance, we received the following phono-typological distances: Korean - Baraba Tatar (5.19); Korean Turkish (5.63); Korean - Ujgur (5.69). It is interesting to find out that the phonostatistical distance to Japanese is greater than to the three Turkic languages mentioned above: Korean - Japanese (6.66). On the other hand, it is important to see that Japanes is also close to the Turkic languages: Japanese - Ujgur (6.77). One can see that Korean is closer to the Turkic languages than Japanese. Let's compare some distances: Korean - Jakut (7.58) and Japanese - Jakut (8.26); Korean - Kazah (7.52) and Japanese Kazah (9.02); Korean – Turkish (5.63) and Japanese - Turkish (9.05); Korean - Baraba Tatar (5.19) and Japanese – Baraba Tatar (9.76); Korean - Uzbek (8.67) and Japanese - Uzbek (10.63); Korean Kazan Tatar (7.52) and Japanese - Kazan Tatar (11.07) and so on. It is interesting to measure the pholological distances between Korean, Japanese, Ket, Nivh, Yukhagir and other isolated languages to American Indian languages since some scholars consider it possible that they are related. We obtained the following ordered series between Japanese and Korean on the one hand, and on the other hand some of the American Indian languages, the first number is for Japanese, the second for Korean: Siriano (11.71 - 12.20), Inga (11.72- 9.13), Nahuatl (12.73 - 10.94), Mam (13.47 - 14.66), Kadiweu (13.52 - 15.69), Cofan (13.57 - 18.48), Guambian (14.91 - 11.69), Kaiwa (16.87 - 19.54), Pokomchi (17.05 - 17.34), Guarani (18.18 - 19.42), Totonak (18.28 - 18.42), Apinaye (19.06 - 15.38), and Secoya (20.87 - 23.57). So, one can see, that Japanese and Korean have more or less the same values of distances, which show that they are not typologically close to the American Indian languages taken for this study. 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. All in all 120 world languages were compared to Japanese. We cannot 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 state that speech sound picture of Japanese is rather far away from the Languages that are geographically close: Korjak (15.35), Nivh (20.23), Itelmen (17.85) or Chookchee (19.53). 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 IXth 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. Our typological results may verify the Altaic hypothesis of Japanese and Korean origin, since Korean showed a greater closeness to the Altaic languages than Japanese. It is especially vivid, when the Austro-Oceanic, Austronesian, Austro-Asiatic and other languages do not show such a closeness: Japanese- Chinese (17.52); Burmanese (20.24); Tagalog (14.36); Indonesian (10.33); Hawaiian (23.97); Samoan (21.56), but for Sea Dajak (8.86), which is a Polynisian language. The Dajaks live on the Kalimantan island.

As a conclusion, we can say that the phonostatistical distances allow us to find out how similar Basque is to the other world languages under investigation.

References:

Crystal, David 1992 *An Encyclopedic Dictionary of Language and Languages*. Oxford: Blackwell

Garry, Jane and Rubino, Carl (eds) 2001 *Facts About the World's Languages: an Encyclopedia of the World's Major Languages, Past and Present*. New York: The H. W. Wilson Company.

Ramstedt G. J. 1951 *A Korean Grammar.* Moskva: IzdInostrLiter.

Ruhlen, Merritt. 1994 *On the Origin of Languages (Studies in Language Taxonomy).* Stanford: Stanford university press.

Tambovtsev, Yuri. 1985 The consonantal coefficient in selected languages. *Canadian Journal of Linguistics*, Vol.30, # 2: 179-188.

Tambovtsev, Yuri.1988 The linguistics distances among some languages of Asia. - In: *The Study of Sounds.* Tokyo. Vol.22: 17 - 34.

Tambovtsev, Yuri. 2001a Kompendium osnovnyh statisticheskih harakteristik funktsionirovanija soglasnyh fonem v zvukovoj tsepochke anglijskogo, nemetskogo, frentsuzskogo 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.

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

Tambovtsev, Yuri 2001c Nekotorye teoreticheskie polozhenija tipologii upor'adochennosti fonem v zvukovoj tsepochke jazyka i kompendium osnovnyh group soglasnyh [Some Theoretical foundations of the typology of orderliness of phonemes in the language sound chain and compendium of statistical characteristics of the basic groups of consonants]. Novosibirsk: Novosibirskij klassicheskij institut.

Trask, R. L. 2001 Basque. - In: Garry, Jane and Rubino, Carl (eds). *Facts About the World's Languages: an Encyclopedia of the World's Major Languages, Past and Present*. New York: The H. W. Wilson Company.

Zagoruiko Nikolay G. 1972 The methods of pattern recognition and their application. [in Russian]. - Moskva: Sovetskoe radio.

Tab.1
Distances between Basque and Indo-European languages Basque

Basque	0
Gyps	7.89
Spanish	8.03
Moldavian	8.63
Bengal	8.96
Norwegian	9.34
Latvian	9.36
Persian	9.36
Osetian	9.36
Italian	9.68
Esperanto	10.04
Marathi	10.14
Latin	10.54
Portuguese	10.70
Dutch	10.71
Russian	10.84
Rumanian	11.21
Old Greek	11.27
SerboCro.	11.44
Lithuanian	11.64
Arminian	11.67
Old English	11.80
Hindi	12.00
Tadjik	12.05
Albanian	12.36
Greek	12.75
Sanscrit	12.97
Czech	13.32
Slovak	13.72
English	13.78
Bulgarian	14.06
Belorussian	14.31
German	14.45
Danish	15.71
French	16.10
Irish	16.16
Prakrit	17.55
Ukrainian	17.68
Gudjarati	17.73
Polish	18.52
Verdic	21.09