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 and description – a task that is becoming urgent as smaller languages fall victim to socio-economic and 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 complicating factor is that presently available data are scattered, making an overview of the family and adequate historical comparisons unfeasible.

Katia Chirkova

Crossing the Himalayas

Tibeto-Burman is the most well-represented language family in the Himalayan region, broadly understood as stretching from the Chinese provinces of Sichuan and Qinghai in the north to the southern extremity of Burma, and from northwestern 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 focusing on languages of the Indian subcontinent (Bhutan, Nepal and India) and the Institute of Ethnography and Anthropology concentrating 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 answers to currently debated questions of historical development, sub-grouping and reconstruction.

Documenting endangered languages

The current programme's documentation of endangered languages builds upon previous work carried out by the HLP and the IEA. We chose to focus on three languages spoken in China: Shixing, a Qiangic language spoken in Muli county of Sichuan province; Bola, a Burmish language spoken in Yingjiang and Lianghe counties of Yunnan province; and rGyal-rong (Mkhono dialect), a Qiangic language spoken in Ma'erkang county of Sichuan province. These three insufficiently documented languages were selected not only for their severely endangered status, but because they constitute important links for historical reconstruction within each language group.

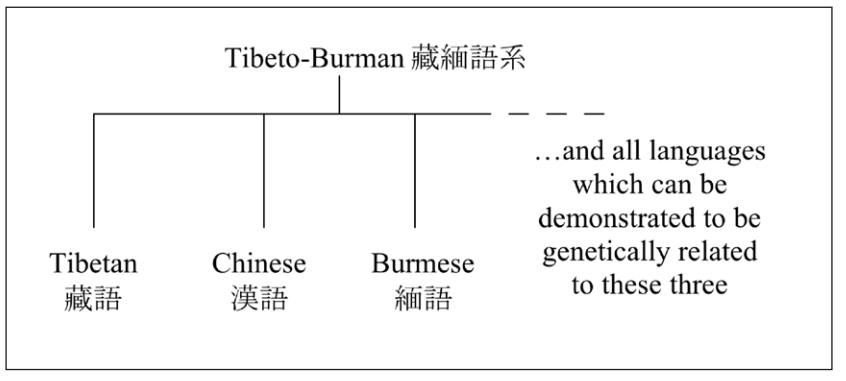

Diagram 1

Tibeto-Burman is one of the language families identified by Julius Heinrich von Klaproth in his 1823 polyphyletic view of Asian linguistic stocks. Klaproth's Tibeto-Burman family explicitly excluded languages today known to be Kra-Dai or Daic (e.g. Thai, Lao, Shan) or Austroasiatic (e.g. Mon, Vietnamese, Nicobarese, Khmer). van Driem, George. 2006. 'The Diversity of the Tibeto-Burman Language Family and the Linguistic Ancestry of Chinese. Paper presented at the 14th Annual Conference of the International Association of Chinese Linguistics and the 10th International Symposium on Chinese Languages and Linguistics, 27 May 2006, Academia Sinica, Taipei.

The documentation of these three languages will be carried out to the extent possible within the three-year project. We expect to complete concise descriptive grammars in Chinese, accompanied by 2,000 word lists and a selection of traditional stories, to be included in the Institute of Ethnography and Anthropology's renowned *Zhōngguó xīn fāxiàn yǔyán yánjiū cóngshū* [New Found Minority Languages in China] series, which aims to document all little-known and endangered languages in China.[1]

A digital database

A crucial step towards a better understanding of Tibeto-Burman languages is assembling existing data in a format that allows for long-term storage and efficient access and modification by multiple users – a goal best achieved through a digital database. Both parties already maintain their own: the IEA, in co-operation with the Hong Kong University of Science and Technology, has since 1998 been compiling a database of cognate words in Sino-Tibetan languages and their dialects (in Chinese)[2] while the HLP hosts a digital database correlating grammatical morphemes in Kiranti and other Tibeto-Burman languages (www.iias.nl/host/himalaya/projects/mld.html). Our goal is to combine data collected by both into one database of over 200 languages so that each branch of Tibeto-Burman will be represented by numerous languages and dialects.

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 relationship. 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 subgrouping of Tibeto-Burman languages abound, our project aims to give 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 different models, or 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 1,332 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 (eg, initial, coda, tone, prefix if any, root, affix if any, suffix 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 1,332 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 co-operation. 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. <

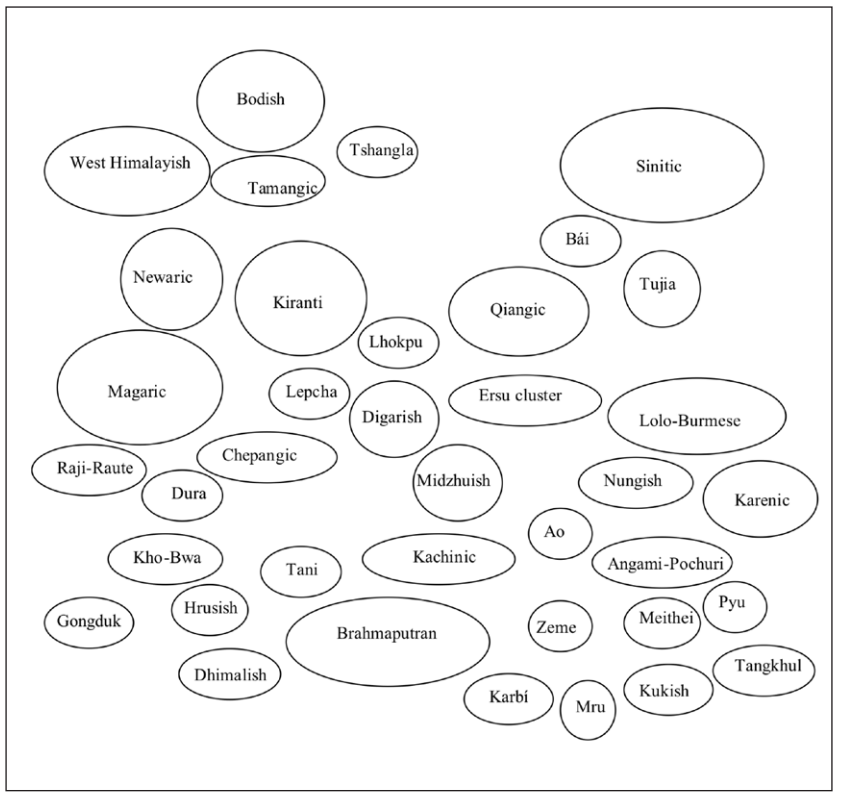

Diagram 2

Tibeto-Burman subgroups identified since Julius von Klaproth. Brahmaputran may include Kachinic and Dhimalish. Competing subgrouping proposals are discussed in the handbook *Languages of the Himalayas*. van Driem, George.2001. Languages of the Himalayas: An Ethnolinguistic Handbook of the Greater Himalayan Region, containing an Introduction to the Symbiotic Theory of Language. Leiden: Brill.

Notes

1. For information on the series see: Sagart, Laurent. 2003. 'A New Collection of Descriptions of Languages of China'. *Cahiers de Linguistique – Asie Orientale* 32(2), 287-298; Thurgood, Graham and Fengxiang Li. 2003. Book notice: Sun Hongkai, ed. *New Found Minority Languages in China Series*. Beijing: Chinese Academy of Social Sciences. *Language* 79-4, 843-845; and Chirkova, Katia. 2006. Review of *Zhōngguó xīn fāxiàn yǔyán yánjiū cóngshū* 《中国新发现语言研究丛书》 [New Found Minority Languages in China Series], 31 Volumes. Sūn Hóngkāi, ed. 孙宏开. Beijing 北京: Chinese Academy of Social Sciences 中国社会科学院. *China Review International* 13-1 (forthcoming).
2. A detailed report on the IEA database can be found in Ting Pang-Hsin and Sun Hongkai, eds. 2004. *Han-Zangyu tongyuanci yanjiu/Cognate Words in Sino-Tibetan Languages*. Nanning: Guangxi Minzu Chubanshe, vol. 3, 396-536, which contains Jiang Di's 'A Development Report on the Sino-Tibetan Cognate Database Retrieval Software' and 'The Sino-Tibetan Cognate Database Project: A Manual for Data Retrieval'.

Katia Chirkova is a fellow at IIAS and co-ordinator 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.chirkova@let.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 IIAS.

- Scientific supervision: George van Driem (Leiden), Sun Hongkai (CASS), Huang Xing (CASS)
- Documentation and description of Shixing, Bola and rGyal-rong: Katia Chirkova, Anton Lustig, Mariëlle Prins
- Data input, translation, annotation: Katia Chirkova, Hu Hongyan, Huang Chenglong Liu Guangkun, Anton Lustig, Mu Shihua, Mariëlle Prins, Wang Feng, Yin Weibin, Zhou Maocao
- Database design and maintenance: Jiang Di and Jean Robert Opgenort