Linguistic Distances Based on Counting of Equal Sounds in Numerals from 1 to 10 in Different Language Groups

Anton Perdih[1]

[1]Faculty of Chemistry and Chemical Technology, University of Ljubljana, Slovenia

Correspondence: Anton Perdih, Faculty of Chemistry and Chemical Technology, University of Ljubljana, Slovenia.
E-mail: a.perdih@gmail.com

Abstract

The relative frequency of equal sounds in pairs of adjacent numerals from 1 to 10 in languages of eleven language groups is a basis for calculation of linguistic distances. By this criterion, the Slavic languages form a cluster separated from all other tested languages. Of other languages, Avestan and Sanskrit are the closest to them. The Germanic languages form another cluster but this cluster is within the space of other tested languages, which are widely dispersed. This is an additional indication that the aboriginal Proto-Indo-European was Proto-Slavic and their speakers were the aboriginal Europeans: mainly the Y Chromosome haplogroup I, mtDNA haplogroup U people. In contact with newcomers of other language groups either the newcomers turned to Proto-Slavic, or the previously Proto-Slavic speakers lost their Proto-Slavic at all, or they turned the non-Indo-European newcomers into Indo-European. A novel time line for Nostratic studies is proposed.

Keywords: Sound frequency, gender sensitive, language clusters, language distance, language group, Nostratic

1. Introduction

During the last two centuries, the comparative linguistics achieved a vast amount of results and it continues refining the knowledge about the languages, their relations, etc. There remained, however, also some possibilities for the niche approaches. One of the niche approaches was the observation that the Slavic numerals rhyme whereas the non-Slavic numerals rhyme to a much lesser extent (Jandáček, 2004). Later this observation was amended (Jandáček, 2013) with an additional intention to see whether it is possible to create on the basis of rhyming of numerals an instrument, which calculates the Space/Time of the Evolution of the many Indo-European languages from their Proto-Slavic Prototype. Considering the Jandáček's (2013) results another question appeared, namely whether there are equal sounds in words for numerals not only in the rhymes but also elsewhere in the words. The objective of present study is thus to see whether besides the rhyming approach also the approach of counting of matching sounds in pairs of adjacent numerals from 1 to 10, regardless whether they form a rhyme or not, would give some additional information.

The Jandáček's (2013) approach left open two questions. One of them was the gender sensitiveness of numerals. Observing that in some languages the numerals are gender sensitive, Jandáček (2013) proposed this feature to be neglected in order to obtain as high numerical values of rhyming as possible.

The other question was the question of the weight of vowels and consonants in order to present the numerical values of rhyming. Jandáček (2013) weighted the matching consonants with the weight of four points, since the consonants are more stable than vowels. For that reason he weighted the matching vowels with the weight of two points. In present paper both of these questions were considered in the methodology.

2. Methodology

2.1 Gender Sensitiveness of Numerals

In English, for example, the numerals are not gender sensitive: one man, one woman etc. In Slavic, Slovene for example, they are gender sensitive up to the numeral 4, Table 1. To be consistent, the masculine form of the numeral is used wherever this dependence is known.

Table 1. Gender sensitiveness of numerals in Slovene.

Numeral	masculine	feminine	neuter
1	en mož	ena žena	eno dete
2	dva moža	dve ženi	dve deteti
3	trije možje	tri žene	tri deteta
4	štirje možje	štiri žene	štiri deteta
5	pet mož	pet žena	pet detet

2.2 Sound Weighting

On the one hand, the consonants are more stable than vowels. On the other hand, however, due to the higher sound impression there are more impressive the vowels. In order not to put deliberate weights it was decided not to distinguish the weights of consonants and vowels but to use the principle of one sound one point. As a result of it here is the weight of the consonants and the vowels of one point each.

In order to avoid the dispute where exists the formal rhyme and where not, here are considered all matching sounds, not only those forming a pure rhyme although this reduces the selectivity of the approach.

2.3 Lenght Compensation

To compensate for the different lenghts of words for numerals in different languages the relative number of matching sounds is used. Namely, in longer words there is in principle a higher probability that more sounds are equal.

2.4 Pair matching

Jandáček (2013) used first of all the comparison of rhyming of the adjacent pairs of odd-to-even numerals, which rhyme in Slavic languages. Here are used strictly both types of matching, i.e. odd-to-even (1-2, 3-4, 5-6, 7-8, 9-10) as well as even-to-odd (2-3, 4-5, 6-7, 8-9).

2.5 Data

The words for numerals of most language groups were taken from Rosenfelder and his grouping of languages is followed here. The words for numerals of Mongol & Tungusic languages were taken from Blažek and Schwarz (2016).

Considered were 16 Slavic languages (including some Slovene dialects), 3 Baltic, 4 Celtic, 6 Germanic, 9 Romance, 12 other Indo-European languages, 8 Caucasian, 21 Turkic, 12 Uralic, 8 Tamil, and 17 Mongol & Tungusic languages.

2.6 The Way of Counting

The counting of equal sounds in the same gender form in adjacent pairs of numerals from 1 to 10 was performed in two ways:

odd-to-even: 1-2, 3-4, 5-6, 7-8, 9-10

even-to odd: 2-3, 4-5, 6-7, 8-9

This has been made in spite of the awarness that the numerals greater than 20, for example 21 and 22, in Czech sound *dvacet jedna* or *jedna a dvacet*, *dvacet dva* or *dva a dvacet*, or, in Slovene *enaindvajset*, *dvaindvajset*. However, these numerals are not gender sensitive.

As the example is presented in Table 2 the way of counting of equal sounds in whole words in adjacent pairs of numerals from 1 to 10 in Slovene and English and its results, including the number of all sounds and the relative number of matching sounds. Presented is the literary writing of words for numerals, whereas counted are the equal sounds.

Table 2. The way of counting and its results

Language	1	2	3	4	5	6	7	8	9	10	Total
Slovene	eden	dva	trije	štirje	pet	šest	sedem	osem	devet	deset	
odd-to-even	1		5		2		3		4		15/44 = 0.341
even-to odd		0		2		2		1			5/35 = 0.143
English	one	two	three	four	five	six	seven	eight	nine	ten	
odd-to-even	0		0		0		1		1		2/35 = 0.057
even-to odd		1		1		1		0			3/29 = 0.103

2.7 Distance Calculation

The distance between languages was calculated in the Pithagorean way,

$$d_{ij} = ((x_i-x_j)^2+(y_i-y_j)^2)^{1/2}$$

where d_{ij} is the distance between the language i and the language j; x is the relative number of matching sounds in odd-to-even numeral pairs; y is the relative number of matching sounds in even-to-odd numeral pairs.

The average distance within a language group was calculated as

$$avg.\ dist. = average\ of\ all\ ((x_i-x_{avg})^2+(y_i-y_{avg})^2)^{1/2}$$

where x_{avg} and y_{avg} are the averages of the relative number of matching sounds in odd-to-even and even-to-odd numeral pairs, respectively.

As the largest distance between two languages within the language group was taken

$$largest\ distance = ((x_i-x_j)^2+(y_i-y_j)^2)^{1/2}{}_{max}$$

3. Results

3.1 The Relative Number of Matching Sounds

The counting results for all tested languages are presented in Figure 1 and the averages within particular language groups are presented in Figure 2.
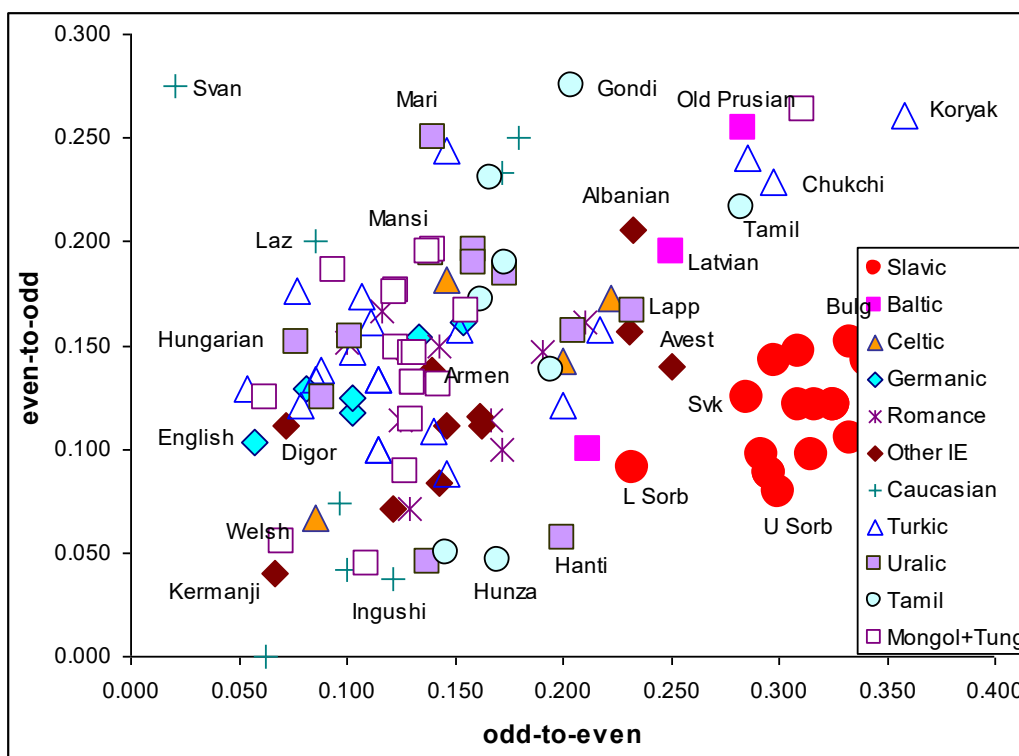

Figure 1. Relative number of equal sounds in pairs of adjacent numerals from 1 to 10 in different languages matching sounds in pairs of numerals from 1 to 10

In Figure 1 we see that the Slavic languages, for which the rhyming of words for numerals is characteristic (Jandáček, 2004, Jandáček, 2013), form a cluster separated from all other tested languages in spite of choosing here the less selective counting of all equal sounds instead of rhymes. The Germanic languages form another cluster but this cluster is within the space of other tested languages.

By the criterion used, all the tested non-Slavic languages are widely spread. The extreme cases are:

- The Kermanji (Kurdish) language with the lowest both odd-to-even and even-to-odd relative number of equal sounds in its numerals.

- The Svan language (Western Georgia) with the lowest odd-to-even relative number of equal sounds in its numerals

and the highest even-to-odd relative number.

- The Koryak language (Siberia) with the highest both odd-to-even and even-to-odd relative number of equal sounds in its numerals.
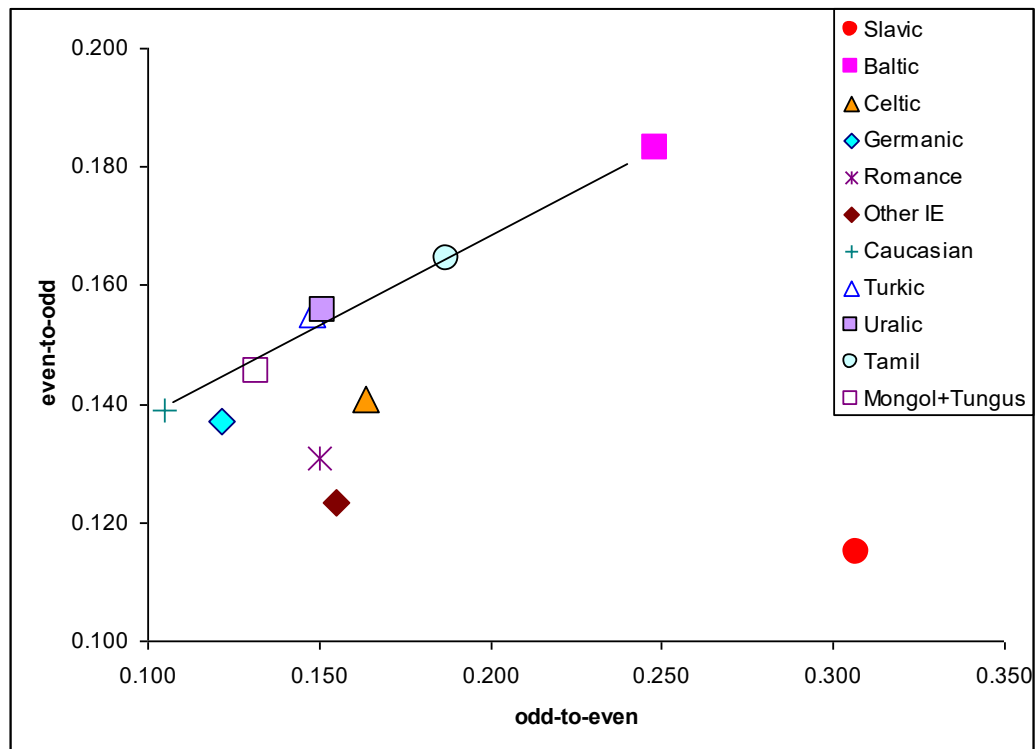

Figure 2. Averages of the relative number of equal sounds in numerals from 1 to 10 in different language groups

The position of the averages within particular language groups is presented in Figure 2. The average position of Slavic languages is positioned far from the other language groups, especially by the odd-to-even criterion, reflecting also the observation of rhyming (Jandáček, 2004, Jandáček, 2013). The average of the Baltic languages is by the odd-to-even criterion positioned between the Slavic and other languages, but by the even-to-odd criterion it departs from all of them. The average of other Indo-European languages is by the odd-to-even criterion close to other language groups, whereas by the even-to-odd criterion it is close to Slavic.

Interesting is also that the average positions of Turkic and Uralic languages are close together and they are positioned close to the line which would connect the averages of Caucasian and Baltic languages. Close to that line are positioned also the averages of Mongol & Tungusic as well as of Tamil languages. Farther from that line in direction towards the other Indo-European languages are positioned the averages of Germanic, Celtic, and Romance languages.

3.2 The Distances between Languages

In Table 3 are presented the average distances of languages from their average position in Figure 1 and 2 within particular language groups.

Table 3. Average internal distance and maximum internal distance between languages within particular language groups in Figure 1

Language group	Avg. distance	Max. distance
Slavic	0.030	0.120
Germanic	0.039	0.151
Romance	0.043	0.121
Mongol & Tungusic	0.053	0.318
Other IE	0.058	0.235
Baltic	0.061	0.171
Celtic	0.064	0.173
Uralic	0.066	0.205
Turkic	0.078	0.331
Tamil	0.084	0.232
Caucasian	0.111	0.278

In Table 3 we see in numerical form that the Slavic languages form the most compact cluster by both criteria presented here, followed by Germanic and Romance ones. If the apparent outlier Lower Sorbian would be omitted, then the average distance within the Slavic languages would amount 0.027 and the maximum distance would amount 0.080.

Whereas according to the spread measured by the average distance the series of increasing distance in Indo-European languages is Slavic < Germanic < Romance < other IE < Baltic < Celtic, by the maximum distance the series is Slavic < Romance < Germanic < Baltic < Celtic < other IE languages. Of the other IE languages, Avestan and Sanskrit are the closest to the Slavic group. Avestan (0.066) is closer than Lower Sorbian (0.083), whereas Sanskrit (0.090) is almost as close as the latter.

Among the non-IE languages, the series are as follows. By the average distance, it is Mongol & Tungusic < Uralic < Turkic < Tamil < Caucasian. By the maximum distance it is Uralic < Tamil < Caucasian < Turkic < Mongol & Tungusic.

4. Discussion

For the separate clustering of Slavic languages there are two conflicting explanations. Among the scholars there is generally accepted the explanation that the Slavic languages are very young (Slovene, for example, less than about 1.500 years) and therefore they are not yet as differentiated as the languages of other language groups. In the case of the Slovene language, it is based on the supposition of arrival of Slavs in the 6th Century AD from a small area east of the Carpathian Mountains. Their settling in isolated valleys would be followed by fast diversification of their previously common language. However, in view of the existence of almost 50 Slovenian quite differing dialects, of which some have very archaic characteristics (Kurkina, 1996), this explanation seems hardly valid. On the other hand, the geolinguists know that in isolation a language is quite stable, whereas in contact with other languages it changes much faster. Thus just the contrary to the above mentioned academic wisdom.

Tomažič (2006) found a report, *Vita Sancti Severini*, where it is explicitly stated that during the Roman occupation, the Romans lived outside Italy in towns and fortresses, whereas around them lived other people. And they were not in friendly relations to Romans. When the Romans left for Italy, there was not a population vacuum. The territory was free of Romans, but not without the other population. And soon thereafter there were records of a Slovenian state. In addition, there exists no report of that time about a Slavic intrusion towards west and south across the Carpathian Mountains and settling there.

In addition, on some Roman monuments there are some names now in usage among Slovenes, but not among non-Slavic peoples. These names existed in Slovenia before the alleged arrival of Slavs (Rant, 2011).

Looking at the development of the language of Aryans in India, there is observed the decrease of likeness to Slavic in the series Vedic > classical Sanskrit > present IE languages in India. They were becoming less and less Slavic-like as if they were losing their Slavic characteristics (Skulj & Sharda, 2002, Skulj, Sharda, Sonina & Jandaček, 2004, Skulj,

Sharda, Narale & Sonina, 2006, Skulj, Sharda, Sonina & Narale, 2008). A similar trend is observed with Avestan > present IE languages in Iran. If one extrapolates this back into past, it appears that the Pre-Vedic Aryans and Avestans spoke a more Slavic-like language. This is in line with the distance by the criterion introduced here. It indicates strongly that Proto-Slavic existed at least in central Europe before the ancestors of Aryans and Avestans left it about 4.500 years ago. This has been additionaly substantiated taking into account the DNA Genealogy data (Jandáček and Perdih, 2017, Perdih, 2016, Perdih, 2018), which indicate that after the catastrophe caused by the cosmogenic mega-tsunami about 68.000 years ago (Yurkovets, 2015, Yurkovets and Vasilenko, 2017) the survived humankind first expanded from Europe to other continents. Later on, the mixing of different branches took place as well as the recurring migrations into Europe.

There is to be drawn the attention also to the Slavic linguistic traces observed all over the southern, western, and northern Europe (references collected in Perdih, 2016, Perdih, 2018). These traces indicate that the aboriginal Proto-Indo-European was Proto-Slavic and their speakers were mainly the Y Chromosome haplogroup I, mtDNA haplogroup U people almost exclusively inhabiting Europe for about four tens of millenia. Other people (mainly G2a and possibly R1a males, as well as a variety of females) arrived there bringing agriculture, which happened about 8,000 to 6,000 years ago. After about 4,800 years ago there intruded into the west Europe the Altaic speaking R1b people eliminating most of the previous males. This was the beginning of the development of Celtic languages. The beginning of development of the Greek language was later, about 3,700 years ago. The beginning of development of the Romance and Germanic languages was later as well, about 3,200 years ago.

Whereas Jandáček (2013) tried to create on the basis of rhyming of numerals an instrument, which calculates the Space/Time of the Evolution of the many Indo-European languages from their Proto-Slavic Prototype, the analysis of matching sounds in numerals presented here has a broader aspect. A much broader aspect, of course, has the Nostratic research. In view of novel data provided by other disciplines, the time-scale of Nostratic explanation should be reconsidered. It seems to had started not approx. 15 millenia BCE in the Fertile Crescent (Bomhard, 2014/2015, Bomhard, 2018) but earlier, at least about 60 millenia ago if not much earlier (Perdih, 2018) in Europe and separately in Africa. It was followed by the primary expansion of people and the accompanying divergence of language during the time span of about 60,000 to 40,000 years ago, followed by the complex events of divergence and convergence on mixing at different times in different localities. This would mean a lot of additional time to reach the present diversity of languages.

5. Conclusion

The separate clustering of Slavic languages based on the relative number of matching sounds in odd-to-even and even-to-odd pairs in numerals 1 to 10 and the gradually increasing distance from them by Avestan, Sanskrit, and other Indo-European languages as well as the languages of other linguistic groups is in line with the conclusions based on DNA Genealogy data and data of Slavic linguistc traces in European languages. It is in line with the conclusion that the Y Chromosome haplogroup I people had been the aboriginal Europeans. In contact with newcomers of other language groups, different developments took place. In some cases the newcomers turned to Slavic and influenced it, as can be observed in present Slavic areas. In other cases the previously Proto-Slavic speakers lost their Proto-Slavic but turned the non-Indo-European newcomers into Indo-Europeans as can be observed in present western, southern and northern areas of Europe. In few cases the Proto-Slavic people lost their Proto-Slavic almost entirely, e.g. in present Finnland, Estonia, and Hungary.

References

Blažek, V., & Schwarz, M. (2016). Numerals in Mongolic and Tungusic languages with notes to code-switching. *Altai Hakpo, 26*, 25pp.

Bomhard, A. R. (2014 / 2015). The Nostratic Hypothesis in 2014. *Мова та історія*, Київ, *Nostratica Kioviensis in honore Vladislav M. Illič-Svityč, 359*, 15-56.

Bomhard, A. R. (2018). *A Comprehensive Introduction to Nostratic Comparative Linguistics: With Special Reference to Indo-European* (3rd ed.), pp. 281-312.

Jandáček, P. (2004). *Base Ten* counting as the extension of the archetypical *Base Five* system of Basques and Slavs. *Zbornik mednarodnega posveta Sledovi evropske preteklosti* [*Proceedings International workshop Traces of European Past*] (pp. 74-84). Ljubljana: Založništvo Jutro. http://www.korenine.si/zborniki/zbornik03/jandacek_baseten.htm

Jandáček, P. (2013). Numerical rhyme as an instrument measuring time/space between Indo-European languages. *Вестник Академии ДНК-генеалогии, 6*(11), 1898-1911. http://dna-academy.ru/proceedings/

Jandáček, P., & Perdih, A. (2017). Words for the Animal Elephant/Mammoth in Relation to the DNA Genealogy Data.

Advances in Anthropology 7, 251-272. https://doi.org/10.4236/aa.2017.74014

Куркина, Л. В. (1996). О лексических архаизмах Толминского диалекта [On the lexical archaisms of the Slovene dialect in the Tolmin area]. *Razprave II. razreda*, Ljubljana: Slovenska akademija znanosti in umetnosti (*Dissertationes classis II*, Academia scientiarum et artium Slovenica) *15*, 35-46.

Perdih, A. (2016). *Izvor Slovencev in drugih Evropejcev* [*Origin of Slovenes and other Europeans*] (2nd ed.). Ljubljana: Založništvo Jutro.

Perdih, A. (2018). Continuity of European Languages from the Point of View of DNA Genealogy. *International Journal of Social Science Studies*, *6*, 18-32. https://doi.org/10.11114/ijsss.v6i1.2809

Rant, J. (2011). *Imena odkrivajo zgodovino: Nekaj jezikovnih dokazov o avtohtonosti Slovencev v Vzhodnih Alpah in okolici*, Ljubljana: Založništvo Jutro.

Rosenfelder, M. *Numbers from 1 to 10 in over 5000 languages*. Retrieved September 6, 2017, from http://www.zompist.com/numbers.shtml

Skulj, J., & Sharda, J. C. (2002). Indo-Aryan and Slavic affinities. *Zbornik prve mednarodne konference Veneti v etnogenezi srednjeevropskega prebivalstva* [*Proceedings of the First International Topical Conference The Veneti within the Ethnogenesis of the Central-European Population*]. (pp.112-121), Ljubljana: Založništvo Jutro. http://www.korenine.si/zborniki/zbornik01/htm/skulj_indo.htm

Skulj, J., Sharda, J. C., Sonina, S., & Jandaček, P. (2004). Linguistic and genetic correspondences between Slavs and Indians (Indo-Aryans) provide insights into their pre-history. *Zbornik mednarodnega posveta Sledovi evropske preteklosti* [*Proceedings; International Workshop Traces of European Past*]. (pp. 43-73), Ljubljana: Založništvo Jutro. http://www.korenine.si/zborniki/zbornik03/skulj_relationship.htm

Skulj, J., Sharda, J. C., Narale, R., & Sonina, S. (2006). 'Lexical Self –Dating' Evidence for a Common Agro-Pastoral Origin of Sanskrit 'Gopati', 'Gospati' and Slavic 'Gospod', 'Gospodin' Meaning *Lord*/*Master*/*Gentleman* more than 8,000 Years Ago. *Vedic Science*, *8*(1), 5-24; and: *Zbornik četrte mednarodne konferenceEvropski staroselci* [*Proceedings of the Fourth International Topical Conference Ancient Inhabitants of Europe*], (pp. 40-58), Ljubljana: Založništvo Jutro. http://www.korenine.si/zborniki/zbornik06/skulj_sanskrt06.pdf

Skulj, J., Sharda, J. C., Sonina, S., & Narale, R. (2008). Indo-Aryan and Slavic linguistic and genetic affinities predate the origin of cereal farming. *Zbornik šeste mednarodne konference Izvor Evropejcev* [*Proceedings of the sixth international topical conference Origin of Europeans*], (pp. 5-39), Ljubljana: Založništvo Jutro. http://www.korenine.si/zborniki/zbornik08/indo_aryan.pdf

Tomažič, I. (2006). Staroselci Norika [The ancient inhabitants of Noricum]. *Zbornik četrte mednarodne konference Evropski staroselci* [*Proceedings of the Fourth International Topical Conference Ancient Inhabitants of Europe*], (pp. 5-11), Ljubljana: Založništvo Jutro. http://www.korenine.si/proceedings.htm#2006

Yurkovets, V. P. (2015). Климатическая катастрофа гаплогруппы «Бета». *Вестник Академии ДНК-генеалогии*, *8*(3) 376-432. http://dna-academy.ru/proceedings/

Yurkovets, V. P., & Vasilenko, S. I. (2017). ДНК-генеалогия, палеоклимат и геоморфология. Гидродинамический карст. *Вестник Академии ДНК-генеалогии, 10*(3) 1412-1442. http://dna-academy.ru/proceedings/

Copyrights