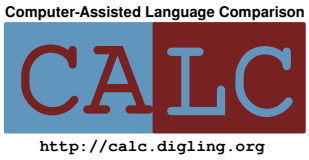

The ancestry of Sino-Tibetan populations and languages

Mei-Shin Wu*

with contributions by Yunfan Lai* and Johann-Mattis List*

July 2018

1 Introduction

1.1 Sino-Tibetan population genetic studies

- The prehistory of major ethnic groups in the area have been extensively studied.

  Studies focus on the peopling of the Tibetan plateau, the origin and migration or expansion of the Han population, or the relationship among Qiang, Han and Tibetan populations (Su et al., 2000; Yao et al., 2015; Kang et al., 2012; Torroni et al., 1994).

- Several large population genetic surveys have been done in East Asia. Thus, there are a lot of genetic datasets available.

  The HUGO Pan-Asian SNP Consortium, for example, provides population genetic data for East Asia, Central-East Asia and South-East Asia (Consortium, 2009).

- Combining various genetic datasets is the "traditional" approach for large-scale population genetic studies. There are guidelines and software available for extracting and merging subsets.

*Computer-Assisted Language Comparison (CALC), Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History

1.2 Sino-Tibetan linguistic studies

- A large language family in an important region, but we do not know much about its past...

  Various factors, including geography, population expansion and migration resulted in highly diversified languages. Most research focuses on the relationships among Sino-Tibetan languages (subgrouping of the language family), with few studies attempting to determine the age and the *Urheimat* of the Proto-Sino-Tibetan.

- Due to idiosyncrasies in documentation, the existing language data is not amenable for a large-scale computer-assisted or computer-based comparison.

  All comparison based on more formal systems (like computer-assisted approaches) require that data is not only readable for humans, but also formalized enough to be processed by machines. The current data is in some form, however, even not readable by humans, since many sources only provide very scarce glosses for the forms documented in word lists and dictionaries, with lexical or grammatical forms being transcribed in highly idiosyncratic orthographies or non-standardized traditions that look like the International Phonetic Alphabet only on the first side. To address these problems, we design workflows to extract common concepts, convert orthographies to phonetic transcriptions, and to combine data from different sources into larger datasets amenable for computer-assisted and computer-based comparison. (List et al., 2017a; Forkel et al., 2017).

- Computational methods for a detailed analysis of large scale cross-linguistic datasets are still in their infancy.

  Many linguistic features can be used to evaluate the similarity of a set of languages, including cognate judgments (List et al., 2017b), aggregated similarity of phonetic sequences (Jäger, 2013), or grammatical features (Longobardi and Guardiano, 2009). When provided manually, by experts, these tasks are very time-consuming. With the assistance of computational methods for automatic sequence comparison, we can speed up the tasks a lot. But since computational methods usually lag behind the accuracy of experts, the best way to combine the speed of computers with the accuracy of experts are computer-assisted frameworks in which software is used to pre- and post-process the data, while experts correct the computational findings (List, 2016b). Therefore, combining the benefit of experts' knowledge and computer algorithms can not only improve the efficiency but also the accuracy of our linguistic datasets.

1.3 Language and population genetic studies

- Systematically integrating techniques and findings from disciplines that deal with the past of human languages and populations could greatly improve our current knowledge about human prehistory.

- But integrating data and results across disciplines faces many challenges

- Different linguistic datasets cannot be directly compared, due to differences in coding and annotation.
- Ethnic groups and languages are not necessarily described by invoking a one-to-one relationship. Due to bilingualism and ethnic identity one ethnic group can represent a very intertwined history both with regard to languages and genes.
- For very few populations we have both population genetic and linguistic data.
- We have only few datasets on certain minority languages and language isolates, and even if we have sources from more than one author, the data may at times vary drastically (consider, for example, the differences in the data on Kusunda as found in (Donohue, 2012; *Kusunda linguistics*) and (Reinhard and Toba, 1970)).
- There are various ways to calculate linguistic similarities and distances, but we cannot tell by now which methods should be preferred.
- We lack appropriate methods to compare similarities and differences in linguistic and population genetic datasets.

2 Methods

2.1 Workflows for population genetic studies

Since the invention of high-throughput technology, screening genome-wide variation (as opposed to variation among selected genes only) has became very popular among population geneticists. Thanks to informatics, standardized workflows have been proposed for the preprocessing of large-scale datasets, and a couple of algorithms are now routinely applied to evaluate genetic similarities among target populations.

- For our genetic data, we use *single nucleotide polymorphisms* (SNPs) on the autosomal chromosome.

  Therefore, this study merely analyses the genetic structure of the target populations and ignores the effect of paternal or maternal inheritance.

- Using the Bioinformatics toolkit PLINK (Purcell et al., 2007), we can extract common SNPs and combine multiple microarray datasets.
- Employing the Fixation index (F_{ST}) (Holsinger and Weir, 2009), we can evaluate the genetic similarities across populations.
- We can use Neighbor-joining tree (Saitou and Nei, 1987) as well as multi-species coalescent methods (Edwards et al., 2007; Liu and Pearl, 2007) to construct Bayesian phylogenies.

2.2 Workflow for linguistic study

The influx of computational algorithms and large amount of digitized linguistic data open up the new era in historical linguistics to process large-scale data efficiently as well as give a consistent

measurement in the aspect of language comparison (List et al., 2017b). We convert selected datasets into cross-linguistic data formats (CLDF) and follow the workflow below.

- Extract lexical items of target languages from individual linguistic studies by using custom Python scripts along with helper functions from the LingPy software package (List et al., 2017a).
- Making use of the Concepticon project's code and data (List et al., 2018), which links different elicitation glosses across more than 200 different questionnaires to unique identifiers (called *concept sets* in the project), we can easily identify which concepts are reflected across the datasets that we want to combine (List et al., forthcoming).
- By linking all data to Glottolog (Hammarström et al., 2017) where possible (proposing new Glottocodes for those varieties that are currently not yet considered in Glottolog), we can easily assemble and compare data for the same language variety from different sources.
- By coding our data in the format proposed by the Cross-Linguistic Data Formats initiative (Forkel et al., 2017), we make sure that our data is provided in a transparent format in which data curation and data comparison is transparent, replicable, and open for improvement.
- We develop and improve existing methods to compute various forms of linguistic distances, going beyond the "normal" level or cognacy by employing and testing methods for *partial cognate detection* (List et al., 2016) and improved annotation of linguistic data within the EDICTOR framework (List, 2017).
- We use standard approaches to compute phylogenetic trees from the data (Bayesian frameworks, distance-based frameworks, etc.) and explore alternative methods for future use, for example for ancestral state reconstruction (Jäger and List, 2018), working in close collaboration with experts on phylogenetic reconstruction from our department and other institutes.

2.3 Mapping genes and languages

Table 1 gives a preliminary list of ethnic groups along with genetic as well as linguistic data currently linked to them. Note that all these mappings are open for improvement, and that we hope that experts in the respective areas will help us to improve our data consistently.

Ethno-Linguistic Group	Language	Language Dataset	Genetic Dataset
CHB	Beijing	Liú, Lìlǐ 刘俐李 et al. 2007	Consortium 2009
Hakka	Meixian	Liú, Lìlǐ 刘俐李 et al. 2007	Consortium 2009
Cantonese	Guangzhou	Liú, Lìlǐ 刘俐李 et al. 2007	Consortium 2009
Han	Fuzhou	Liú, Lìlǐ 刘俐李 et al. 2007	Consortium 2009
	Suzhou	Liú, Lìlǐ 刘俐李 et al. 2007	
Tujia	Tujia	Sūn 1991	Lazaridis et al. 2014
Jinuo	Jinuo	Sūn 1991	Consortium 2009
Tibetan	Tibetan (Written)	Sūn 1991	Simonson et al. 2010
	Tibetan (Lhasa)	Sūn 1991	Yao et al. 2017
Naxi	Naxi (Lijiang)	Sūn 1991	Lazaridis et al. 2014
Karen	Karen	Huáng 1992	Consortium 2009
Yi	Yi (xide)	Sūn 1991	Lazaridis et al. 2014
Kusunda	Kusunda	Reinhard and Toba 1970	Lazaridis et al. 2014
	Kusunda linguistics		Lazaridis et al. 2014
Burmese	Burmese (Written)	Sūn 1991	1000 Genome project
Lahu	Lahu (Black)	Sūn 1991	Lazaridis et al. 2014
Ao Naga	Naga (Chungli)	Marrison 1967	Reich et al. 2009

Table 1: The selected ethno-linguistic populations and the datasets.

3 Chances and Challenges

Chances

- We try to establish a new framework to study to which degree language and genetic diversity correlate.
  - Previous studies on the co-evolution of genes and languages only investigated phoneme inventories and genetic diversities of ethnic groups (Creanza et al., 2015). Phoneme inventories, however, are an extremely bad predictor of language history (see, e.g., the discussions around the approach by Atkinson 2011).
  - If we want to fully explore to which degree languages and genes co-evolve, we need to assemble the best of available linguistic data and the best of available genetic data and analyze them with the best methods available in both fields. Following classical historical linguistics, we assume that lexical data is most indicative for language history and language diversity, although we try to code our data in such a way that we can search beyond the pure lexicon by investigating, for example, the distribution of sound correspondences across related languages with methods that we currently develop.
- By presenting our work in this very preliminary stage, where we do not have any initial results yet, we hope to draw linguists' attention to the importance of increasing the comparability of linguistic datasets when documenting the diversity of the world's languages. We also hope to instigate a debate on best methods and best practices when trying to identify ethnic groups with population genetic and linguistic data samples.

Challenges

- Increasing the comparability of linguistic data.
  - We need to work much harder on increasing the comparability of linguistic data. Data on Kusunda, an almost extinguished language can be taken as an example for current problems. We have found two resources from two separate studies (Donohue, 2012; *Kusunda linguistics*; Reinhard and Toba, 1970), however, we often found vastly different orthographies linked similar concepts. It is possible that the studies document different dialectal varieties of the language, or different historical stages, but we can hardly tell from the scarce descriptions we find in the literature or the sources themselves. For our study, we currently include all resources without merging them, hoping that future documentation can help to resolve these issues.
- Improving automatic methods for historical language comparison.
  - We need to work much on improved automatic methods for both language-internal and language-external comparison and search for cognates. Especially language-internal cognates (Hill and List, 2017) were so far largely ignored in recent studies (see Arnaud et al. 2017 for a recent approach), although automatic methods could be designed in a rather straight-forward way. Further challenges for automatic methods are the improvement of measures for linguistic distances, and an improved handling of phylogenetic reconstruction, especially in those cases where partial cognacy is predominant (List, 2016a).
- Improving explicitness of qualitative linguistic annotation.
  - We need to make classical linguistic work much more explicit than it is at the moment. Scholars know a lot about their data, but the way in which they share their findings is not amenable for formal analysis and therefore rarely machine-readable. The only way to address this problem is to start from existing examples and to try hard to improve the annotation of linguistic analyses that are so far most often written in prose form in numerous articles devoted to language history. But since historical linguistics is a rather formal endeavor, we think it should be possible for linguists to improve on this, and we will try to propose new methods for the annotation of important phenomena like sound correspondences, derivation patterns, and analogy in our future work.

Author contributions

This project is a multidisciplinary study involving experts with various backgrounds. To achieve the goal, it requires historical linguists with a deep understanding of the Sino-Tibetan language family, computational linguist who can help with the linguistic data preparation and the design of experiments, as well as bioinformaticians who collect genetic data and analyses to investigate the genetic structure of populations. MSW was responsible for the population genetic data preparation and analysis. MSW and JML were responsible for linguistic data preparation and automatic analyses. YFL helped in answering specific questions related to the Sino-Tibetan language family and its history. All authors wrote the draft and agreed on its final version.

Acknowledgements

This research was supported by the ERC Starting Grant 715618 "Computer-Assisted Language Comparison" (CALC). We sincerely thank Dr. Hugo Reyes-Centeno from the DFG Center for Advanced Studies "Words, Bones, Genes, Tools: Tracking Linguistic, Cultural and Biological Trajectories of the Human Past" (University of Tübingen), and Dr. Silvia Ghirotto from the Department of Life Science and Biotechnologies (University of Ferrara) for their help with population genetic analyses.

Statement

Data and code are under heavy construction and may change drastically during the next time. Nevertheless, for those interested in our initial implementations, you can check out the codes at GitHub under `https://github.com/LinguList/Peiros2004`. We kindly ask all people interested in using parts of the code or the data for their own studies to contact us before you submit any papers on it, so that we can make sure to give you proper instructions on how to quote our data and code, and also submit a draft version to a stable archive to guarantee that the data will be long-term archived.

References

Arnaud, Adam S., David Beck, and Grzegorz Kondrak (2017). "Identifying cognate sets across dictionaries of related languages". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. (Copenhagen, Sept. 7–11, 2017). Association for Computational Linguistics, pp. 2509–2518.

Atkinson, Quentin D (2011). "Phonemic diversity supports a serial founder effect model of language expansion from Africa". In: *Science* 332.6027, pp. 346–349.

Consortium, HUGO Pan-Asian SNP et al. (2009). "Mapping human genetic diversity in Asia". In: *Science* 326.5959, pp. 1541–1545.

Creanza, N. et al. (2015). "A comparison of worldwide phonemic and genetic variation in human populations". In: *Proc. Natl. Acad. Sci. U.S.A.* 112.5, pp. 1265–1272.

Donohue, Mark. *Kusunda linguistics*. Webpage. Online: `http://kusunda.linguistics.anu.edu.au`. 2013.

– (2012). "Kusunda Project-Kusunda-ANU". In:

Edwards, Scott V, Liang Liu, and Dennis K Pearl (2007). "High-resolution species trees without concatenation". In: *Proceedings of the National Academy of Sciences* 104.14, pp. 5936–5941.

Forkel, Robert et al. (2017). *CLDF. Cross-Linguistic Data Formats. Version 1.0*. Jena: Max Planck Institute for the Science of Human History. DOI: `10.5281/zenodo.1117644`. URL: `https://doi.org/10.5281/zenodo.1117644`.

Hammarström, Harald, Robert Forkel, and Martin Haspelmath (2017). *Glottolog*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: `http://glottolog.org`.

Hill, Nathan W. and Johann-Mattis List (2017). "Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages". In: *Yearbook of the Poznań Linguistic Meeting* 3.1, 47–76.

Holsinger, Kent E and Bruce S Weir (2009). "Genetics in geographically structured populations: defining, estimating and interpreting FST". In: *Nature Reviews Genetics* 10.9, p. 639.

Huáng, bùfán 黄布凡 (1992). *Zàngmiǎnyǔzú yǔyán cíhuì 藏缅语族语言词汇 [A Tibeto-Burman Lexicon]*. zhōngyāng mínzú xuéyuàn 中央民族学院出版社.

Jäger, Gerhard (2013). "Phylogenetic inference from word lists using weighted alignment with empirically determined weights". In: *Language Dynamics and Change* 3.2, pp. 245–291.

Jäger, Gerhard and Johann-Mattis List (2018). "Using ancestral state reconstruction methods for onomasiological reconstruction in multilingual word lists". In: *Language Dynamics and Change* 8.1, pp. 22–54.

Kang, Longli et al. (2012). "Y-chromosome O3 haplogroup diversity in Sino-Tibetan populations reveals two migration routes into the eastern Himalayas". In: *Annals of human genetics* 76.1, pp. 92–99.

Lazaridis, Iosif et al. (2014). "Ancient human genomes suggest three ancestral populations for present-day Europeans". In: *Nature* 513.7518, p. 409.

List, Johann-Mattis (2016a). "Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction". In: *Journal of Language Evolution* 1.2, pp. 119–136. DOI: `http://dx.doi.org/10.1093/jole/lzw006`. URL: `http://jole.oxfordjournals.org/content/1/2/119`.

– (2016b). *Computer-Assisted Language Comparison: Reconciling Computational and Classical Approaches in Historical Linguistics*. Tech. rep. Jena: Max Planck Institute for the Science of Human History.

– (2017). "A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*. Valencia: Association for Computational Linguistics, pp. 9–12.

List, Johann-Mattis, Philippe Lopez, and Eric Bapteste (2016). "Using sequence similarity networks to identify partial cognates in multilingual wordlists". In: *Proceedings of the Association of Computational Linguistics*. Vol. 2, pp. 599–605.

List, Johann-Mattis, Simon Greenhill, and Robert Forkel (2017a). *LingPy. A Python library for quantitative tasks in historical linguistics*. Jena: Max Planck Institute for the Science of Human History. DOI: `https://doi.org/10.5281/zenodo.1065403`. URL: `http://lingpy.org`.

List, Johann-Mattis, Simon J Greenhill, and Russell D Gray (2017b). "The potential of automatic word comparison for historical linguistics". In: *PloS one* 12.1, e0170046.

List, Johann-Mattis et al. (2018). *Concepticon. A Resource for the linking of concept list*. Jena: Max Planck Institute for the Science of Human History. URL: `http://concepticon.clld.org/`.

List, Johann-Mattis et al. (forthcoming). "CLICS²: An improved database of cross-linguistic colexifications. Assembling lexical data with help of cross-linguistic data formats". In: *Linguistic Typology* 22.2, ??–?? URL: `http://clics.clld.org`.

Liu, Liang and Dennis K Pearl (2007). "Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions". In: *Systematic biology* 56.3, pp. 504–514.

Liú, Lìlǐ 刘俐李, Wáng, Hóngzhōng 王洪钟, and Bǎi, Yíng 柏莹 (2007). *Xiàndài Hànyǔ fāngyán héxīncí, tèzhēng cíjí 现代汉语方言核心词·特征词集*. Nánjīng 南京: Fènghuáng 凤凰.

Longobardi, Giuseppe and Cristina Guardiano (2009). "Evidence for syntax as a signal of historical relatedness". In: *Lingua* 119.11, pp. 1679–1706.

Marrison, Geoffrey Edward (1967). "The classification of the Naga languages of north-east India." Accessed via STEDT database. PhD thesis. School of Oriental and African Studies (University of London).

Purcell, Shaun et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American Journal of Human Genetics* 81.3, pp. 559–575.

Reich, David et al. (2009). "Reconstructing Indian population history". In: *Nature* 461.7263, pp. 489–494.

Reinhard, Johan and Tim Toba (1970). *A preliminary linguistic analysis and vocabulary of the Kusunda language*. Summer Institute of Linguistics and Tribhuvan University.

Saitou, N. and M. Nei (1987). "The neighbor-joining method: A new method for reconstructing phylogenetic trees". In: *Molecular Biology and Evolution* 4.4, pp. 406–425.

Simonson, Tatum S et al. (2010). "Genetic evidence for high-altitude adaptation in Tibet". In: *Science* 329.5987, pp. 72–75.

Su, Bing et al. (2000). "Y chromosome haplotypes reveal prehistorical migrations to the Himalayas". In: *Human genetics* 107.6, pp. 582–590.

Sūn, Hóngkāi 孙宏开, ed. (1991). *Zàngmiǎnyǔ yǔyīn hé cíhuì* 藏缅语语音和词汇 *[Tibeto-Burman phonology and lexicon]*. Běijīng: Zhōngguó Shèhuì Kēxué 中国社会科学 [Chinese Social Sciences Press].

Torroni, Antonio et al. (1994). "Mitochondrial DNA analysis in Tibet: implications for the origin of the Tibetan population and its adaptation to high altitude". In: *American Journal of Physical Anthropology* 93.2, pp. 189–199.

Yao, Hong-Bing et al. (2015). "Genetic structure of Sino-Tibetan populations revealed by forensic STR loci". In: *arXiv preprint arXiv:1503.01880*.

Yao, Hong-Bing et al. (2017). "Genetic structure of Tibetan populations in Gansu revealed by forensic STR loci". In: *Scientific Reports* 7, p. 41195. DOI: 10.1038/srep41195. URL: https://doi.org/10.1038/srep41195.