

HUMAN DIVERSITY AND LANGUAGE DIVERSITY

WILLIAM S.-Y. WANG

Department of Electronic Engineering, City University of Hong Kong

Since language is the defining trait of our species, human evolution and linguistic evolution are obviously closely intertwined. Recent studies in genetics suggest that anatomically modern humans emerged at a very late date, perhaps 50 kys (Bertrapetit 2000; Thompson 2000). This dating is consistent with the onset of an unprecedented degree of cultural innovations, in both quality and quantity, as revealed in the archaeological record (Klein 1999). We share the belief with many students of human prehistory that the evolution of anatomically modern humans, the emergence of language, and the burst of cultural innovations, including extensive cave art and sailing across broad expanses of water are events which are all closely linked to each other.

Culturally, there have been several major transitions separating us from our prehistoric ancestors—such as the use of fire, the invention of tools, the advent of agriculture, etc. Similarly, there must have been major transitions which led from the primitive growls and howls of our ancestors to the intricate languages we have today. We cannot recover language evolution in the very distant past in ways comparable to those of the archaeologist, since the earliest ‘material remains’ of language, i.e., ancient texts, date back no farther than several millennia. However, linguists have developed methods of reconstruction and taxonomy which are helpful toward an interdisciplinary understanding of the diversity of peoples.

Indeed the identity of a people is often intimately coupled to the language it speaks. Linguistic grouping has been taken, time and again, to be the first criterion for sorting out human diversity. The celebrated diagram published by Cavalli-Sforza et al. (1988), comparing a genetic tree with a linguistic tree, was an eloquent statement on the important parallelisms between genetic evolution and linguistic evolution on a global scale. More locally, when a method developed to quantify genetic affinity was applied to a chain of languages in Micronesia, it was found to yield comparable results (Cavalli-Sforza and Wang 1986, reprinted in Wang 1991).

At the same time, however, languages and genes do go their separate ways, and such cases are not hard to find. When one ethnic group conquers another ethnic group, the common language eventually arrived at may be that of the conqueror, or that of the conquered. The latter is clearly the case with the Manchus, an Altaic people from northeastern China who founded the Qing dynasty and ruled the entirety of China for nearly 300 years. Although there are numerous monuments and documents which attest to the glory of their long reign, the Manchu language has been all but replaced by the language of the Han majority. Li (2000, p. 15) describes the situation this way.

"A survey done in the People's Republic of China in the 1950's found that quite a few elderly Manchus who lived in the more remote regions of Manchuria could still speak Manchu. Those over thirty years old were likely to understand it, while the younger generation could neither speak or [sic] understand it. Since then, anthropologists and linguists doing research in northern Manchuria have been reporting on a rapidly dwindling number of Manchu speakers. By the 1990s Manchu speakers have become nearly non-existent."

Such cases of language displacement, by no means rare, remind us that genes and languages can and do go separate ways. While they match in the default case, we should not be disturbed when their phylogenies do not agree. In fact, the cases of mismatch are in a sense more interesting since they may reveal displacement events long ago which would be difficult to uncover otherwise.

Potential contributions from linguistics on the question of human diversity come under three headings:

1. To establish genetic groups and subgroups of languages.
2. To locate the homeland of speakers of ancient languages.
3. To date splits among languages.

The study of language prehistory has a distinguished tradition in many cultures. In China, reconstructing the rhymes of ancient poetry reached a high level of scholarship in the 16th century. In the West, historical linguistics traces its roots to a famous lecture given in 1786 by William Jones. The following paragraph with which he announced the genetic relatedness among some of the languages in Europe and in Asia is perhaps the most often quoted in linguistics:

“The Sanskrit language, whatever be its antiquity, is of a wonderful structure; more perfect than the Greek, more copious than the Latin, and more exquisitely refined than either, yet bearing to both of them a strong affinity, both in the roots of verbs and in the forms of grammar, than could possibly have been produced by accident; so strong indeed, that no philologer could examine them all three, without believing them to have sprung from some common source, which perhaps no longer exists; there is a similar reason, though not quite so forcible, for supposing that both the Gothick and the Celtick, though blended with a very different idiom, had the same origin with the Sanskrit; and the old Persian might be added to the same family, if this were the place for discussing any questions concerning the antiquities of Persia.” (Quoted in Cannon 1991:31)

Building upon Jones’s insight, a great deal has been achieved toward clarifying the relationships among the 6000 or so languages spoken in the world today. The reconstruction of the Proto-Indo-European, the “common source” that Jones conjectured in the above paragraph, together with the light it sheds on civilizations of some 7,000 years ago, has become a standard in scholarship to be emulated everywhere. Many proto-languages of similar time depths have been reconstructed.

Currently, there is a spectrum of positions on how much time depth is recoverable in language for determining genetic relationships. At one end of the spectrum, some linguists have been reluctant to venture beyond the time depth established by Indo-European studies. Since a living language is constantly changing, these linguists believe that nothing reliable will be left of the original language after 7,000 years to be of diagnostic value. Although this ceiling of 7,000 years has never been objectively justified, it seems to reflect a bias from Indo-European studies. At the other end of the spectrum, some linguists propose global etymologies, roots of words which can be found in all major phyla. These linguists believe that all the world’s languages can be traced to a single monogenetic source.

While monogenesis is the dominant view today, probabilistic considerations actually favor a scenario in which language was invented independently at many sources, i.e., polygenesis (Freedman and Wang 1996). In pondering these issues, we should also take into account the effects of global events such as major glaciations, which must have scrambled human populations extensively by forcing distant migrations. It would be difficult to establish linguistic lineages across such barriers of panmixia.

Although methods of taxonomy are not nearly as well developed in linguistics as in biology, nonetheless a general picture is emerging, largely

thanks to the pioneering efforts of Joseph H. Greenberg of Stanford University. Figure 1 shows the dozen or so phyla he proposes for the languages of the world. This classification is discussed by Ruhlen (1991). While most of the details remain to be worked out, his proposal is the first major framework within which future research can be anchored. The phylum that Greenberg has been investigating in depth himself is one he calls Eurasiatic. As shown in Figure 2, the Eurasiatic phylum has Indo-European as one of its branches, but also comprises many other branches as well, including the enigmatic Ainu language, which has been considered by most to be a linguistic isolate. Greenberg's results (2000), which have been just published, are sure to elicit very different responses from linguists of various persuasions.

Quite independent of Greenberg's research, a group of Russian linguists, led by the late Illich-Svitych, have also proposed a large phylum of languages, which they call Nostratic. For some discussion of the Nostratic proposal, see the anthology edited by Salmons and Joseph (1998). It is instructive to compare the memberships of the two proposals, as seen in Table 1. Much of the original work on the two proposals was done during the decades when communication across the continents was hampered by political curtains, and the sharing of data was difficult. Recent years have seen closer interactions between the linguists of the U.S. and Russian, with the encouraging result of increasing convergence in their views.

Another phylum of great interest is Dene-Caucasian. The proposal by Sergei Starostin (1990), a linguist at the Moscow State University, is shown in Figure 3. Again, while some members of the phylum may be firmly established, such as Sino-Tibetan, much work needs to be done for the proposal to reach general acceptance. An example of recent progress here is the finding of Ruhlen (1998), on the Yeniseian and Na-Dene, which are two branches of the Dene-Caucasian. This finding of 36 common etymologies is of special interest since it definitively connects languages which are currently distributed on opposite sides of the Pacific.

There is still no consensus regarding the distant affiliations of the Chinese language. This is reflected in a monograph edited by Wang (1995), in which E.G.Pulleyblank discusses the connection between the Chinese and Indo-European. Laurent Sagart (see Wang 1995) discussed the Chinese and Austronesian. In the same monograph, Starostin shows the number of basic words, defined by Sergei Yakhontov (see Wang 1995), shared among these language groups. In Table 2, Starostin's numbers have been converted to

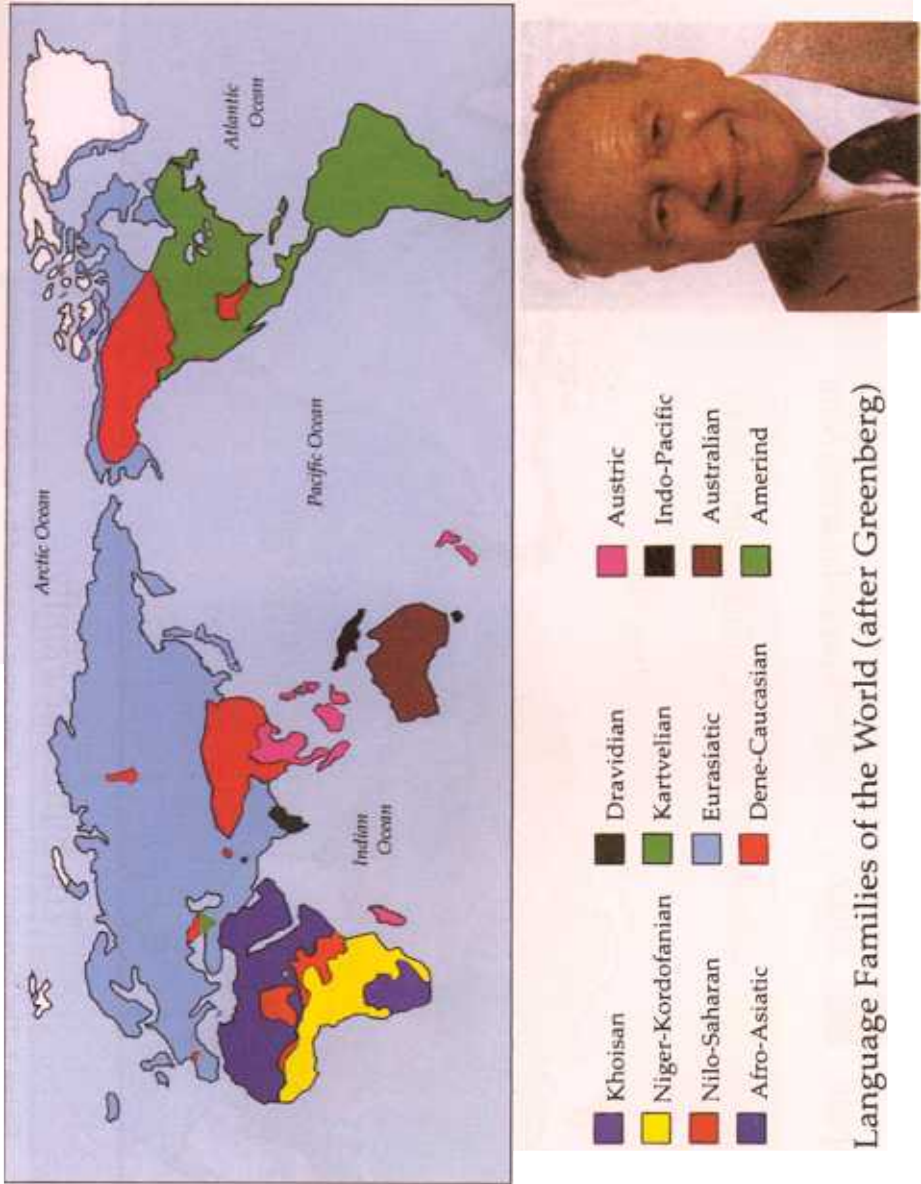
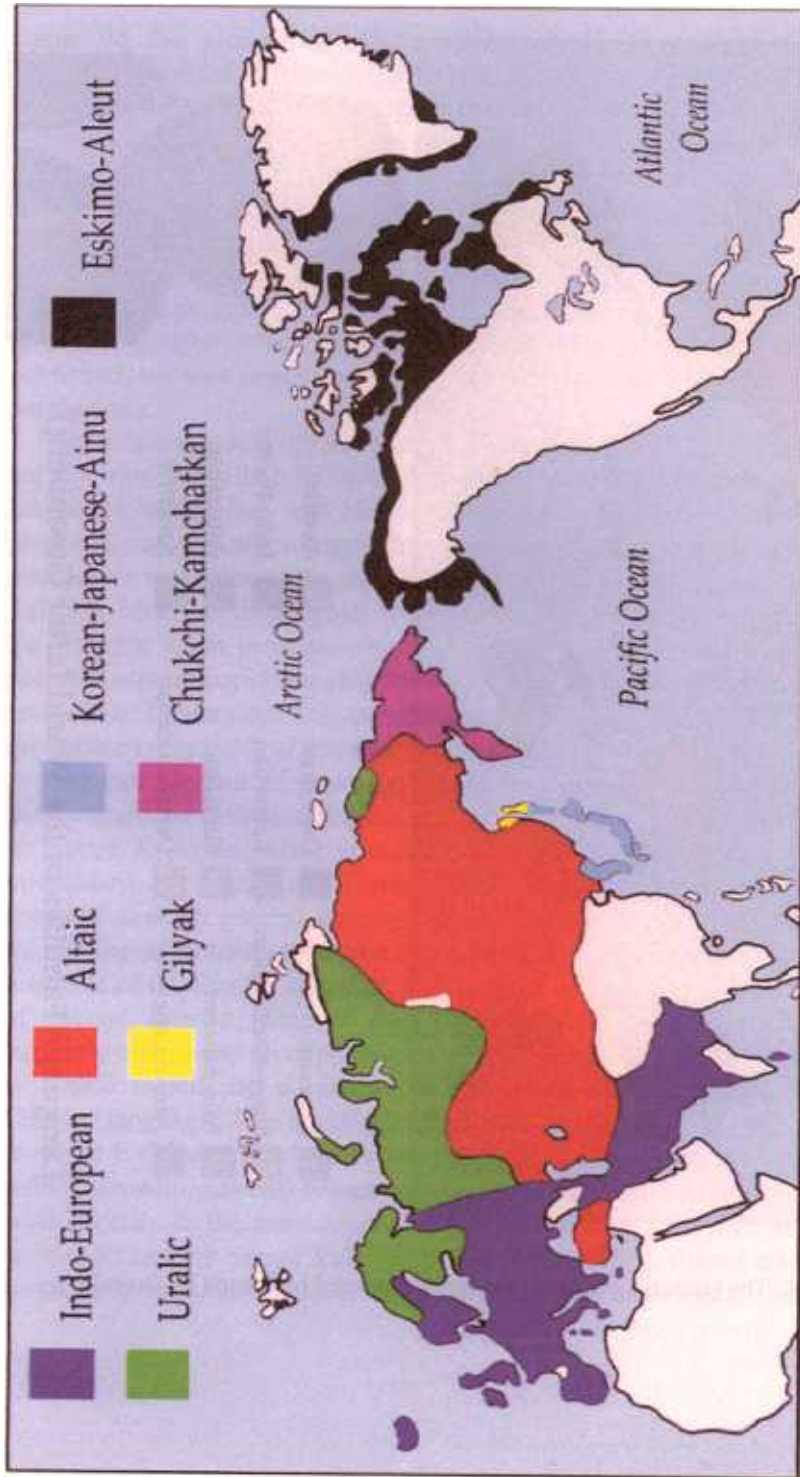
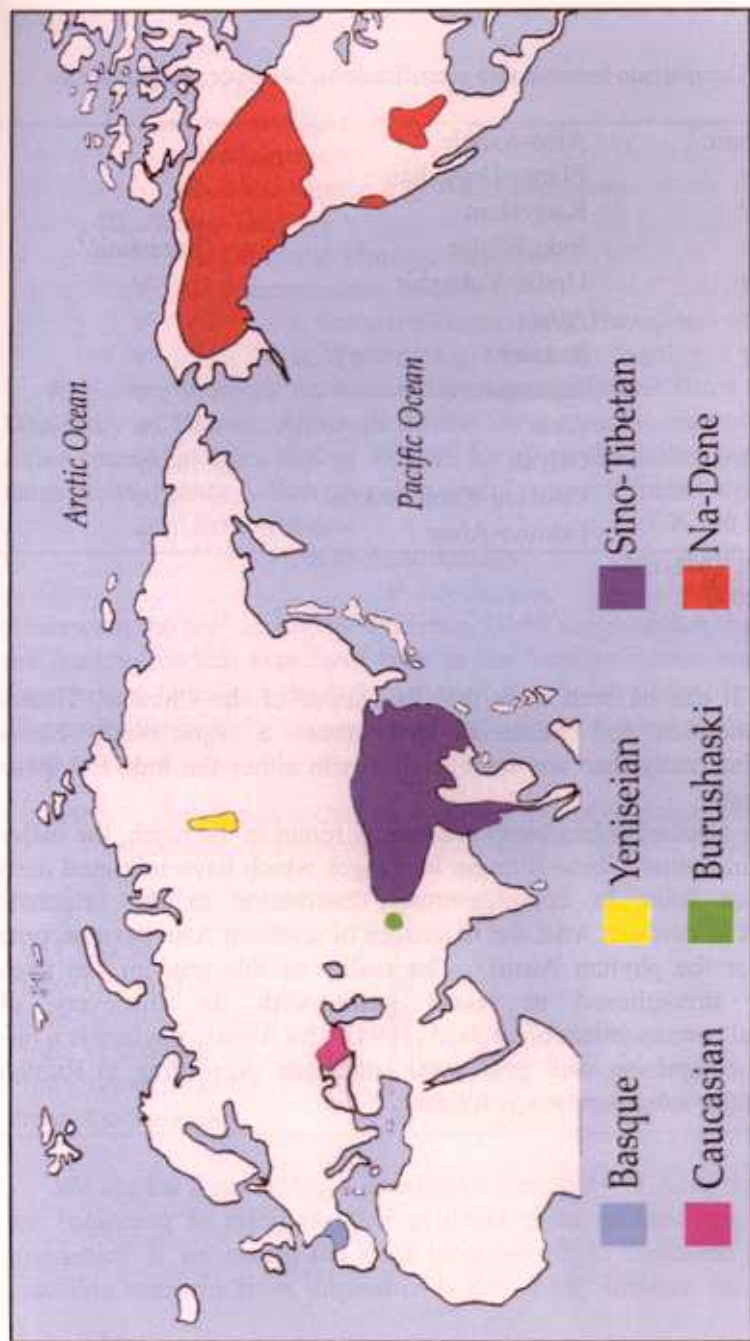


Figure 1. The language phyla of the world, proposed by Joseph H. Greenberg.



The Eurasian Family

Figure The Eurasian phylum languages, proposed by Joseph Greenberg.



The Dene-Caucasian Family

Figure Map Dene-Caucasian phylum of languages, proposed by Sergei Starostin.

Table Comparison between two classifications, Nostratic and Eurasiatic

Nostratic	Afro-Asiatic	Eurasiatic ²
∇	Elamo-Dravidian	
∇	Kartvelian	
∇	Indo-Hittite	
∇	Uralic-Yukaghir	∇
∇	Altaic	∇
∇	Korean	∇
	Japanese	∇
	Ainu	∇
	Gilyak	∇
	Chukchi-Kamchatkan	∇
	Eskimo-Aleut	∇

¹ Illich-Svitych, 1971-1984

² Greenberg, 2000

percentages. It can be seen there that the subset of the Chinese, Tibeto-Burman, Caucasian and Yeniseian does show a significantly closer relationship internally than any member has with either the Indo-European or Austronesian.

The Dene-Caucasian languages are largely found in the north; the major exception being some Tibeto-Burman languages which have migrated deep into Southeast Asia. In complementary distribution to the linguistic developments in northern Asia, the languages of southern Asia have become grouped under the phylum Austric. The reality of this phylum has been considerably strengthened in recent years with the discovery of morphological correspondences by Reid (1994). The Austric phylum is a far-flung group, comprising well over 1000 languages. According to Ruhlen (1991), the major subgroups are as follows:

Austriac:

5. I. Miao-Yao
 II. Austro-Asiatic
 a. Munda
 b. Mon-Khmer. e.g. Wa, Vietnamese.
 III. Austro-Tai
 a. Daic. e.g. Zhuang, Thai, Lao.
 b. Austronesian.
 i. Eastern = Oceanic, e.g. Hawaiian.
 ii. Western. e.g. Malagasy, Tagalog.

A leading authority on Austriac languages is Robert Blust (1996) of the University of Hawaii. Although Blust's latest classification of Austriac may differ somewhat from that of Ruhlen, he offers the following approximate dates of divergence, which provide a useful temporal framework.

Proto-Austriac	8,500 BP
Proto-Austronesian	6,500
Proto-Oceanic	4,000

Reviewing the archaeological evidence, Blust suggests that the last unity of the Austriac phylum may have been at the Yunnan-Burma border, splitting into various families, which then spread into South China, Southeast Asia. The paths these early migrants took probably followed the courses of the great rivers of Asia.

Table 2. The relation of Chinese to other groups of languages, shown as the percentage of apparent cognates from 35-word list of Yakhontov

	OC	PTB	PNC	PY	PIE
Old Chinese					
Proto-Tibeto-Burman	74				
Proto-North-Caucasian	43	51			
Proto-Yenisseian	34	40	57		
Proto-Indo-European	23	14	17	7	
Proto-Austronesian	14	11	11		14

We are far from having a conclusive prehistory of Asia, though scholars are beginning to bring together evidence from archaeology, genetics and linguistics. If we accept the three language phyla discussed above, then a plausible scenario from linguistics is this. Early humans entered East and

Southeast Asia, bringing with them two linguistic phyla, the Dene-Caucasian in the north and the Austric in the south. Their domains were later supplanted by the Eurasiatic phylum, particularly the Altaic family and the Indo-Iranian branch of the Indoeuropean family.

The Altaic family of languages stretches like a belt across Central Asia, stretching from Turkey in the west and extending to the Pacific in the east over several millennia. Only in recent centuries did Russian, a member of the Slavic branch of the Indo-European family, colonize large regions of northern Asia. The Indo-Iranian languages have moved into West Asia and South Asia, where they claim large communities of speakers in Iran, India and Pakistan. The expansion eastward of the Eurasiatic phylum covers over much of the territory earlier occupied by speakers of the Dene-Caucasian and Austric. With a few notable exceptions, such as Chinese, the earlier languages have been consistently shrinking as the Eurasiatic languages gained the upper hand.

Much of the evidence linguists offer is based on vocabulary. In any language, the vocabulary contains words which are more cultural, such as: tennis, television, tea, etc. Cultural words are frequently adopted from language to language, and hence are not stable indicators of genetic relations. On the other hand, all languages also have basic words which are much more stable, such as: water, hand, and tree. Although basic words do get adopted, they are relatively stable. As Morris Swadesh (1952) proposed in the 1950s, they provide a source of quantitative data for studying relations among languages.

Table 3 presents in tabular form one of the lists of 100 basic words Swadesh (1952) proposed that has gained wide acceptance in linguistic research. Various criticisms have been voiced against the concept of basic words in general, and against this list of 100 words in particular. Some scholars feel that the list is too inclusive, and whittle it down to fewer words. The table Starostin constructed, upon which Table 2 is based, uses a list of 35 words proposed by Yakhontov. In Table 3, these 35 words are shown in italics. As can be seen in the table, 32 of the 35 are in the Swadesh list. The three words Yakhontov proposes not in the Swadesh list are: salt, wind, and year.

Basic words as a method in studying linguistic prehistory has been used primarily in two contexts. One is to show degrees of affinity, as Starostin (1990) does in Table 2. The other is to estimate dates of the linguistic split. A central problem in the historical study of language is that of sorting out

linguistic traits which are vertically transmitted as opposed to those which are horizontally transmitted. The former mode is also called inheritance, and the latter mode is also called borrowing. The problem is extremely difficult because any linguistic trait can be transmitted either vertically or horizontally.

Figure 4 illustrates one approach to this problem in the form of a family tree for the Austronesian languages of Taiwan. Using standard methods of cluster analysis, I constructed a tree on the basis of a table of numbers of shared words among these languages (Wang 1989). Such trees are of course

Table 3. List of 100 basic words, proposed by Morris Swadesh. A smaller subset of 32 words - plus *salt*, *wind*, and *year* - proposed by Sergei Yakhontov are shown in italics

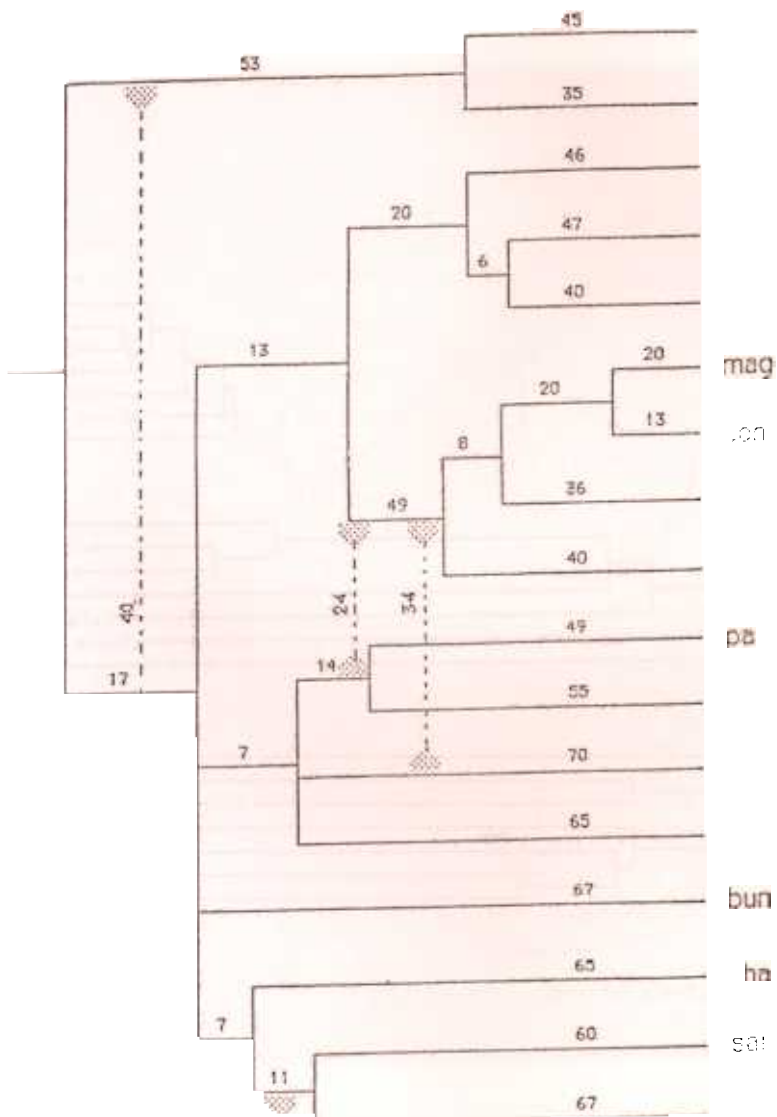
	Nature	Body	Animal	Verb	Adjective	Misc
1	ashes	belly	bird	bite	all	earth
2	bark	<i>blood</i>	claw	burn	big	<i>I</i>
3	cloud	<i>bone</i>	<i>dog</i>	come	black	<i>name</i>
4	<i>fire</i>	breast	feather	<i>die</i>	cold	night
5	leaf	<i>ear</i>	<i>fish</i>	drink	dry	not
6	man	<i>egg</i>	<i>horn</i>	eat	fat	<i>one</i>
7	<i>moon</i>	<i>eye</i>	<i>louse</i>	fly	<i>full</i>	road
8	mountain	foot	<i>tail</i>	<i>give</i>	good	
9	person	hair		hear	green	that
10	rain	<i>hand</i>		kill	long	<i>this</i>
11	root	head		<i>know</i>	many	<i>thou</i>
12	sand	heart		lie	<i>new</i>	<i>two</i>
13	seed	knee		say	red	we
14	smoke	liver		see	round	<i>what</i>
15	star	meat		sit	small	<i>who</i>
16	<i>stone</i>	mouth		sleep	warm	
17	<i>sun</i>	neck		stand	white	
18	tree	<i>nose</i>		swim	yellow	
19	<i>water</i>	skin		walk		
20	woman	<i>tongue</i>				
21		<i>tooth</i>				
	<i>salt</i>					<i>year</i>
	<i>wind</i>					

time-honored ways of graphing vertical transmission. On the basis of the resulting tree, I was able to make another table of the *presumed* number of shared words among these languages. Comparing these two tables enabled me to detect regions of mismatch, which I interpret to be due to horizontal transmission. These horizontal transmissions are indicated on the tree by broken lines. While such a modified tree does capture both modes of transmission, the method of its construction appears to give dominance to vertical transmission.

Using similar methods, I made an attempt to estimate the date of the split of the Sino-Tibetan family of languages, as shown in Figure 5. Details of this exercise are discussed more fully in (Wang 1998). I first constructed a tree of the major dialects of Chinese, which is shown at the top of the figure. The tree shown in the middle of the figure is one I constructed for Indo-European, following identical procedures. The encouraging result when comparing the two trees is that the 'height' of the tree for Chinese dialects is approximately the same as that for the three Germanic languages in the Indo-European tree. Based on these rough yardsticks, it would seem that the Sino-Tibetan tree at the bottom of the figure should be somewhat younger than the Indo-European tree. This means that if we assume that the Indo-European tree is 7,000 years old, then the Sino-Tibetan tree would be 6,000 years old.

Although definitive support for this date of 6,000 years, arrived at from linguistic data, is hard to come by from other disciplines, there is a map drawn by the Harvard archaeologist K. C. Chang (1986) which is very suggestive. This map, shown here as Figure 6, illustrates the period of 6,000 years ago in China when for the first time there was wholesale interaction among the many cultural spheres, based on archaeological finds. The melting together of these many cultures led Chang to refer to the period as 'initial China'. There is, then, an encouraging convergence of results here between archaeology and linguistics.

With the dramatic advances made by genetics in recent years, there is accumulating an ever increasing body of genetic data that can be compared with archaeological and linguistic hypotheses. Such comparisons will surely deepen our understanding of the nature of human diversity and linguistic diversity, whether or not genetic and linguistic maps always agree. In either case, it is certain that we had only one past, and mismatches between the maps can yield important insights on when genes and languages went separate ways.



verti horizontal transmission the languag

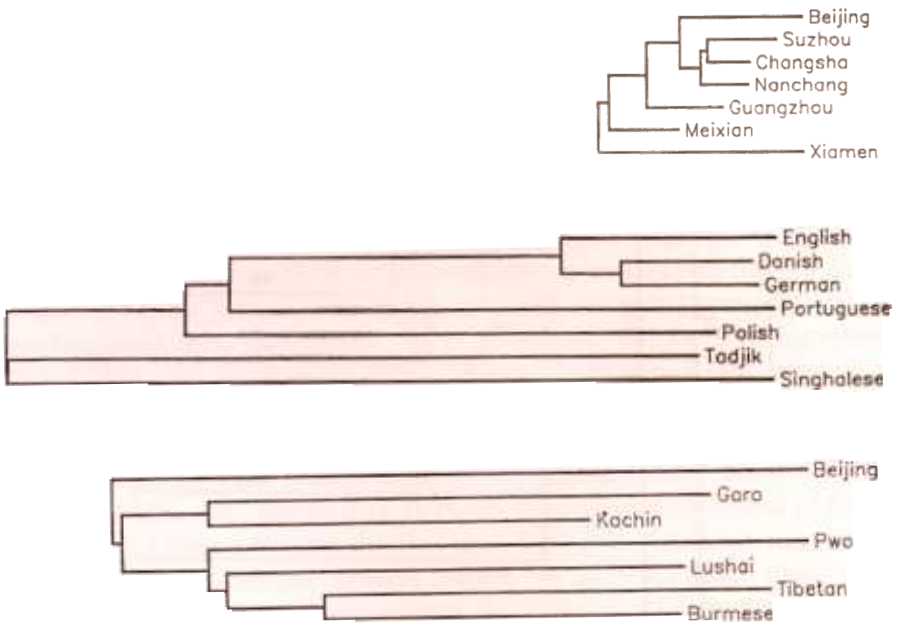


Figure 5. Additive trees of Chinese, Indo-European, and Sino-Tibetan.

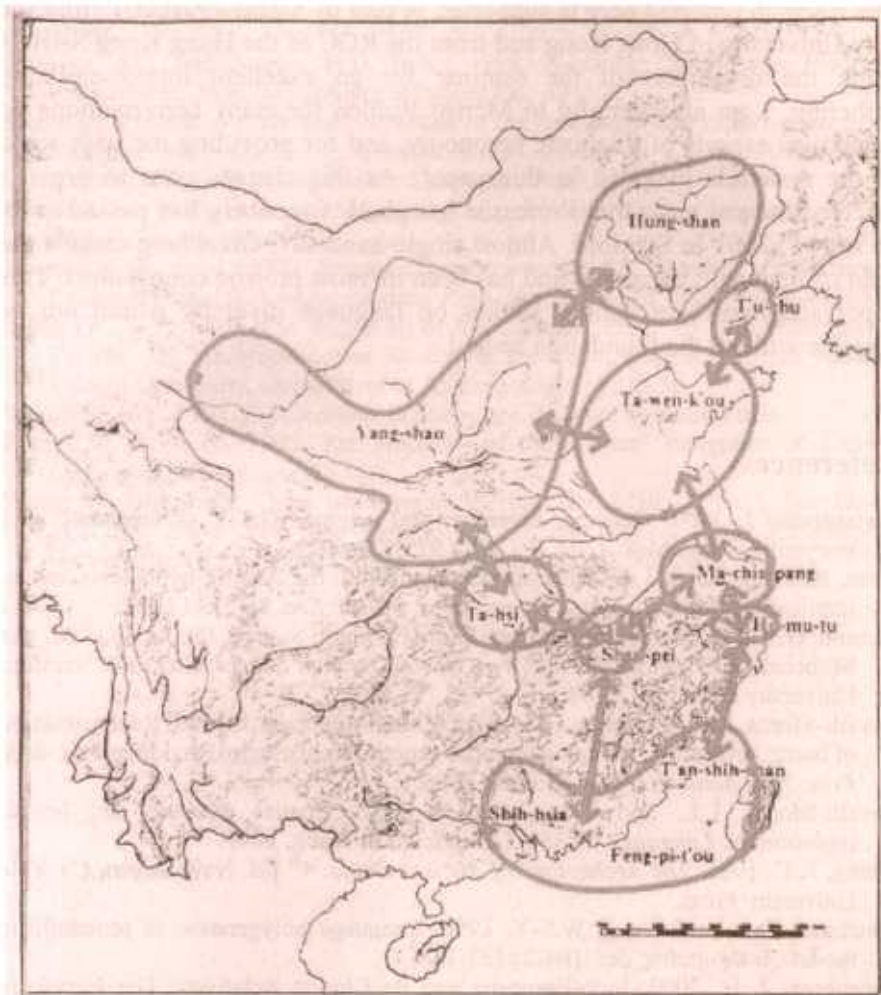


Figure 6. China in prehistory as revealed by archaeology. [Figure adapted from K.C. Chang, p. 235.]

Acknowledgements

The research reported here is supported in part by Grant #9010001 from the City University of Hong Kong and from the RGC of the Hong Kong SAR. I thank the organizers of the seminar for an excellent interdisciplinary gathering. I am also grateful to Merritt Ruhlen for many conversations on theoretical aspects of linguistic taxonomy, and for providing me with some of the materials included in this paper. As this chapter goes to press, I received the sad news that Professor Joseph H. Greenberg has passed away on May 7, 2001 in Stanford. Almost single-handedly, Greenberg created the field of linguistic taxonomy and has been its most prolific contributor. This paper and numerous similar studies on language diversity would not be possible without the foundation he laid.

References

- Bertranpetit, J. 2000. Genome, diversity, and origins: The Y chromosome as a storyteller. *Proc. Natl. Acad. Sci. USA* 97:6927-6929.
- Blust, R. 1996. Beyond the Austronesian homeland: the Austric hypothesis and its implications for archeology. *Trans. Amer. Philos. Soc.* 86(5):117-160.
- Cannon, G. 1991. Jones's Sprung from Some Common Source. IN: Lamb, S.M. and Mithcell, E.D. (eds.), *Sprung from Some Common Source*. Stanford: Stanford University Press, pp. 23-47.
- Cavalli-Sforza, L.L., Piazza, A., Menozzi, P. and Mountain, J. 1988. Reconstruction of human evolution: Bringing together genetic, archeological and linguistic data. *Proc. Natl Acad. Sci. USA* 85:6002-6006.
- Cavalli-Sforza, L.L. and Wang, W.S-Y. 1986. Spatial distance and lexical replacement. *Language* 62:38-55. Reprinted in Wang, 1991.
- Chang, K.C. 1986. *The Archeology of Ancient China*. 4th Ed. New Haven, CT: Yale University Press.
- Freedman, D.A. and Wang, W.S-Y. 1996. Language polygenesis: A probabilistic model. *Anthropolog. Sci.* 104(2):131-138.
- Greenberg, J. H. 2000. *Indo-European and its Closest Relatives: The Eurasiatic Language Family*. Stanford: Stanford University Press.
- Klein, R. G. 1999. *The Human Career*. 2nd ed. Chicago: University of Chicago Press.
- Li, G. R. 2000. *Manchu: A Textbook for Reading Documents*. Honolulu: University of Hawaii Press.
- Reid, L. A. 1994. Morphological evidence for Austric. *Oceanic Linguistics* 33:323-344.

- Ruhlen, M. 1991. *A Guide to the World's Languages*. Stanford: Stanford University Press .
- Ruhlen, M. 1998. The origin of the Na-Dene. *Proc. Natl. Acad. Sci. USA* 95:13994-13996.
- Salmons, J.C. and Joseph, B.D. (eds.) 1998. *Nostratic: Sifting the Evidence*. Philadelphia: John Benjamins Publishing Co.
- Starostin, S. 1990. *A statistical evaluation of the time-depth and subgrouping of the Nostratic macrofamily. Symposium on Molecules to Culture*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, p. 33.
- Swadesh, M. 1952. Lexicostatistic dating of prehistoric ethnic contacts. *Proc. Am. Philos. Soc.* 96:452-463.
- Thompson, R. et al. 2000. Recent common ancestry of human Y chromosome: Evidence from DNA sequence data. *Proc. Nat. Acad. Sci. USA* 97:7360-7365.
- Wang, W. S-Y. 1989. The migration of the Chinese people and the settlement of Taiwan. IN: *Anthropological Studies of the Taiwan Area*. Taiwan: National Taiwan University, Department of Anthropology, pp. 15-36.
- Wang, W. S-Y. 1991. *Explorations in Language*. Taiwan: Pyramid Press.
- Wang, W. S-Y., ed. 1995. The Ancestry of the Chinese Language. *J. Chinese Linguistics Monograph* 8.
- Wang, W. S-Y. 1998. Three windows on the past. IN: Mair, V. (ed.), *The Bronze Age and Early Iron Age Peoples of Eastern Central Asia*. Philadelphia: University of Pennsylvania Museum Publications, pp. 508-534.