Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

Computer-Assisted Comparison of Gelong and Hlai using Cross-Linguistic Data Formats

Fiona R. Grond and Ayten Tüfekci
Faculty for Linguistics and Literary Sciences
Bielefeld University

In this study, we discuss the sparsely studied Gelong language of Hainan Island and its affiliation to the Hlai languages. Our work is based on Andy Chin's article "The Gelong Language in the Multilingual Hub of Hainan". We extracted Chin's data and processed it with the help of various computer-assisted methods in order to make it more accessible, machine-readable, and comparable with other datasets.

1 Background

Our project is based on Chin's work from 2015 in which he compared the data he himself collected on the Gelong language in 2011/12 with that of Fu (1996) and Ouyang (1998) and thus examined the linguistic affiliation of the Gelong language and its relationship with the Hlai languages of Hainan, as well as any changes in the Gelong language that might have taken place since the 90s. In that second regard, he came to the result that no significant changes had taken place, but he found some cases of lexical variations where the Gelong vocabulary recorded in the 90s had been expanded or replaced with loanwords from Hainan Min, which is a neighboring language with higher status (it is typical for substrate varieties with lower socioeconomic status to be influenced by more prestigious superstrate languages).

According to Chin (2015: 144), the genetic affiliation of Gelong, spoken in a "complex multilingual community" in Dongfang city (ibid. 143) remains disputed among scholars. Although most classify it as a Tai-Kadai language with heavy influence from

Computer-Assisted Language Comparison in Practice
Volume 4, Number 7
https://calc.hypotheses.org/2827

Published on 05/07/2021

Chinese (Sun et al. 2007, Fu 1983), its place in the Hlai subgroup of Tai-Kadai remains unclear.

2 Materials

The data we worked with is based on Chin's table of "Basic Vocabulary in Gelong (on the basis of the 100 items in the Swadesh list)" (Chin 2015, Appendix 1 p.151-155). The table not only lists the 100 items by their English gloss and the corresponding Gelong words, but also the words in nine different Hlai varieties (Northern Hlai: Xifang, Baisha, Yuanmen, Southern Hlai: Heitu, and Central Hlai: Baoding, Zhongsha, Tongzha, Quiandui, and Baocheng). In a few cases, a specific concept does not seem to appear in one or more of the languages, or Chin could not find out the respective word; the concept KNOW for instance is missing in Baisha and Yuanmen, and BARK does not have corresponding words listed in any of the ten languages. In other cases, a language has more than one word for the same concept (for example there are two words for SUN in all but three of the languages).

We converted the raw data we extracted from Chin's article to a spreadsheet and pasted it into a Google spreadsheet document to allow easy access for multiple editors. The data had to be cleared up manually to account for the irregularities in the data set mentioned above as well as general problems that arose from the conversion into spreadsheet format. We added the header information for concepts where an additional row could be found because the data contained two separate words for a given concept, like it was the case with SUN). We also had to specifically control the morpheme boundaries by distinguishing carefully those cases where a word consists of two morphemes (in which case we wrote the word into the same cell, with the morphemes separated by a space) and those cases where there are two words (which we wrote into the same cell, separated by two slashes with an empty space before and after the slashes: "word // word") to create a comprehensible overview of the cleaned data (compare the raw/data.tsv sheet on GitHub).

3 Methods

In order to standardize our data we followed the suggestions made by the Cross-Linguistic Data Formats initiative (https://cldf.clld.org, Forkel et al. 2018). CLDF provides reference catalogues which offer standards for the identification of languages, concepts and phonetic transcriptions. Besides, CLDF also offers several software-API's for the quantitative analysis of linguistic data. In a first step on our way to standardisation we needed to link the given languages to Glottolog (Hammarstrom et al. 2021, https://glottolog.org) and map the concepts to Concepticon (List et al. 2021,

https://concepticon.clld.org). Afterwards our data was ready for the CLDF conversion which allows further processing in the form of automated analysis, such as partial cognate detection and phylogenetic reconstruction.

3.1 Language Linking

Our data consists of ten languages, out of which nine are different Hlai varieties. In order to link these languages to Glottolog (Hammarstrom et al. 2020) their Glottocodes needed to be looked up on the official website of the reference catalogue. Unfortunately, as of now, there are no Glottocodes for the different Hlai varieties, so we linked the varieties to the same standard Glottocode hlai1239. In order to distinguish the varieties, we provided the geographic locations of all varieties, including Gelong, although they are close to each other since they are all located on Hainan Island. All this information about the Hlai varieties was put into a CSV-file in preparation for the CLDF conversion. This information can be seen in Table 1. As the language affiliation of Gelong has yet to be determined, we could not provide a Glottocode for it. Figure 1 shows where the languages are located (according to our interpretation of Chin's data). Gelong itself is the left-most variety on the map. An interactive version of this map is available in form of the file languages.geojson on GitHub.

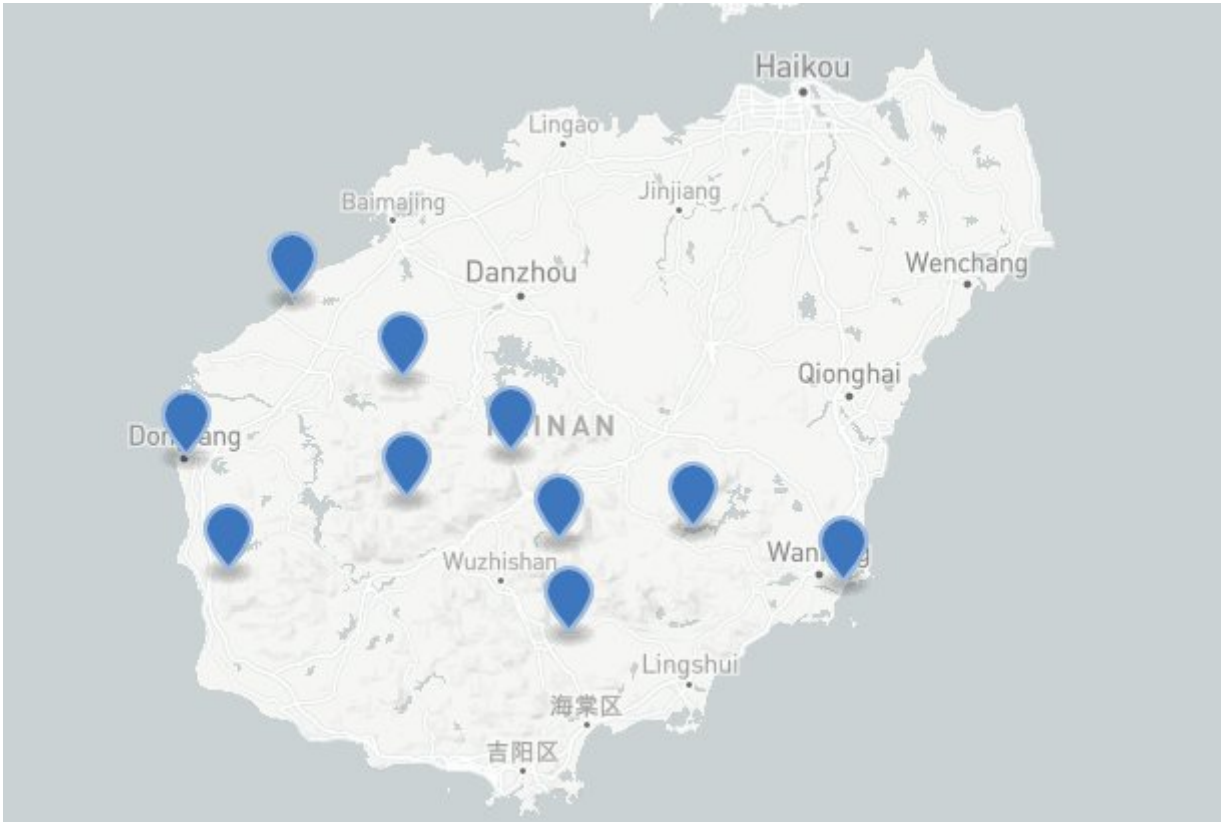

Figure 1: Geographic locations of the language varieties in Chin's sample.

3.2 Concept Mapping

The words in Chin's data are based on Swadesh's most famous concept list of 100 items (Swadesh 1955) and therefore can be easily mapped to Concepticon (List et al. 2021). Concepticon is a catalogue of concept sets which is essential for the aggregation of data. Although linking our data to the data provided by Concepticon could be done manually for only 100 concepts, we used the semi-automated workflow of linking concepts automatically first and correcting this later (see Tjuka 2020) for this purpose. Chin's concept list has by now already been included in version 2.5 of the Concepticon project, where it can be accessed under the ID Chin-2015-100.

3.3 CLDF Conversion

In our next step, we converted the raw data that we now have to the CLDF formats. The Cross-Linguistic Data Formats initiative (CLDF) seeks to provide standards for the identification, transcription and comparison of linguistic data to make them accessible for computer-aided processes. To transform our data, we used the cldfbench software for Python along with the lexibank-extension for wordlist data (https://github.com/lexibank/pylexibank). The major code that we used can be found on GitHub.

Our data was provided in the form of a phonetic transcription which did not completely conform to the CLDF guidelines, which follow the standards proposed by the Cross-Linguistic Transcription System initiative (https://clts.clld.org, Anderson et al. 2018). Thus, we needed to convert the source transcription into the target transcription. To do so, we created an orthography profile (Moran and Cysouw 2018), which is a replacement table that can be used to automatically convert the phonetic transcription in our source data into the target transcription used in CLDF, while at the same time segmenting the data (with boundaries between sound segments being represented by a space). The automatically created orthography profile needed to be manually checked and where necessary corrected (see the file etc/orthography.tsv on GitHub).

3.4 Partial Cognate Detection

Having converted our dataset into CLDF, we could now use the LingPy library (https://lingpy.org, List and Forkel 2021) in order to carry out an automated search for partial cognates in the data. For this purpose, we used the partial cognate detection algorithm by List et al. (2016), which offers to basic modes, one which searches for deeper signals by computing sound correspondences between all language pairs first, and one searching for surface similarities, based on the SCA algorithm for phonetic alignment analysis (List 2012, see List 2014 for details). Having applied this algorithm to the dataset, the results can be written to a TSV file which can in turn be inspected with

the web-based EDICTOR tool for manipulating wordlist data (https://digling.org/edictor/, List 2017).

3.5 Phylogenetic Analysis

In our final step we used the cognate sets for phylogenetic analysis by calculating the distances between languages through shared cognate percentages and computing phylogenetic trees directly from our data. Since we had run a partial cognate detection, we had to convert our data to "normal" cognates first. In this strict conversion, only those words which had partial cognates for all their morphemes were judged to be full cognates. There are several methods to compute phylogenetic trees from these full cognates. We used the UPGMA algorithm which is a statistical method for evaluating systematic relationships (Sokal and Michener 1958). It generates results that can directly be visualized as a language tree or saved as a distance matrix which can then be uploaded into further phylogenetic software. The distance matrix (see results/distances.dst on GitHub) shows all pairwise distances between all examined languages represented by a number between one and zero where a smaller indicates a higher similarity. We imported the distance matrix to SplitsTree (Huson 1998) to visualise our data with the help of the Neighbor-Net algorithm (Bryant and Moulton 2004), resulting in the unrooted phylogenetic network shown below (see Figure 2).

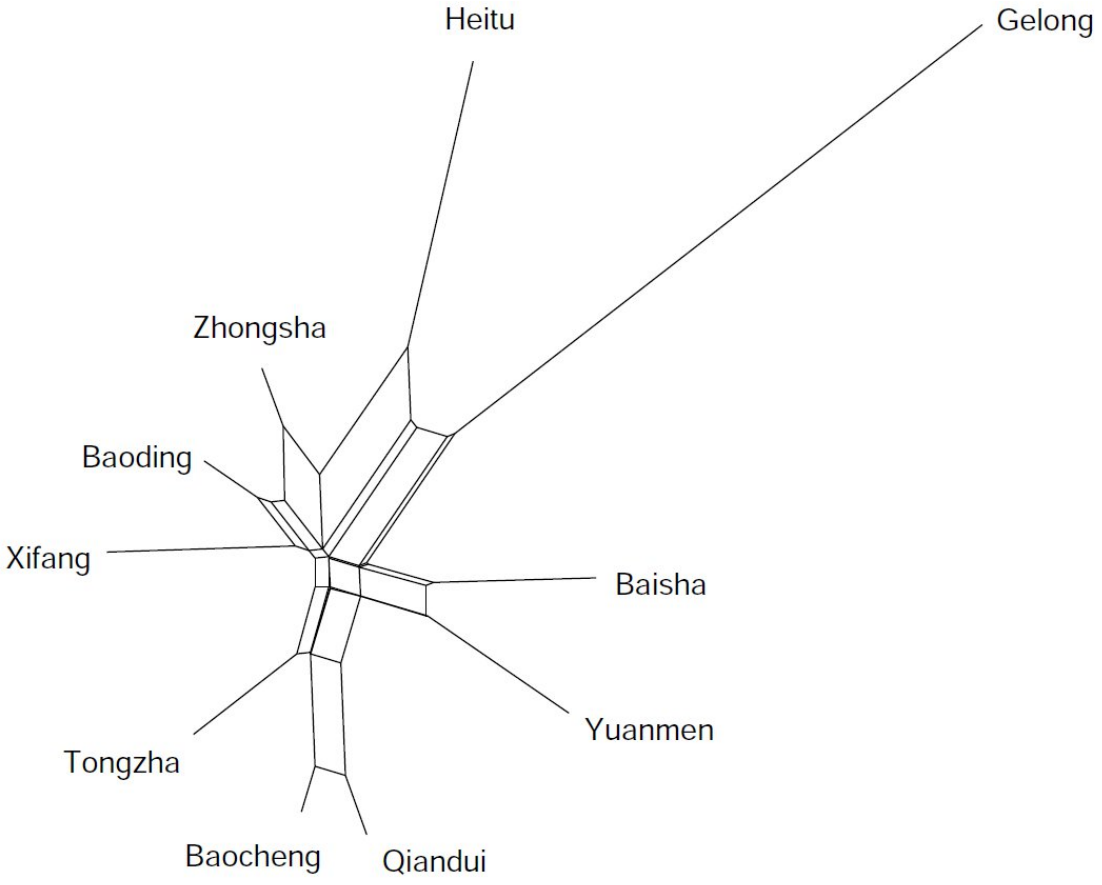

Figure 2: Phylogenetic network.

4 Results

Examining the results of our partial cognate analysis (see cognate_groupings.tsv on GitHub) it becomes apparent that roughly 40% of the examined Gelong vocabulary is cognate with all Hlai dialects, but around a quarter of the Gelong words are not related to Hlai at all. The Hlai languages among themselves share a lot more cognate sets, and we can see this already in the first dozen lines which compare the different words for the concept “I”. The words 1-10 are cognates with the same cognate ID (with the exception of 6, where Baoding has a second word for “I”) whereas 11—Gelong—has a totally different word. The existing cognates hint at a relationship between Gelong and the Hlai languages, though not a very close one.

We can investigate the relationships between the languages further by looking at our distance matrix: Gelong has a distance from 0.38 to 0.41 to the other languages, which is significantly larger than all other distances; in comparison, the greatest distance among the nine Hlai dialects to each other is 0.30, many are closer (between 0.10-0.20), the closest distance is only 0.05 between two members of the same sub-group.

We also see this distance visualized in the phylogenetic network (see Figure 2), that positions Gelong far away from the other languages which show a strong affiliation to each other. Out of the nine Hlai varieties, Heitu is the language which is the closest to Gelong, but still diverges greatly from it.

We can definitely say that Gelong is not as closely related to any of the Hlai languages as they are to each other. Maybe that means Gelong is a distant descendant from the proto-Hlai language which separated from the development of the other Hlai varieties fairly early. Alternatively, it might not be related to the Hlai languages at all and the occurring similarities can be ascribed to language contact. At this point, it seems still difficult to give a definite conclusion to any of the existing hypotheses.

5 Conclusion

Chin (2015) concludes that the Gelong language is rather closer related to Hlai than to the Sinitic languages. However, Gelong also possesses many unique features which may suggests an independent development from the Hlai languages. As our final analysis shows, we can neither confirm nor deny the affiliation of Gelong with the Hlai languages since the data we worked with cannot be considered representative. The data contains only 100 basic vocabulary items which is not sufficient to draw conclusions for an entire language. More research is needed not only in this regard, but also e.g. in comparing Gelong to Sinitic languages and other potential donor varieties spoken on Hainan Island.

Data and code discussed in this study have been curated on GitHub (https://github.com/digling/chingelong) and archived with Zenodo (https://doi.org/10.5281/zenodo.5040168).

References

Anderson, C., T. Tresoldi, T. Chacon, A.-M. Fehn, M. Walworth, R. Forkel, and J.-M. List (2018): A Cross-Linguistic Database of Phonetic Transcription Systems. Yearbook of the Poznań Linguistic Meeting 4.1. 21-53.

Bryant, D. and Moulton, V. (2004): Neighbor-Net. Molecular Biology and Evolution. 21.2: 255-265.

Chin, Andy C. (2015): The Gelong Language in the Multilingual Hub of Hainan. Bulletin of Chinese Linguistics. 8. 140-156.

Forkel, R. et al. (2018): Cross-Linguistic Data Formats, advancing data sharing and reuse in comparative linguistics. Sci. Data. 5:180205.

Forkel, Robert & List, Johann-Mattis (2020): CLDFBench: Give Your Cross-Linguistic Data a Lift. Proceedings of the 12th Language Resources and Evaluation Conference. 6995-7002. Marseille, France: European Language Resources Association.

Fu, Changzhong. 符昌忠. (1996):《海南村話》[The Cun language of Hainan]。 廣州：華南理工大學出版社.

Fu, Zhennan. 符鎮南. (1983): 〈海南島西海岸的村話〉 [The Cun language spoken in Western Hainan].《民族語文》4: 68–71.

Hammarstrom, Harald & Forkel, Robert & Haspelmath, Martin & Bank, Sebastian (2020): Glottolog 4.3. Jena: Max Planck Institute for the Science of Human History. DOI: 10.5281/zenodo.4061162, URL: http://glottolog.org

Huson, D. H. (1998): SplitsTree. Analyzing and visualizing evolutionary data. Bioinformatics 14.1. 68-73.

Huson, D. H. & Bryant, D. (2006): Application of Phylogenetic Networks in Evolutionary Studies. Mol. Biol. Evol. 23(2). 254-267.

List, J.-M., P. Lopez, and E. Bapteste (2016): Using sequence similarity networks to identify partial cognates in multilingual wordlists. In: Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers). Association of Computational Linguistics 599-605.

List, J.-M. (2012): SCA: Phonetic alignment based on sound classes. In: Slavkovik, M. and D. Lassiter (eds.): New directions in logic, language, and computation. Springer: Berlin and Heidelberg. 32-51.

List, Johann-Mattis (2014): Sequence Comparison in Historical Linguistics. Dusseldorf: Dusseldorf University Press.

List, Johann-Mattis (2017): A Web-Based Interactive Tool for Creating, Inspecting, Editing, and Publishing Etymological Datasets. Valencia, Spain: Proceedings of the EACL 2017 Software Demonstrations, April 3-7 2017. 9–12.

List, Johann-Mattis & Anderson, Cormac & Tresoldi, Tiago & Greenhill, Simon J. & Rzymski, Christoph & Forkel, Robert (2019): Cross-Linguistic Transcription Systems (Version 1.2.0). Jena: Max Planck Institute for the Science of Human History. URL: https://clts.clld.org

List, Johann-Mattis & Forkel, Robert (2021): LingPy. A Python library for historical linguistics. Version 2.6.8. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://lingpy.org

List, Johann-Mattis & Rzymski, Christoph & Greenhill, Simon & Schweikhard, Nathanael & Pianykh, Kristina & Tjuka, Annika & Hundt, Carolin & Forkel, Robert (2021): Concepticon 2.5.0. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: http://concepticon.clld.org

List, Johann-Mattis (2021): EDICTOR. A web-based interactive tool for creating and editing etymological datasets. Version 2.0.0. Max Planck Institute for Evolutionary Anthropology: Leipzig. URL: https://digling.org/edictor/.

Ouyang, Jueya. 歐陽覺亞. (1998):《村語》[The Cun language]。 上海: 遠東出版社。

Sokal, Robert. R. & Michener, Charles. D. (1958): A statistical method for evaluating systematic relationships. Kansas: University of Kansas Scientific Bulletin. 28. 1409-1438.

Sun, Hongkai 孫宏開, Hu, Zengyi 胡增益, Huang, Xing 黃行. (2007):《中國的語言》。 北京：商務印書館。

Tjuka, Annika (2020: Adding concept lists to Concepticon: A guide for beginners, in Computer-Assisted Language Comparison in Practice, 29/01/2020, https://calc.hypotheses.org/2225.