

# Mathematical Models of the Distribution and Change of Linguistic Information in Language Communities: a Case of Proto-Indo-European and Proto-Chinese Language Communities

Maia Egorova<sup>1\*</sup>, Alexander Egorov<sup>1</sup>, and Tatiana Solovieva<sup>1</sup>

<sup>1</sup>Peoples' Friendship University of Russia (RUDN University), RU-117198, Moscow, Russian

**Abstract.** The paper presents a theoretical analysis and computer simulations of the distribution and changes of the linguistic information in two model language communities: Proto-Indo-European and Proto-Chinese. Simulations show that out of two main hypotheses of the formation of the Proto-Indo-European languages, the Anatolian hypotheses and the Kurgan hypotheses, the latter is better consistent with the time estimates obtained in this study. The results obtained for Proto-Indo-European communities may also be used in the analysis of Asian language communities. In particular, the similarity of Chinese and Proto-Indo-European languages in terms of the relationship between the verb and the noun opens the possibility of applying our method to the analysis of the Proto-Sino-Tibetan language family. A possibility of creating a single national language Pǔtōnghuà (普通话) in the modern China was investigated. The results of the present study also suggest that the developed models look like a quite promising new instrument for studying linguistic information transfer in complex social and linguistic systems.

## 1 Introduction

To study the distribution and change of linguistic information in Proto-Indo-European (PIE) language communities, as well as to search for the ancestral homeland of the peoples that were carriers of the PIE language we investigated two mathematical models: the first is a dynamic system model described by a nonlinear equation; the second is the model described by the system of integral differential equations (see, for example, [1-3]).

Within these models, the distribution and change of linguistic information in the model of Indo-European (IE) language communities were numerically studied, including at the initial stage of its formation, and both regular and chaotic phenomena were discovered in the transmission of the linguistic information (see, for example, [2]).

The Chinese language has more than three thousand years of history (see, for example, [4, 5, 11-13]). It is one of two branches of the Sino-Tibetan (ST) language family of languages. Primarily, it was the language of the main ethnic group of China – the Han people (dominates the national composition of the PRC: more than 90% of the country's population). In its standard form, Chinese is the official language of the PRC and Taiwan, as well as one of the six official and working languages of the UN.

ST language family is one of the world's largest and most prominent families, spoken by nearly 1.4 billion

people [13]. As it is noted in [13], despite the importance of the ST languages, their prehistory remains controversial, with ongoing debate about when and where they originated. The authors of this paper [13] tried to shed light on this debate and developed a database of comparative linguistic data, and applied the linguistic comparative method to identify sound correspondences and establish cognates. Then they used phylogenetic methods to infer the relationships among these languages and estimate the age of their origin and homeland. So they pointed that Sino-Tibetan originating with north Chinese millet farmers around 7200 B.P. and suggest a link to the late Cishan and the early Yangshao cultures.

In our work, we rely on data generally recognized by specialists at the moment. So why we assumed that the start time of the separation of Proto-Sino-Tibetan (PST) language family, i.e. in fact, the time of the “appearance” of the hypothetical proto-Chinese language occurred approximately no later than 7000 years ago (see, for example, [4, 5, 12]).

From this point of view, the relevance of our study is obvious. Moreover, linguists have long noted the wave nature of many linguistic phenomena (see e.g. [15]). However, therefore far few mathematical models have been proposed that allow this to be taken into account.

In the case of the Proto-Chinese (PC) language communities we investigated a dynamic system model described by a nonlinear equation. It was assumed that the time of separation of the Proto-Sino-Tibetan (Proto-

\* e-mail: Mey1@list.ru

Chinese-Tibetan) language family, i.e. in fact, the time of the “appearance” of the hypothetical Proto-Chinese language occurred around 6500 years ago [4, 5].

The mathematical models of the distribution and change of linguistic information in language communities under study were built using the information obtained from independent studies as a priory data, both from linguistics and other scientific fields, for example, from history, archeology, and genetics [1-15].

Our results made it possible to re-evaluate the hypotheses put forward earlier and made it possible to propose new promising approaches to the study of the phenomenon of the transmission of linguistic information in complex socio-linguistic systems.

## 2 Dynamic nonlinear model of distribution and changes of linguistic information: PIE communities

### 2.1. Mathematical model and analysis

The dynamic model of distribution and change of the linguistic information  $I$  in some community can be described by the following nonlinear equation [1, 2]:

$$y(x) = \lambda_1 x(1-x) + \lambda_2 x(1-x)^2, \quad (1)$$

where  $y = I_{m+1}/M$ ,  $x = I_m/M$ ,  $\lambda_1 = a_1 M \lambda$ ,  $\lambda_2 = a_2 M \lambda$ ,  $m = 1, 2, \dots$  ( $m = 1$  corresponds to the first “measurement”, i.e.  $I_1$  is the initial value of the given information, for example, during some initial moment of time  $t_1$ );  $a_1$  is the factor characterizing distribution of the linguistic information at contacts of “ignorant” with “knowing” given information  $I_n$ ;  $a_2$  is the factor characterizing influence only on “ignorant”;  $M$  is the maximum value of the given linguistic information;  $\lambda$  is problem parameter (according to the catastrophe theory it can be named by the operating parameter);  $\lambda_{1,2}$  are new operating parameters of the system under study;  $0 \leq x \leq 1$ .

The equation (1) and its variants can be used in particular at research of the process of training, for example, children by adults in some linguistic community.

When parameter  $\lambda_1$  increases (from 0 to 4) the system following laws are observed ( $\lambda_2 < 1$ ). At  $\lambda_1 < 1$ ,  $y(x) \rightarrow 0$  (iterations of  $I_n$  converge to  $I^* = 0$ , a steady motionless point). At the further growth of the operating parameter  $\lambda_1$  occurs the bifurcation: the point  $I^* = 0$  loses stability, and again appeared point becomes steady (iterated values  $I_m \rightarrow const$ , i.e. the dynamic regime is stationary or has the period equal to 1, the cycle  $S^1$  is observed).

The further increase in parameter  $\lambda_1$  again leads to the system change: occurs bifurcation period doubling

(the motionless point is split and  $I_m$  oscillates between two values forming steady attractor with the period 2). In this case mapping (1) has a steady cycle with the period 2:  $S^2$ . At the further increase in parameter  $\lambda_1$  there are cycles of the kind  $S^{2^p}$ :  $S^1, S^2, S^4, S^8, S^{16}, S^{32}, S^{64} \dots$ , etc. (where  $p = 0, 1, 2, 3, 4, \dots$ ). As a result it is possible to observe branching process: the initial branch is divided on two, new two split again on two etc. Therefore, with reference to process of propagation of the linguistic information, it is possible still to name this model as a “tree model”.

The further growth of the operating parameter  $\lambda_1$  allows observing the system transition in a chaotic mode: acyclic, statistic process arises as a limit of more and more difficult structures (cycles of kind  $S^{2^p}$ ). Thus, the chaos arises as a limit of the super complicated organization. Between the order and chaos a deep internal communication is observed. Similar laws exist in any systems where the sequences of doubling period bifurcations are observed.

At such approach as one of quantitative characteristics of considered process of distribution of the “linguistic information” a number of arising cycles  $S^{2^p}$  can be consider as a number of the arisen languages (or dialects) in the given (simulator) language community; so cycles  $S^1, S^2, S^4, S^8, S^{16}, S^{32}$  give accordingly: 1, 2, 4, 8, 16 and 32 languages.

Then quite naturally it is possible to choose some “internal” (“built in”) time scale in the given process, namely – we will take advantage of known data in linguistics: each cycle in dynamics of linguistic (model) community corresponds on a time scale to the size of 500 years – time of divergence of two related languages.

As a result, the given nonlinear model (1) where there are branching (treelike) cycles of kind  $S^{2^p}$  allows us to use the aprioristic linguistic information on the time of divergence of languages as time scale. For example, between cycles  $S^2$  and  $S^4$  the distance on time is 500 years, thus each of two languages of predecessors in an initial cycle  $S^2$  have divided during 500 years on 2 new, related to initial language (two new “linguistic populations”), and as a result them became 4; and so on.

The cited data on a number of languages allow receiving following time estimations in the basis of which lays the assumption made above: the usage of the given nonlinear model where there are cycles of the kind  $S^{2^p}$ , permit us to receive not only the “built in” internal time scale within the step to 500 years, but also to define time “length” of given linguistic time length along which dynamics of researched linguistic system develops.

Actually, for the cycles of kind  $S^{2^p}$  the following number of values for number of languages  $L$  corresponding to the given number of consecutive them doubling (starting from one possible PIE (parent) language, a basis of all modern IE languages): 1, 2, 4, 8,

16, 32, 64, 128, 256, 512, 1024, 2048, etc; proceeding from it, it is easy to receive the following simple formula for an estimation of possible number of languages in such dynamic process:  $L = 2^p$  [2]. Whence we receive the formula for a numerical estimation of quantity of languages of cycles corresponding to the given number  $p$ :  $p = \log_2 L$ . For example, for 6000 languages it is received:  $p = \log_2 6000 \approx 12.55$ , i.e. approximately 13 generations (as well as for 8000 languages:  $p = \log_2 8000 \approx 12.97$ ), that corresponds approximately to time of years:  $13 \cdot 500 = 6500$  years. Only for IE languages we have:  $p = \log_2 500 \approx 8.97$ , i.e. approximately 9 generations or 4500 years [2]. Comparison of the obtained data with the data of independent researchers shows their good agreement [4, 6, 7, 9]. This confirms the validity of our approach to the study of such social-linguistic communities.

We use the data obtained in [1, 2] during numerical research of distribution and change of the linguistic information in some model Indo-European language community, including the initial stage of its formation. We will consider thus, that the time of the beginning of splitting (i.e. as a matter of fact “disappearance”) of the hypothetical PIE language has occurred approximately not later than 6500 (Kurgan hypothesis), or not later than 9500 (Anatolian hypothesis) years ago [5].

We should note that mathematical model of the process of wave propagation and change of linguistics information in this system is described by a system of integral-differential equations and it discussed in sufficient detail in our paper [2], therefore we do not consider it here.

## 2.2. Results for PIE communities

The most characteristic of the simulation results according Eq. (1) (taken in the following form:  $x_{m+1} = \lambda_1 x_m (1 - x_m) + \lambda_2 x_m (1 - x_m)^2$ ), that we obtained within the framework of this nonlinear model are shown in our papers [1, 2].

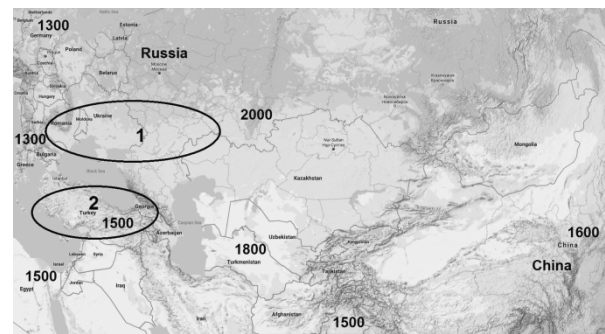
The results of computer simulations correspond perfectly in time to the two main hypotheses about the formation of the Proto-Indo-Europeans: Anatolian and Kurgan hypotheses [6-9] (see in fig. 1 ellipses marked by “1” and “2”). Let us briefly recall these two hypotheses.

The *Anatolian hypothesis* localizes the Indo-European ancestral homeland in western Anatolia (modern Turkey; see in fig. 1 ellipse marked by “2”).

The *Kurgan hypothesis* was proposed by Maria Gimbutas in 1956 and now is the most popular theory. To determine the location of the ancestral homeland of the carriers of the Proto-Indo-European language the data from archaeological and linguistic studies were used. According to the hypothesis, the Proto-Indo-European peoples existed in the Black Sea steppes and southeastern Europe (see in fig. 1 ellipse marked by “1”) in the period from approximately the 5th to the 3rd millennium BC [6-9].

The most important stage in the development of the Kurgan culture was the domestication of the horse and the use of carts, which made the culture carriers mobile and significantly expanded their influence. We used this fact as a foundation for the construction of the theoretical models. In particular, the ratio of the coefficients  $\lambda_{1,2}$  chosen in computer modeling based on the ratio of the average speeds (in km/h) of horsemen  $v_1$  and pedestrians  $v_2$  in the ancient Indo-European communities:  $\lambda_1 : \lambda_2 \approx v_1 : v_2 \approx 15 : 3$ .

As an example, figure 1 shows the so-called “Country of cities” – the conventional name of the territory in the South Urals, within which the ancient sites of the Sintashta culture of the Middle Bronze Age were found (about 2000 BC, i.e. comparable in time to the ancient Egyptian the pyramids).



**Fig. 1.** Map showing “Country of cities” (territories with a diameter of approximately 350-500 km around the number 2000 in the center) and the chariot spread: 2000–1300 BC.

A comparison of the data we obtained on both hypotheses with the data of independent researchers (see e.g. [6-9]) allows us to draw an unambiguous conclusion about the preference of the Kurgan hypothesis.

We should note that the period of existence of one or two languages is approximately 1000–1500 years (the time interval from 4500 to 3500 years ago), which is quite consistent with the data [2].

## 3 Dynamic nonlinear model of distribution and changes of linguistic information: PC communities

### 3.1. Mathematical model and analysis

In this article, we use the possibility of applying the results obtained for IE communities as a certain standard (reference community) in the analysis of other language communities, for example, Asian language communities [10, 16], where a similar approach to the analysis of the dissemination of language information is possible.

Indeed, Chinese and PIE languages refer to active languages, where there is a division of nouns into “active” and “inactive”, verbs into “active” and “stative”, and adjectives are usually absent. Note that the “stative” verbs (meaning literally “be sick”, “be funny”, etc.) do not express action and do not imply duration, but give only a description of the state. We emphasize that in

modern Indo-European languages, adjectives are usually used instead of similar verbs.

We assume that the similarity of modern Chinese and PIE languages in terms of the relationship between the verb and the noun (they are both active languages) can be used as a justification for the possibility of using our method to analyze the Sino-Tibetan language family of languages (see e.g. [4, 5, 7, 10]).

In the computer modeling according Eq. (1), the ratio of the coefficients was chosen based on the ratio of average speeds and (in km/h) different types of pedestrians  $v_1$  and  $v_2$ :  $\lambda_1 : \lambda_2 \approx v_1 : v_2 \approx 3 : 3 \approx 1$ . So, we took into account the fact that the domestication of the horse and the use of horse-drawn carts occurred in the PST linguistic community much later than in the PIE linguistic society (see about the domestication of the horse above and look at the fig. 1) [1, 2, 6, 7].

Another important condition for the task we consider is the linguistic policy of China on linguistic unification, i.e. the creation of one national language *Putonghua* (Pǔtōnghuà 普通话) in the near future [16, 17]. According the “Law of the People’s Republic of China on the Standard Spoken and Written Chinese Language”:

“**Article 9** Putonghua and the standardized Chinese characters shall be used by State organs as the official language, except where otherwise provided for in laws.

**Article 10** Putonghua and the standardized Chinese characters shall be used as the basic language in education and teaching in schools and other institutions of education, except where otherwise provided for in laws.

Putonghua and the standardized Chinese characters shall be taught in schools and other institutions of education by means of the Chinese course. The Chinese textbooks used shall be in conformity with the norms of the standard spoken and written Chinese language.” (see Chapter II in Ref. [17]).

We assume that the time of the beginning of the separation of the great Sino-Tibetan language family, i.e. in fact, the time of the “appearance” of the hypothetical Proto-Chinese language occurred approximately 6500 years ago, i.e.  $6500 : 500 = 13$  generations back (500 years is the time of the divergence of two related languages) [4, 5].

### 3.2. Results for PC communities

Figure 2 gives one example of the graphical illustration of the results of computer simulation in the case of PC linguistic communities.

The following parameters were used during computer modeling (see Fig. 2):  $\lambda_1 = 0.6$  and  $\lambda_2 = 0.59$ , initial value  $x_0 = 10^{-4}$ . Equation (1) was taken during computer simulation in the following form:

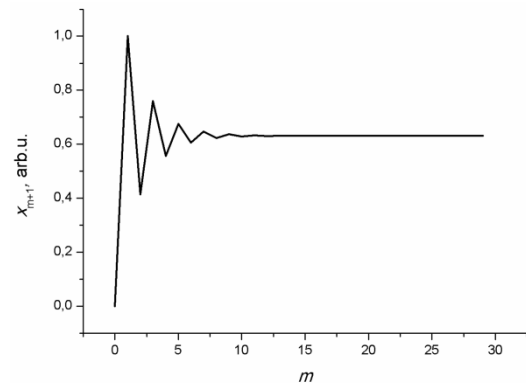
$$x_{m+1} = \lambda_1 x_m (1 - x_m) + \lambda_2 x_m (1 - x_m)^2. \quad (2)$$

As can be seen from the figure 2, iterated values of  $x_{m+1}$ , i.e. information  $I_m \rightarrow const$ , so here the dynamic mode is stationary or has a period equal to unity, – a

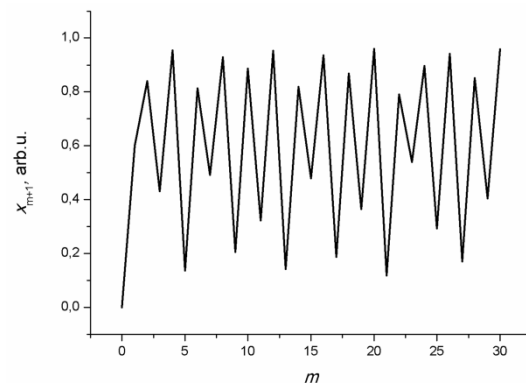
cycle  $S^1$  is observed. Consequently, the appearance of one language can be achieved, at about 13-16 generations (i.e. to the years 2019-3500), but unification, most likely, will be achieved only by 16-18 generations, i.e. not earlier than 3500.

The results obtained during computer modeling (see Fig. 2) allow us to make the following assumption: about 5500-6000 years ago, after 1-2 generations, i.e. fast enough, since the beginning of the existence of the linguistic PST community, namely after 500-1000 years after co-development of the Proto-Sino-Tibetan language family, two main linguistic populations could emerge in the considered linguistic PST community, which are now characterized as a division of the Sino-Tibetan group of languages into Tibeto-Burmese (or Tibeto-Burman) and Proto-Chinese.

Figure 3 gives a graphical illustration of the results of computer simulation in the case of PC linguistic communities when chaos begins to arise. The system sets the cycle type  $S^{2^p}$ ,  $p \rightarrow \infty$ . The following parameters were used in Eq. (2) during computer modeling (see Fig. 3):  $\lambda_1 = 3$  and  $\lambda_2 = 0.6$ , initial value  $x_0 = 10^{-4}$ .



**Fig. 2.** Plot of the dependence of the function  $x_{m+1}$  on  $m$  (number of generations), that characterizes the kind of cycles  $S$  occurring in the dynamic linguistic system: in the PC linguistic community, chaos begins to arise.

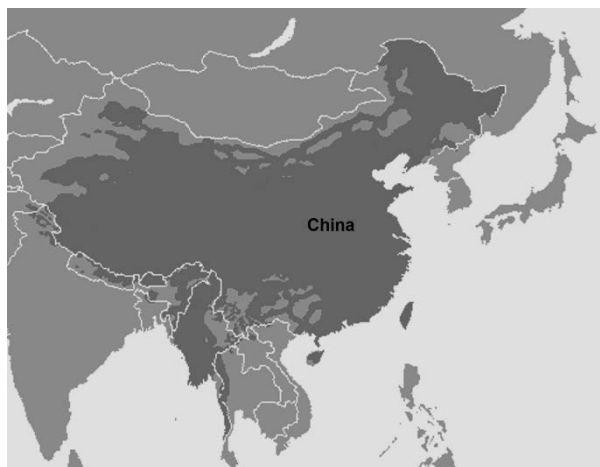


**Fig. 3.** Plot of the dependence of the function  $x_{m+1}$  on  $m$  (number of generations), that characterizes the kind of cycles  $S$  occurring in the dynamic linguistic system: in the PC linguistic community, chaos begins to arise.



Our model allows us to investigate the phenomenon of the spread of language information and obtain similar results with some reasonable variation of the initial data, in particular, the moment when the separation of languages in the Proto-Sino-Tibetan language family occurred [4, 5, 10-13, 18, 19].

Figure 4 shows the spread of the Sino-Tibetan languages in modern days.



**Fig. 4.** A map of the Sino-Tibetan languages (dark gray areas).

A more detailed analysis of the situation in the considered linguistic PST community and other aspects of the problem, for example, possibility of creation of one national Chinese language in the near future, will be considered in our subsequent works.

## 4 Discussion

In our earlier paper [2] we found the probable place of the location of the ancestral homeland (so-called *Urheimat*) of the peoples that were carriers of the Proto-Indo-European language. In our opinion a prehistorical PIE homeland was located in the area marked by “1” on the Fig. 1. In subsequent times, their carriers spread to the west and east, north and south.

According to archeology, speakers of PIE languages (mostly carriers of the Y-DNA R1a haplogroup) came into contact with speakers of PST languages most likely in the west of modern China (e.g. in the “*Tarim Basin*”, which is located in the northwest part of modern China, not far from the modern day Kazakhstan, Kyrgyzstan, Tajikistan, Pakistan, Afghanistan, and India) [6, 8-10, 20-27].

We should note that the Northern Silk Road on one route bypassed the “*Tarim Basin*” north of the Tian Shan Mountains and traversed it on three oases-dependent routes: one north of the Taklamakan Desert, one south, and a middle one connecting both through the Lop Nor region (see e.g. [20-27]).

The earliest inhabitants of the “*Tarim Basin*” may be the *Tocharians* whose languages are the easternmost group of Indo-European languages. Caucasoid mummies have been found in various locations in the Tarim Basin such as Loulan, the Xiaohe Tomb complex, and Qäwrighul [20]. These mummies have been suggested to

be of Tocharian origin, and these people may have inhabited the region since at least 1800 BC [20-22].

The existence of the Afanasievo culture near Altai around 3000 BC could also provide an explanation for the mysterious presence of one of the oldest Indo-European languages, Tocharian in the Tarim basin in China [22, 23, 24].

Another people in the region besides Tocharian are the Indo-Iranian Saka people who spoke various Eastern Iranian Khotanese Scythian or Saka dialects, i.e. IE languages [22, 23].

## 5 Results

The new method of research of rather different linguistic communities as dynamic dissipative systems is described: Proto-Indo-European and Proto-Chinese language communities.

We assume that a certain similarity of Chinese and PIE languages in terms of the relationship between the verb and the noun can be used as a justification for the possibility of using our method to analyze the Sino-Tibetan language family of languages. Besides, we also take into account that speakers of PIE languages came into contact with speakers of PST languages in the west of modern China.

Another basis that we used in the analysis is the time of the appearance of wheeled carts with domesticated horses in the prehistoric China.

We analyzed the spread of language information in the Proto-Chinese linguistic community. The main results of computer simulation are presented. The possibility of creating a single state language (Pütōnghuà 普通话) in modern China is briefly analyzed. It is concluded that the language unification will be achieved not earlier than the 3000s.

We are planning to give more comprehensive study of these problems in our further researches. Indeed, our approach requires taking into consideration many heterogeneous factors, which are not always possible to take into account explicitly in our mathematical models.

Such an interdisciplinary approach is undoubtedly promising and can give good quality results that cannot be obtained in other areas of knowledge. At the same time, data from other subject areas (archeology, history, linguistics, mathematical modeling in linguistics) should certainly be taken into account to verify the results obtained where it is possible.

It must be emphasized that this approach is especially promising primarily for a qualitative analysis of the behavior of such social-linguistic systems; therefore, all the numerical estimates obtained must be considered precisely as some preliminary evaluations.

Acknowledgements. The publication was supported in part by the “RUDN University Program 5-100”.

## References

1. M. Egorova, A. Egorov, T. Solovieva, L. Stezhenskaya, *Proc. of the 4th International*

- Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM'2017*, 28–31 March, 2017, Extended Scientific Sessions Vienna, Austria: STEF92 Technology Ltd., Sofia, Book 3, **1**, 155 (2017).
2. M.A. Egorova, A.A. Egorov, *Automatic Document. and Math. Linguistics*, **53** (3), 127 (2019).
  3. A.A. Egorov, M.A. Egorova, *Proc. of the XXI All-Russian Conf. "Theoretical Foundations of Designing Numerical Algorithms and Solving Problems of Mathematical Physics"*, September 5–11, 2016, Novorossiysk, Abrau-Durso, Moscow, 82 (2016).
  4. *Atlas of world languages. The origin and development of languages around the world* (Lik press, Moscow, 1988).
  5. S.E. Yakhontov, *Ancient Chinese language* (Nauka, Moscow, 1965).
  6. A. Pereltsvaig, M.W. Lewis, *The Indo-European Controversy: Facts and Fallacies in Historical Linguistics* (Cambridge University Press, Cambridge, 2015).
  7. D.W. Anthony, *The Horse, the Wheel, and Language: How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World* (Princeton University Press, Princeton, 2007).
  8. W. Haak, I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, S. Nordenfelt, E. Harney, K. Stewardson, Q. Fu, A. Mittnik, E. Banffy, Ch. Economou, M. Francken, S. Friederich, R.G. Pena, F. Hallgren, V. Khartanovich, A. Khokhlov, M. Kunst, P. Kuznetsov, H. Meller, O. Mochalov, V. Moiseyev, N. Nicklisch, S.L. Pichler, R. Risch, M.A.R. Guerra, C. Roth, A. Szecsenyi-Nagy, J. Wahl, M. Meyer, J. Krause, D. Brown, D. Anthony, A. Cooper, K.W. Alt, D. Reich, *Nature*, **522**, 207 (2015).
  9. W. Chang, C. Cathcart, D. Hall, A. Garrett, *Language*, **91** (1), 194 (2015).
  10. *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics*. Eds. Blench R., Sagart L., Sanchez-Mazas A. (Routledge Curzon, London, 2005).
  11. G. Thurgood, R.J. LaPolla. *The Sino-Tibetan Languages* (Routledge, London, 2003)
  12. R. Shafer, *Classification of the Sino-Tibetan Languages*, *Word*, **11** (1), 94 (2015).
  13. L. Sagart, G. Jacques, Y. Lai, R.J. Ryder, V. Thouzeau, S. J. Greenhill, J.-M. List, *PNAS*, **116** (21) 10317 (2019).
  14. A. Kornai, *Mathematical Linguistics* (Springer, London, 2008).
  15. W. Labov, *Language*, **83**, 344 (2007).
  16. M.A. Egorova, *Language Policy in a Globalizing World* (RUDN, Moscow, 2017, 44).
  17. Law of the People's Republic of China on the Standard Spoken and Written Chinese Language (Order of the President No. 37). URL: [http://www.gov.cn/english/laws/2005-09/19/content\\_64906.htm](http://www.gov.cn/english/laws/2005-09/19/content_64906.htm)
  18. Chinese languages. *Encyclopædia Britannica*. URL: <https://www.britannica.com/topic/Chinese-languages>
  19. Tibeto-Burman languages. *Encyclopædia Britannica*. URL: <https://www.britannica.com/topic/Sino-Tibetan-languages/Language-affiliations>
  20. J.P. Mallory, V.H. Mair, *The Tarim Mummies: Ancient China and the Mystery of the Earliest People from the West* (Thames & Hudson, N.Y., 2000).
  21. C. Li, H. Li, Y. Cui, C. Xie, D. Cai, W. Li, V.H. Mair, Z. Xu, Q. Zhang, I. Abuduresule, L. Jin, H. Zhu, H. Zhou, *BMC Biology*, **8** (15), 1 (2010).
  22. M.E. Allentoft, M. Sikora, K.-G. Sjogren, S. Rasmussen, M. Rasmussen, J. Stenderup, P.B. Damgaard, H. Schroeder, T. Ahlstrom, L. Vinner, A.-S. Malaspinas, A. Margaryan, T. Higham, D. Chivall, N. Lynnerup, L. Harvig, J. Baron, P.D. Casa, P. Dabrowski, P.R. Duffy, A.V. Ebel, A. Epimakhov, K. Frei, M. Furmanek, T. Gralak, A. Gromov, S. Gronkiewicz, G. Grupe, T. Hajdu, R. Jarysz, V. Khartanovich, A. Khokhlov, V. Kiss, J. Kolar, A. Kriiska, I. Lasak, C. Longhi, G. McGlynn, A. Merkevicus, I. Merkyte, M. Metspalu, R. Mkrtychyan, V. Moiseyev, L. Paja, G. Palfi, D. Pokutta, L. Pospieszny, T.D. Price, L. Saag, M. Sablin, N. Shishlina, V. Smrcka, V.I. Soenov, V. Szeverenyi, G. Toth, S.V. Trifanova, L. Varul, M. Vicze, L. Yepiskoposyan, V. Zhitenev, L. Orlando, T. Sicheritz-Ponten, S. Brunak, R. Nielsen, K. Kristiansen, E. Willerslev, *Nature*, **522**, 167 (2015).
  23. E. Heyer, P. Balaesque, M.A. Jobling, L. Quintana-Murci, R. Chaix, L. Segurel, A. Aldashev, T. Hegay, *BMC Genetics*, **10** (49), 1 (2009).
  24. J. Novembre, *Nature*, **522**, 164 (2015).
  25. R. Qamar, Q. Ayub, A. Mohyuddin, A. Helgason, K. Mazhar, A. Mansoor, T. Zerjal, C. Tyler-Smith, S.Q. Mehdi, *Am. J. Hum. Genet.*, **70**, 1107 (2002).
  26. Horolma Pamjav, Tibor Feher, Endre Nemeth, Zsolt Padar, *Am. J. Phys. Anthropol.*, **149** (4), 611-615 (2012).
  27. P.A. Underhill, N.M. Myres, S. Rootsi, M. Metspalu, M.A. Zhivotovsky, R.J. King, J.D. Cristofaro, H. Sahakyan, D.M. Behar, A. Kushniarevich, J. Sarac, T. Saric, P. Rudan, A.K. Pathak, G. Chaubey, V. Grugni, O. Semino, L. Yepiskoposyan, A. Bahmanimehr, S. Farjadian, O. Balanovsky, E.K. Khusnutdinova, R.J. Herrera, J. Chiaroni, C.D. Bustamante, S.R. Quake, T. Kivisild, R. Villems, *Eur. J. Hum. Genet.*, **18**, 479 (2010).