https://doi.org/10.1038/d41586-019-01214-6

LINGUISTICS

The origin and spread of Sino-Tibetan languages

A robust computational approach with added finesse provides evidence to support the view that the Sino-Tibetan languages arose in northern China and began to split into branches about 5,900 years ago.

RANDY J. LAPOLLA

The location and timing of the emergence of the Sino-Tibetan language family has long been debated. This family has around 1.5 billion speakers worldwide, the second largest number of speakers globally after those who speak languages in the Indo-European family. One school of thought is that the ancestral language (Proto-Sino-Tibetan) from which all the Sino-Tibetan languages evolved originated in northern China around 4,000–6,000 years ago[1,2]. An alternative view is that it arose 9,000 years ago in southwest China or northeast India[3,4].

Writing in *Nature*, Zhang *et al.*[5] report a study that might settle this debate. The authors gathered evidence about the Sino-Tibetan language family and its speakers from disciplines including genetics, computational biology, linguistics, archaeology and anthropology, and also compiled information about the development of agriculture and its possible effects on human migrations in the region. They then used a method of probability testing to assess the different language family trees that could be made on the basis of this evidence.

Historical linguists seek to determine the relationships between languages, and usually take an approach called the comparative method. They look for cognate words in different languages — words that have similar meanings and that can be shown to have a shared origin in a word from an earlier, ancestral language. Linguists then try to explain why the words often don't look exactly alike: the changes that the sounds went through, what additions were made to the words, and what led to the words being used, in some cases, for different meanings in related languages. For example, work in Indo-European linguistics has determined that the English word cow and the French word *boeuf* are part of a family of cognate words that have descended from a reconstructed Proto-Indo-European root word, *gwou- (the asterisk indicates a reconstructed form and the hyphen that it is a root that formed a number of different words)[6]. Understanding such changes enables language families such as the Indo-European family to be split into branches, such as the Romance, Germanic and Slavic languages, on the basis of shared changes.

The use of particular words found to be cognate, together with evidence from other fields, can help inferences to be made about the relationship of languages to human migrations, and the emergence of human cultures. This can then aid efforts to determine the home of the speakers of an ancestral proto-language, when these people and their language dispersed and the different branches of the language family formed. However, the vagaries of history that have led to criss-crossing migrations, contact between different languages and cultures and other sociological factors have often meant that it is difficult to identify the family tree that correctly represents the history of a language family. Competing interpretations of the same data can lead to the generation of different trees and to different models of the origin and dispersal of a particular language. And it has previously been difficult to evaluate all of the possible trees that could be made on the basis of the available data.

Modern computers now make it possible to handle large amounts of data and calculations rapidly. Software developed for biosciences research that applies a particular model of probability testing known as Bayesian phylogenetic modelling can also be used in linguistics. This software can test the many possible language trees that could be made from a data set, and thereby determine the most likely tree and the most probable time frame for language diversification.

Zhang and colleagues focused on the Sino-Tibetan family, which encompasses hundreds of languages, including Chinese, Tibetan, Burmese and many other, less widely spoken, languages. The authors used data on cognate terms that have been assembled over the past 30 years in a project called the Sino-Tibetan Etymological Dictionary and Thesaurus (see go.nature.com/2uombqo). This provided a solid basis of relevant data for their calculations, and set Zhang and colleagues' study apart from earlier work that applied similar computational techniques but used random word lists from word families that had not been evaluated for cognacy, affecting the reliability of those studies.

The authors used these language data together with information from other fields, such as anthropology, and ran millions of iterations of their computer program. They determined the most likely location of the homeland of the ancestors of the modern Sino-Tibetan-speaking peoples, and the most probable time

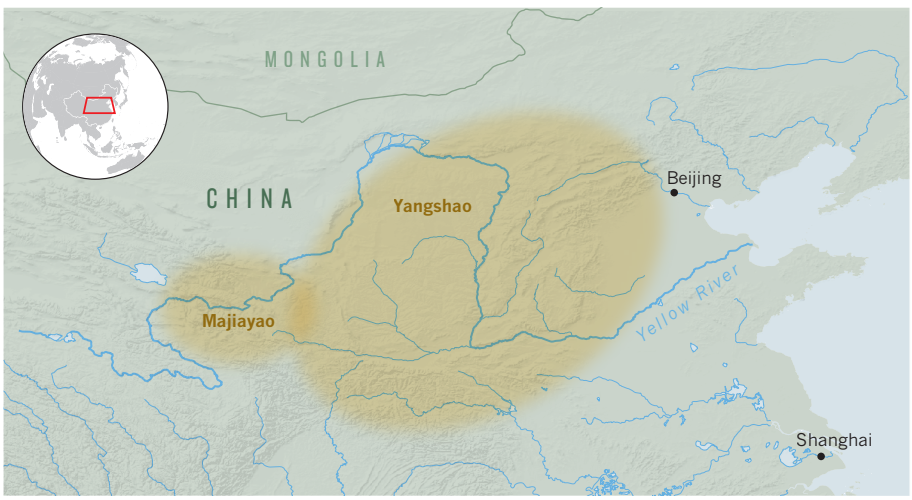

Figure 1 | Site of origin of the Sino-Tibetan languages. Zhang *et al.*[5] present the results of a probability-testing approach used to analyse data relating to the origins and spread of the Sino-Tibetan languages, which are spoken today by 1.5 billion people. Their analysis indicates that, consistent with one current model[1], the ancestral form of the language originated approximately 5,900 years ago in northern China, in the basin of the Yellow River. They identify the origin and earliest spread of the languages as being associated, respectively, with the Yangshao culture and the later Majiayao[7] (cultures indicated in shaded regions).

frame for when this language family began to diverge into subgroups as some members of the group of early Sino-Tibetan speakers migrated away from where the language originated. The authors also determined the most probable language family tree and which type of branching structures had the highest probability of representing the relationships between the languages.

Zhang *et al.* compared the two competing views of where the earliest Sino-Tibetan speakers originated. Their results support the theory that the homeland of the Proto-Sino-Tibetan language was in the Yellow River basin region (Fig. 1) of present-day northern China, and that the dispersal and diversification of this language family began around 5,900 years ago. At that time, this region was associated with the Yangshao culture and the later Majiayao (a culture thought to have arisen after a westward migration of people from the Yangshao culture)[7]. These cultures were associated with pottery and silk production, and the communities kept domesticated animals and had large, fixed settlements.

The results indicate that there was a major initial split between the Sinitic languages and the Tibeto-Burman languages before each of these two groups split further into linguistic sub-branches. This contrasts with one current model[3] suggesting that these two branches did not form from a major initial bifurcation. That model proposes instead that many branches formed at the same time. It suggests that the Sinitic languages do not form a major branch that is split from all of the other languages, and that what are commonly referred to as the Tibeto-Burman languages do not group into a single branch[3].

Zhang and colleagues' work is important in many ways. The history of the Sino-Tibetan languages has not been studied for as long as has the history of the Indo-European languages. Thus, by comparison, there has been much less certainty about some of the key points that provide a foundation for this area of research, such as the origins of the language. The authors' work provides more certainty on such fundamental issues, freeing researchers to build on this and to explore the history of this language family more deeply. The work should also help to identify connections between these language studies and findings from other related fields, such as archaeology and history. ■

Randy J. LaPolla *is in the School of Humanities, Nanyang Technological University, Singapore 639818, Singapore.*
e-mail: randylapolla@ntu.edu.sg

1. LaPolla, R. J. in *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics* (eds Aikhenvald, A. Y. & Dixon, R. M. W.) 225–254 (Oxford Univ. Press, 2001).
2. Bradley, D. *10th Int. Conf. Evol. Linguistics* Nanjing Univ. (2018); go.nature.com/2udgyy9
3. van Driem, G. in *Trans-Himalayan Linguistics* (eds Owen-Smith, T. & Hill, N. W.) 11–40 (de Gruyter, 2014).
4. LaPolla, R. J. *Linguist. Tibeto-Burman Area* **39**, 282–297 (2016).
5. Zhang, M., Yan, S., Pan, W. & Jin, L. *Nature* https://doi.org/10.1038/s41586-019-1153-z (2019).
6. Buck, C. D. *A Dictionary of Selected Synonyms in the Principal Indo-European Languages* 152 (Univ. Chicago Press, 1949).
7. Li, L. & Chen, X. *The Archaeology of China: From the Late Paleolithic to the Early Bronze Age* (Cambridge Univ. Press, 2012).