中國語言學集刊

BULLETIN OF CHINESE LINGUISTICS

紀念李方桂先生中國語言學研究學會
Li Fang–Kuei Society for Chinese Linguistics

香港科技大學中國語言學研究中心
Center for Chinese Linguistics, The Hong Kong University Of Science and Technology

中國語言學集刊

（第二卷 第一期）

BULLETIN

OF

CHINESE LINGUISTICS

中國語言學集刊

(第二卷第一期)

Editorial Board

目 錄
CONTENTS

Articles

Interview Transcription

Appendices

Languages and Genes in China and in East Asia

Alain Peyraube

CNRS & EHESS (Paris, France)

This article poses two main questions: can the history of genes help us understand better what the Chinese linguistic situation was some 5,000 years B.P., not to mention the population distribution in China? Consequently can the history of genes help us in grouping the languages of China and East Asia into families and macro-families?

Languages and genes have two different histories and two different types of evolution – one being natural, the other one largely cultural – with different mechanisms of origin and reproduction. Nonetheless, there are indeed many clear analogies in the mechanisms of transmission: mutation, natural selection, migration, and chance. These have lead population geneticists and linguists to look for any congruence in genetic and linguistic evolution, in order to correlate genetic and linguistic distance.

In light of these congruences, but also of non-correlations existing between the genetic classification of populations and the classification of languages, the different hypotheses concerning the traditional grouping of languages (Sino-Tibetan, Austronesian, Austro-Asiatic, Tai-Kadai , Miao-Yao or Hmong-Mjen, Altaic), as well as the new groupings in macro-families (Austric, Austro-Tai, Sino-Tibetan-Austronesian, Sino-Indo-European, Sino-Caucasian, Proto-East-Asian, etc.) will be discussed.

It will be concluded that while we have various hypotheses, we are not sure of anything. The considerable accumulation of data in population genetics has rendered the landscape much less simple, all the more so since the theoretical models of evolution necessary to interpret the genetic data in historical context are still being refined.

Key-words: languages, genes, Sino-Tibetan, Austronesian, Austro-Asiatic, Tai-Kadai, Hmong-Mjen, Altaic, Austric, Austro-Tai, Sino-Tibetan-Austronesian, Sino-Indo-European, Sino-Caucasian, Proto-East-Asian.

Shang China was inhabited by Chinese populations speaking Sinitic languages. It was also inhabited by non-Chinese populations. We do not know however who

precisely these non-Chinese populations were, while the evidence for the origins of the Chinese language is still also scanty and controversial.[1]

Can the history of genes help us in understanding better what the Chinese linguistic situation was some 5,000 years B.P., not to mention the population distribution?

Languages and genes do not have the same history, but as Darwin already noticed in the second half of the 19th century, there is an "interesting parallelism between the development of species and the development of languages."

If we have two different types of history, two different types of evolution – one being natural, the other one largely cultural – with different mechanisms of origin and reproduction, there are indeed many clear analogies in the mechanisms of transmission: mutation, natural selection, migration, and chance (called "drift" in genetics) (Piazza 1995). It is not surprising then that the similarities of these mechanisms of transmission have lead scholars to look for the congruence of genetic and linguistic evolution.

Correlations between genetic distance and linguistic distance

For almost two decades, population geneticists – whose main aim is to put forward a reconstruction of the history of modern humans since their origins, through a systematic study of the genetic variability of our species – have been very active in correlating genetic distance, a central notion in genetic population, and linguistic distance. As a result, the amount of genetic data has increased considerably as the development of bio-chemical and molecular techniques progressed. Some of the most important studies in this respect – out of more than one hundred – have been the following ones: Greenberg *et al.* (1986) on American populations; Excoffier *et al.* (1987) on sub-Saharan Africa; Sokal *et al.* (1988), Barbujani *et al.* (1990) on Europe,

---

[1] Wang (1999) gives the following scenario for Ancient China. The Yi to the East may have been speakers of Austronesian and Austro-asiatic languages. The Rong to the west were probably mostly speakers of Tibeto-Burman. In the north, there were the Di who spoke various languages of the Altaic family. To the south, the name Man covers the most diverse collection of peoples, speaking languages which are now grouped under Kadai, Miao-Yao, and others. As for the Yue, the term apparently was used in late Shang times to refer to peoples in Northwest China, before being used from the late Warring States period through the first centuries of the Common Era, to refer to a wide spectrum of peoples in southern China, especially under the name Baiyue. Wang nevertheless adds: "These identifications of ancient labels with modern peoples are no more than speculations, of course, based on the extrapolations from contemporary population distributions."

especially showing that language affiliation of European populations plays a major role in maintaining and probably causing genetic differences; Poloni *et al.* (1997) on various populations of Africa and Europe; and significantly the now classical study by Cavalli-Sforza, Piazza, Menozzi and Mountain (1988), who have built a tree of 42 world populations representing the world's aborigines.[2]

The genetic information for this last work came from a very large collection of gene frequencies for "classical" (i.e. non-DNA) polymorphisms. 120 alleles were studied. They showed that linguistic superfamilies show remarkable correspondence with most of the clusters, indicating what they called a "considerable parallelism between genetic and linguistic evolution (p. 6002)."

In most of these studies, if not all, the linguistic hypotheses adopted by the geneticists have been the ones for the macro-families advocated by Greenberg and Ruhlen and/or by the Russian "unificationists".[3] This is most probably due to the fact that geneticists need to work on large geographical areas for reasons of sampling and because the local variations for gene frequencies are lower.

Many other later works have invalidated the existence of indisputable correlations between a genetic classification of populations and a classification of languages. Cavalli-Sforza himself recognized that the correlation between genetic distance and linguistic distance is extremely weak. The "considerable parallelism" he was speaking in 1988 became a "certain non random association between linguistic families and human genetic history" (Cavalli-Sforza *et al.* 1992: 5623). As a good illustration of non- correlations, one may cite the two well-known cases in the Caucasus. The first one concerns the Armenians and the Azeris: they are genetically very close to each other, but they speak entirely different languages (Armenian being Indo-European and Azeri belonging to the Altaic family). The second one deals with Chechen and Ingush who speak very closely related languages belonging to the Northeast Caucasian branch of North-Caucasian while being quite genetically distinct (Nasidze *et al.* 2001, Nasidze *et al.* 2003).

If the first studies of genetic variability were done indirectly by analyzing the polymorphism at the phenotypical level (i.e. at the level of the products of genes),

---

[2] See also Cavalli-Sforza *et al.* (1994), Cavalli-Sforza (1996), Mountain *et al.* (1992) for China.

[3] Ruhlen (1997) proposed that all languages of the world belong to the following 12 macro-families : Khoisan, Nigero-Kordofanian, Nilo-Saharan, Afro-Asiatic (in Africa), Austric, Indo-Pacific, Australian (in Southeast Asia and Oceania), Amerind (in the Americas), Dravidian, Kartvelian, Eurasiatic, and Dene-Caucasian (in Eurasia).

starting in the 1980s, the analysis of the polymorphism was mainly realized at the DNA level, more complex but by a direct manner.[4]

Mitochondrial DNA (mtDNA), transmitted through the mothers, and the Y chromosome, strictly paternally inherited, have thus proved to be potent tools in our understanding of human evolution, owing to characteristics such as high copy number, apparent lack of recombination, high substitution rate and maternal or paternal mode of inheritance. Most of the time, the analyses of mtDNA and the Y chromosome have given similar results, but in some cases men and women seem to have contributed in a different manner to the constitution of human populations. This is, for instance, the case for the Dama population in Africa speaking a language of the Khoisan macro-family. The mt genes are shared in the population but there are also strong affinities, concerning the Y chromosome, with populations of the Niger-Congo linguistic group (belonging to the Nigero-Kordofanian macro-family).[5]

Hence, for the moment, we have to recognize that while we have various hypotheses, we are not sure of anything. The considerable accumulation of data in population genetics has rendered the landscape less simple, all the more so since the theoretical models of evolution necessary to interpret the genetic data in a historical context are still being refined.

The situation in East and Southeast Asia

What is the situation for China, for East Asia, and for Southeast Asia? It is probably even more confused.

From a linguistic point of view, the classical way is still to group the languages of the region into the six following language families:

- Sino-Tibetan (ST) dated ca. 4500 BCE, with its two sub-families, namely Tibeto-Burman (TB) and the Sinitic languages. Wang (1996) estimates that the Chinese languages diverged from the Tibeto-Burman languages around 6,000 years BP. This is also the date that Chang (1986: 234) identifies as the formation of what he calls "the initial China", when various Neolithic sites all over China began to emerge significantly.
- Austronesian (AN), ca. 4000 BCE

---

[4] At every generation level, there are mistakes in copying the DNA transmitted from parents to children. These mistakes are called polymorphisms. Those errors or polymorphisms are the differences between populations we can investigate today to reconstruct the past of our species.

[5] See Dupanloup *et al.* (2002).

- Austro-Asiatic (AA), ca. 5.000-4.000 BCE
- Tai-Kadai (TK), ca. 3000-2000 BCE
- Miao-Yao (or Hmong-Mien), ca. 500 BCE
- Altaic (AT)

There is now a general consensus, though not generally accepted, that the Sinitic languages (and dialects) are genetically related to the Tibeto-Burman languages, constituting probably one branch of the Sino-Tibetan family. Beyond this relationship, however, there is little agreement. And different possible scenarios for macro-families grouping these different families are under discussion today, most of them being revivals of earlier proposals. Six main hypotheses have been put forward during the last decades.

- Austric, a combination of AA and AN, proposed by Schmidt (1905) and advocated today by Reid (1994). Miao-Yao and Tai-Kadai are considered as belonging to Austric by some (Ruhlen 1997).
- Austro-Tai, a grouping of Austronesian and Tai-Kadai, proposed already by Schlegel (1901), then by Benedict (1942, 1975), and supported today by Ostapirat (2005). Miao-Yao is also usually included under Austro-Tai.
- Sino-Austronesian, and then Sino-Tibetan-Austronesian, grouping Sino-Tibetan and Austronesian, already proposed by Conrady (1916) and strongly advocated today by Sagart (1994).
- Sino-Indo-European, grouping Sino-Tibetan and Indo-European, a hypothesis put forward by Pulleyblank (1995, 1996) who noticed traces of shared phonological and morphological correspondences at a very deep level, hard to explain except as evidence of a common origin. If they are indeed related to one another, the time at which the two languages formed a single community must be very remote, prior to Proto-Indo-European.
- Sino-Caucasian, grouping Sino-Tibetan, North-Caucasian and Yenisseian, proposed by Starostin (1989, 1995), a theory that can be traced back to Donner (1916). Na-Dene languages have been added to Sino-Caucasian, which is why the family is called Dene-Caucasian most of the time.
- Proto-East-Asian; a macro-macro family composed of Sino -Tibetan- Austronesian and Austric, including Miao-Yao (in Austric) and Tai-Kadai (in Austronesian). This proposition has been made by Starosta (2005). Proto-East-Asian (PEA) could have been spoken in Central China, around the Han River and the Yellow River, ca 9,000 years BP or even earlier.

Why such a divergence of opinions? This is because it is extremely difficult to sort out the effects due to contact between the different languages once spoken in this vast region from the effects of internal evolution, that is, due to horizontal transmission as opposed to vertical transmission. It is therefore difficult to favour one scenario rather than another. The tenants of each of these macro-families all have some arguments in support of their positions. [6]

Can the study of genes help us in choosing one or the other hypothesis?

One of the first important genetic analyses of Chinese populations studied the distribution of the immoglobuline allotypes Gm and Km ("classical" markers) in 74 populations of China. See Zhao and Lee (1989), and most importantly Zhao *et al.* (1991). The two papers showed that a major distinction should be made between Northern and Southern Han people due to genetic differences. A division could be in fact made along the latitude of 30 degrees north.

Southern Chinese show greater affinity with the peoples of Southeast Asia, at least according to the genes studied. Within the major north-south division, genetic studies have also shown that the so-called Han peoples do not form groups among themselves, but rather group with the minority peoples who co-inhabit the region. In other words, the Han unity is a cultural rather than a biological one.

In addition, a comparison with 33 other populations outside of China showed that the Northern Han Chinese could belong to a group comprising the A thapascans (who speak the Na-Dene languages of North America, from the Dene-Caucasian macro-family), the Eskimos, the Japanese, the Koreans and the Mongolians (who

---

[6] For instance, Starostin (1995), who advocates the Sino-Caucasian hypothesis, after having compared the reconstructions of the 35 basic words of the Jakhontov list across different families, concludes that there is a greater affinity between Chinese and Proto-Tibeto-Burman, Proto-North-Caucasian and Proto-Yenisseian, as shown by the following table, adapted in Wang (1996):

[OC = Old Chinese; PTB = Proto-Tibeto-Burman; PNC = Proto-North-Caucasian;
PY = Proto-Yenisseian; PIE = Proto-Indo-European; PAN = Proto-Austronesian]

	OC	PTB	PNC	PY	PIE	PAN
Old Chinese	100	74	43	34	23	14
Proto-Tibeto-Burman		100	51	40	14	11
Proto-Noth-Caucasian			100	57	17	11
Proto-Yenisseian				100	11	11
Proto-Indo-European					100	14
Proto-Austronesian						100

The figures are percentages.

speak Eskimo-Aleut or Altaic languages belonging to the Eurasiatic macro-family). In contrast to this, the Southern Han would be closer to the Thais (Tai-Kadai languages) and to the Vietnamese (Austro-Asiatic languages), or even to the Indonesians and the Filipinos (Austronesian languages), i.e. to people speaking languages subsumed under the Austric macro-family.

Chu *et al.* (1998) also studied genetic profiles of 28 populations sampled in China and supported the distinction between southern and northern populations, the latter being biphyletic. Furthermore, the phylogeny suggested that it is more likely that ancestors of the populations currently residing in East Asia entered from Southern China. Finally, genetic evidence does not support an independent origin of *Homo sapiens* in China.

Su *et al.* (1999) collected DNA samples from members of 21 Chinese ethnic-minority populations (the project was undertaken as a part of the big Chinese Human Genome Diversity project) with Han Chinese samples collected from persons living in 22 provincial areas, using Y-chromosome biallelic markers. They confirmed the northern-southern division while similarly concluding that northern populations are derived from the southern populations. Their data indicate that southern populations in Eastern Asia are much more polymorphic than northern populations, which have only a subset of southern haplotypes. This pattern probably indicates that the first settlement of modern humans in Eastern Asia occurred in mainland Southeast Asia during the last Ice Age, i.e. 50,000 to 100,000 years ago, from where they expanded northwards to other parts of East Asia, Northern China and even Siberia.

Ding *et al.* (2000), however, after having studied three human genetic marker systems, find support for neither a major north-south division in these markers nor a southern origin of Northeast Asian populations; rather, the marker patterns suggest simple isolation by distance. In the map of diversity in mtDNA, they found that some southern populations such as the Dai are much more similar to other northern populations than they are to other southerners, such as the Vietnamese. The putative northern and southern clusters appear to blend across a cline; there is no abrupt change. They conclude that the lack of patterning in East Asia suggests that many of the anthropological trends previously held to define pervasive regional distinctions are strictly cultural phenomena with no implications for genetic differentiation.

Yao *et al.* (2002a) studied and compared the mtDNA variation in Han Chinese from several provinces in China. The comparison revealed an obvious geographic differentiation in the Han Chinese, shown by the haplogroup-frequency profiles. The south-to-north cline observed in the frequency of some haplogroups (F1, B, and D4) is quite similar to the distributions of immunoglobulin Gm allotypes in Chinese populations (Zhao and Lee 1989). However, the grouping of different Han populations

into just "Southern Chinese" and "Northern Chinese" (Su *et al* 1999) or the use of one or two regional populations to stand for all Han Chinese does not appropriately reflect the genetic structure of the Han. Intriguingly, despite the numerous historically recorded migrations and substantial gene flow across China from the Bronze Age to the present time (Ge *et al.* 1997), differences between geographic regions have been maintained. The regional difference is more pronounced in south and southwest China: southern and southwestern populations show a more diverse pattern than the populations from central, east and northeast China. Yao *et al* (2002a) conclude that an initial pioneer colonization of China ca. 60,000 years ago from Southeast Asia is a conceivable scenario (as proposed by Su *et al* 1999) but still leaves much room for speculation about the population dynamics during the long period between then and the Late Glacial Maximum (some 50,000 years ago). The contrast between the northern and southern genetic pools might have its roots in this period. Subsequent migration events may have somewhat blurred this early distinction, with the genetic pools of central China possessing mtDNA features of both the northern and the southern pools.

Xue *et al.* (2003) studied the distribution of MYS2 polymorphism in 26 Chinese populations. The results show that there are indeed differences in genetic structure between southern and northern populations. The geographical distinction between south and north is thereby confirmed. A distinction between western and eastern populations in Northern China is also suggested.

He *et al.* (2003) using molecular biotechnology methods have described the polymorphism analysis for the mtDNA D-loop high variable region in ancient human bones (coming from the tumulus in Wupu, Hami, Xinjiang). The results show that there were Asians and Europeans in these tumuli and that they might have coexisted in that district of Hami, Xinjiang, 3200 years ago.

Quintana-Murci *et al.* (2001) state that the geographical distribution, observed clines, and estimated ages of HG-9 and HG-33 chromosomes in southwestern Asia all support a model of demic diffusion of early farmers from southwestern Iran – and nomads from western and central Asia – into India, bringing the spread of genes and culture (including language) to southwestern Asia. Although alternative, more complex explanations are possible, the analysis of the modern male-specific gene pools in these populations suggests that major demographic events, involving migration and admixture, accompanied these important historical and linguistic events.

For Southeast and East Asia, a new study by Poloni *et al.* (2005) on the GM system and the Rhesus factor shows that the Austro-Asiatic group is the most highly

differentiated by far, followed by Sino-Tibetan.[7] In comparison to this, Tai-Kadai and Austronesian are the least differentiated. For the GM, there is a strong genetic differentiation between Sino-Tibetan on the one hand and all the others, a divergence which is above all a fact of northern populations (Northern Han Chinese, Tibetans). The southern Sino-Tibetan (Southern Han Chinese, Tibeto-Burman of India, Burma, Thailand) show genetic similarities with populations from other Southeast Asian groups. For all that, the levels of differentiation for these southern groups are less than those between the populations which compose them.

Another interesting finding is that the North Caucasian populations are genetically highly differentiated from all the others.

The analysis of HLA genetic variability supports a grouping of Altaic populations (Mongolian and Manchurian) as well as Japanese, Korean, Austro-Asiatic and Tai-Kadai, as opposed to the more extreme differentiation for the Austronesian populations. With regard to variation within each group, the most diversified is Austronesian (for GM and Rhesus, Austronesian is the least differentiated group, see above). The study of this system does not support the hypothesis concerning a common origin for all the southern populations, nor that concerning the northern populations derived from those of the south. Another finding, finally, is that there is no genetic differentiation between Austronesian and Kadai.

Furthermore, the GM and HLA-DRB1 polymorphisms, those of the Y chromosome, and, to a lesser extent, those of the mtDNA match up with the genetic closeness between the Altaics, the Tibetans and the Northern Han. These populations could have shared a common Altaic origin while certain of the populations (Tibetans, Northern Chinese) might have replaced their own languages after prolonged contact with the original speakers of Sino-Tibetan.

In contrast with the RH, GM and HLA-DRB1 systems, genetic distance calculated on the basis of HVS1 (a segment inside the D-loop control region) polymorphism is not correlated with geographical distance.

As a matter of fact, new approaches have come to the fore, which focus on research into spatial discontinuities linked with genetic variation in populations. It needs to be verified, however, whether or not such discontinuities have their origin in the ones created by the distribution of languages. Nonetheless, a great deal of emphasis has been given to the relationship between spatial organization and geographical distance. Dupanloup et al. (2002) have confirmed that degrees of genetic

---

[7] To have information on the history of populations, Poloni et al. claim that it is necessary to study several genetic systems, given that the transmission of genes from one generation to the next is subject to stochastic processes: the GM and Rhesus factor systems.

differentiation between different population groups are indeed closely related to their geographical distribution. Yao *et al* (2002 b: 63), however, conclude that "in general, linguistic and geographical classifications of the populations did not agree well with classification by MtDNA variation."

Rosser *et al*. (2000) have studied the polymorphism of the Y chromosome in Europe. Using a sample of 47 populations speaking 37 languages, the researchers noted that genetic diversity is significantly correlated with geographical distance, but not with linguistic diversity. The linguistic effect of such differentiation is thus only secondary to the geographical one.

The above sketch assuredly shows a blurred picture. The two following hypotheses are nevertheless probable: (i) There is a distinction between northern and southern Chinese populations and the Han Chinese tend to group with minority peoples who co-inhabit the regions where they live; (ii) The northern Han population is likely derived from the southern Han population.

As for the correlations between linguistic distance and genetic distance, the genetic analyses are not sufficiently clear to validate one or the other hypothesis concerning possible macro-families in East and Southeast Asia.

There are also some serious methodological and theoretical problems briefly discussed below.

Some remaining issues

The first issue is: why should we expect genetic boundaries and linguistic ones to necessarily coincide? It seems to be the case that any parallelism between language evolution and genes can only be convincing when co-evolution has been demonstrated. If the evolutionary history of genes and linguistic typology coincide a few times, this does not constitute any proof of congruence. In other words, the presence of a correlation between genes and languages can only be used as evidence if a convergent evolutionary history can be shown to exist between them. A single correlation does not necessarily establish any direct causal link, nor does it lead to congruence in the philogenetic sense of this word from a methodological viewpoint.

The second issue concerns linguistic reconstruction. If the time depth in question is more than 10 millennia ago, we cannot even identify the descendants of a prehistoric language as belonging to the same family. Ringe (1999) claims that one universal type of language change is the replacement of old linguistic material by completely different words and affixes. As it happens, even the most basic words and affixes get replaced by totally new ones distressingly fast, so that after 10,000 years (at