Creating a database for Tibeto-Burman languages

In the number of its speakers, Tibeto-Burman is one of the largest language families in the world. The language family, however, has received little scholarly attention and its composition and history remain poorly understood. Many languages are still awaiting detailed documentation – a task that is becoming urgent as serious demographic pressures. Given the dazzling linguistic diversity and sheer number of languages yet to be studied, a thorough understanding of the Tibeto-Burman language family poses great challenges. One complicate factor is that presently available data are scattered, making an overview of the family and adequate historical comparisons unfeasible.

Katie Chirikos

Crossing the Himalayas

Tibeto-Burman languages are the most well-represented language family in the Himalayan region, broadly understood as stretching from the Chinese provinces of Qinghai in the north to the southern tip of Burma, and from northeastern Vietnam in the east to northern Pakistan in the west. Our research programme – Trans-Himalayan Database Development: China and the Subcontinent – pools the expertise of two renowned centres of Tibeto-Burman research. George van Driem’s Himalayan Languages Project (HLP) at Leiden University and the Chinese Academy of Social Sciences’ Institute of Ethnography and Anthropology (IEA) in Beijing. Both have worked for years on the documentation and description of Tibeto-Burman languages, with the Himalayan Languages Project concentrating on languages of the Indian subcontinent (Bhutan, Nepal and India) and the Institute of Ethnography and Anthropology on languages spoken within China’s borders (in the Tibetan Autonomous Region and the provinces of Qinghai, Sichuan and Yunnan).

Both parties have thus been simultaneously working north and south of the Himalayan range that divides China and the Indian subcontinent. The mutual aspiration of acquiring a better understanding of the Tibeto-Burman family has moved them, figuratively, to cross the Himalayas in combining their achievements and sharing their research: a wealth of data on over 80 Himalayan languages amassed by the HLP, and on 57 (out of over 80) Tibeto-Burman languages spoken in China collected by the IEA.

Our current research programme targets the two major challenges of Tibeto-Burman research. On the one hand, it contributes to the documentation of endangered languages. On the other, it aspires to assemble all collected data digitally to enable multi-leveled research and, ultimately, balanced and well-documented foundation for historical reconstruction and sub-grouping within the family, and ultimately, influence studies in the history of Tibeto-Burman and enrich the theoretical foundations and methodology of comparative linguistics.

Notes


Katie Chirikos is a fellow at IAS and co-editor of the programme Trans-Himalayan Database Development: China and the Subcontinent. Her field of research is Chinese linguistics and Tibeto-Burman languages in China.

k.chirikos@leidenuniv.nl

Trans-Himalayan Database Development: China and the Subcontinent

The programme, launched in June 2005, pools the expertise of two internationally renowned centres of Tibeto-Burman research: George van Driem’s Himalayan Languages Project at Leiden University and the Chinese Academy of Social Sciences’ Institute of Ethnography and Anthropology. It receives funding from the Royal Netherlands Academy of Arts and Sciences, the Chinese Academy of Social Sciences, and IAS.

- Scientific supervision: George van Driem (Leiden), Sun Hongkai (CAS), Huang Xing (CAS)
- Documentation and description of Shina, Bola and i-gal-rong: Katie Chirikos, Anton Lustig, Marielle Prins
- Data input, translation, annotation: Katie Chirikos, Hu Hongzhan, Huang Chengfeng, Liu Guangkan, Anton Lustig, Mu Shuishuai, Marielle Prins, Wang Feng, Yin Weinan, Zhou Mocao
- Database design and maintenance: Jing Di and Jean Robert Opgenort

Diagram 1

West Himalayas

East Himalayas

Tibet

Burma

China

Funding

International Institute for Asian Studies

Theoretical and practical challenges abound. The present programme brings together two distinct scholarly traditions with different understandings of Tibeto-Burman languages. The prototype of the Sino-Dutch database, that of the IEA, is structured in accordance with the Sino-Tibetan model as accepted in China, and includes data on Chinese, Tibeto-Burman, Tai-Kadai and Hmong-Mien languages spoken within China’s borders. This model presents each language family as consisting of hierarchically organised subgroups with an a priori implied appreciation of their phylogenetic relationships. In the agnostic model advocated by Van Driem, the precise phylogenetic relationships between the recognised subgroups of Tibeto-Burman languages (which in his understanding also includes Chinese) have not been precisely determined. As different models of the exact sub-grouping of Tibeto-Burman languages abound, our project aims to provide a fair overview of diversity within the Tibeto-Burman family and to let the data speak for themselves rather than formatting it in accordance with any model. Ultimately, it would be beneficial to create a system allowing users to group data according to the model they feel is most relevant even to create their own model to test against data in the database.

The precise structure of the database is currently under negotiation. The prototype of the proposed database, that of the Institute of Ethnography and Anthropology, is searchable both by Chinese translations and by semantic field and includes 1,532 predefined basic lexical items for each language. Words are grouped by semantic fields, such as body parts or celestial bodies, and are accompanied by their Chinese translation and morphological analysis (e.g., initial, codas, tone, prefixes if any, roots if any, affix if any). A logical continuation of work on the IEA database would be to increase the number of languages in the joint Sino-Dutch database by the standard number of 1,532 basic lexical entries per language. Given the overall goal of this project and the elaborate nature of the vocabularies collected by the HLP, however, it would be desirable to include as much lexical data as possible (viz. words, accompanied by an English translation and lexical, morphological and grammatical comment when available). It also seems sensible to create a database which can be expanded over time.

Our project is still in its initial stage; the very idea of an open and online database is to lay the foundation for further scientific cooperation. Given the depth and breadth of our aims, we invite participation from all individual researchers and teams working in the field so that we may create a broad, well-documented foundation for historical reconstruction and sub-grouping within the family, and ultimately, influence studies in the history of Tibeto-Burman and enrich the theoretical foundations and methodology of comparative linguistics.