Automated phylogeny of Palaung dialects

Junsung Lee
Kent School, USA

Improved methods in automatic cognate detection have recently been used by historical linguists to help determine the subgrouping of a clade of languages or dialects, capitalizing on the efficiency of computers when handling substantial amounts of data. In this paper, 16 Palaung dialects are examined using various methods of automatic cognate detection. *Partial* (List, 2016) and *whole* cognate detection are used together with Lexstat and Sound Class Alignment to generate a phylogenetic tree of these dialects. The results of these methods are compared to the results using cognate detection decided by human experts (Deepadung et al., 2015). These results are substantially similar, suggesting that automatic cognate and phylogeny detection using algorithms is a viable complement to historical linguistic research. Accompanying this paper is a tutorial for the automated cognate detection and phylogeny procedure that was used. By following the steps, users can create results based on segmenting the morphemes of Palaung dialects differently.

1. Introduction

Automatic cognate detection has been often doubted because of its past inaccuracy. Especially, the use of edit distance, a measurement of how many characters of a word would need to be changed in order to turn it into another word, is judged to be shallow and inaccurate for figuring out the relationships among languages (Greenhill, 2011). Nevertheless, algorithms for automatic detection have developed apace in the past eight years, and now they can be of substantial help for historical linguists.

Baxter & Manaster Ramer (2000) applied a probabilistic approach to compare modern Hindi and modern English. They used a 33-word list of basic concepts and identified phonetic matches between the two languages. Dolgopolsky's sound classes (Dolgopolsky, 1964), which separate all phonetic segments into eleven different classes (or in this implementation, ten), were used for the initial segment of each word. To establish a baseline, a computer program paired the Hindi and English lists randomly; then, when each pair of concepts was coordinated, it found the probability that the minimal total number of matches, assuming they arose by chance. The suggestion made by the probabilistic approach is that a correlation between modern Hindi and modern English is significant at the p=0.02 level, meaning that modern Hindi and modern English matched better than random pairs of languages. This example demonstrates how a computerized approach could yield the same results as drawn by human experts.

Gray et al. (2009) used a Bayesian phylogenetic method and lexical data to model the origin and spread of the Austronesian languages; a phylogenetic tree was created for 400 languages. A penalized likelihood rate-smoothing approach was applied to estimate the time-depth of Austronesian languages. The generated phylogenetic tree supports the "Pulses and Pauses" theory by Gray et al. (2009), describing their origin and spread. The diversification of the Austronesian languages is also correlated with geographic expansions. Again, this is an example showing how automatic methods could assist human reasoning and theorizing.

Indeed, researchers have tried to determine whether there is a high congruence between traditional language subgroups and automated language phylogenies. For example, Greenhill et al. (2010) used automated language phylogeny on wordlists from several hundred Austronesian languages, and then researchers matched the results by comparing them with the results of the traditional method. The results, though not perfect, were promising, as there is substantial congruence between traditional language subgroups and the automated language phylogeny. Among the many methods under fast development, phylogenetic reconstruction via sequence comparison and pair group clustering is becoming one of the key methods in computational historical linguistics. This technique first uses alignment analyses to identify many potential cognates across sets of two or more languages. By utilizing a sufficient amount of data, this automatic method can identify not only potential cognates but probable cognates as well (List, 2012; Baxter & Manaster Ramer, 2000).

Cognate detection can be used to estimate how distant languages are, and these distances can be used to construct a phylogenetic tree. Comparing the cognate sets produced by the algorithm with those produced by human experts can indicate how close the program is capable of performing the same analysis that the researchers do. The trees created from two different methods can also be compared (Jäger & List, 2016). One database that is well-suited for testing the accuracy of a cognate detection algorithm is Concepticon, which provides the concept list used for comparison in the present paper. It is a database that contains various concept lists for different branches of linguistics studies, such as historical linguistics and psycholinguistics. Each concept is given a concept ID, a unique label, and a definition, and these can be used to keep track of relationships between concepts. Concepticon not only labels the concept lists but links the concept sets in a reliable way to enable the creation of new lists of concepts as well. Researchers can test the comparability between languages and link concept sets to enhance their research (List et al., 2016). A convenient database used in accordance with the Concepticon database is Glottolog. It is a repository of documents including grammar, wordlists, and phonological studies coordinated by dialect, language, and language family. After integrating with Glottolog, data will contain meta-linguistic information such as the longitude and latitude of the language and other details (Forkel & List, 2020; Hammarström et al., 2020).

In May 2020, I had been writing an automated phylogeny script from scratch when I came across the Lingpy tools created by the Max Planck Institute. These tools, Lexstat and SCA, were able to detect cognates to a certain extent automatically. I noticed that if I added rules to split words by syllable/morpheme, I could employ lingpy.compare's (a Python library) Partial class in conjunction with each of these methods, thus diversifying the methods that can be used to create phylogenetic trees. The goal of this paper is to compare the results given by these four automated cognate detection methods with each other and with manual cognate detection.

An online supplement to this paper at github.com/Juunlee/Palaung_scripts contains a tutorial for the automated cognate-based phylogeny procedure used in this paper. By following the steps, users can create results based on segmenting the morphemes of Palaung dialects differently. The scripts can also be adapted for analysis of other dialects.

2. Materials

The data set provided by Deepadung et al. (2015) contains each of the 100 concepts attested in 16 dialects of the Palaung dialect cluster. The 100 items are the first 100 of the comparative basic concept list created by Mann (2004). This concept list consists of the 118 concepts found in Matisoff (1978) and either Swadesh (1952) or Swadesh (1955). The list of the 100 concepts can also be found at the Concepticon (List et al., 2016).

As background information, Palaung, with more than half a million speakers in countries ranging from China to northern Thailand, is part of the Palaungic branch of the Austroasiatic (or Mon–Khmer) language family. The dialects in the data set can be classified into four subgroups according to the lexical comparison method used by Deepadung et al. (2015), which lines up well with endonyms and exonyms commonly used to refer to Palaung groups: Ta-ang, Pule, Dara-ang, and Rumai. The lone outlier, Cha Ye Qing, is a language spoken by a people commonly referred to as Rumai, but who call themselves Raokot instead (Deepadung et al., 2015).

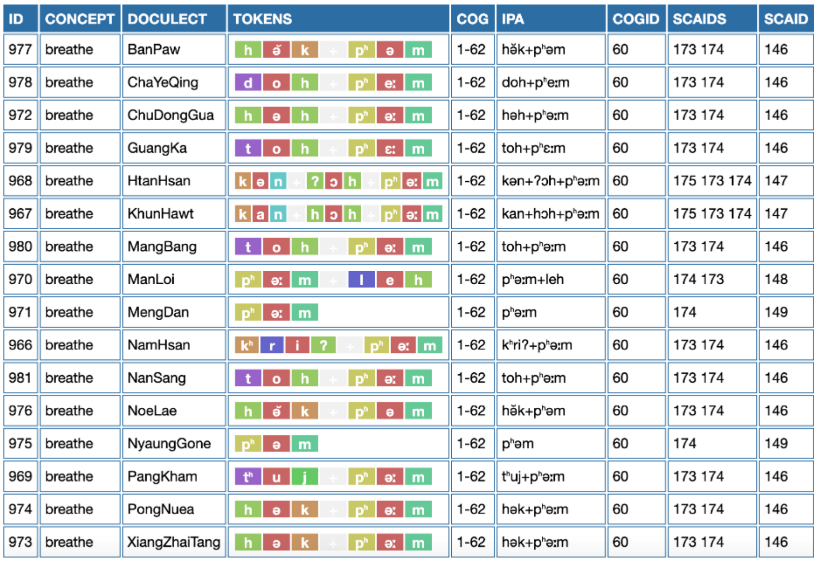
Ta-ang, which is often referred to as “Golden Palaung,” is a subgroup of Palaung spoken by people in the area of the northern Shan State. It contains dialects such as Nam Hsan, Khun Hawt, and Htan Hsan as constituents. Among them, Nam Hsan is recognized as a central dialect of this Palaung group. Pule dialects are found in the villages surrounding Shan state. Pule consists of Pang Kham and Man Loi in the south, and Meng Dan and Chu Dong Gua in the north. The Rumai subgroup consists of dialects found on the border of China and Myanmar. In Deepadung et al. (2015), Guang Ka, Mang Bang, and Nan Sang are classified as this subgroup. Lastly, Dara-ang is a subgroup whose core is spoken in a large diverse area in the south of the Palaung region. Dialects such as Ban Paw, Pong Nuea, Nyaung Gone, and Noe Lae are classified in this group, and Xiang Zhai Tang is a Dara-ang dialect spoken much further north, closer to Cha Ye Qing and northern Pule. (Deepadung et al., 2015)

3. Procedure

The methods outlined in Wu et al. (2020) for Hmong–Mien languages were applied to the Palaung dataset as mentioned above. First, to *tokenize* the data (split it up into sound segments), an *orthography profile*, as outlined in Wu et al. (2020), was used by the Cross-Linguistic Data Formats Bench (Forkel and List, 2020) on the raw data. The CLDF Bench uses CLTS, or Cross-Linguistic Transcription Systems (Anderson et al., 2018), to consolidate transcriptions of words done by different linguists.

Partial cognate detection is the process of comparing parts of words and deciding when those parts seem to be cognate (List et al., 2016). The lexicons of languages of the Southeast Asian linguistic area overwhelmingly tend to consist of words in which each syllable corresponds to a morpheme (Enfield, 2005). It is, therefore, appropriate to use Partial cognate detection for these languages since each syllable so often has an independent meaning.

Upon inspection of the word list, it became clear that the syllabic segments of each word are almost always morphemes and often seem to be related as units among multiple dialects (see Figure 1 for an example). These syllables of words seem to follow an i[m]n[c] template: they require an initial consonant (i), sometimes contain a medial glide (m), and always have a nucleus vowel (n), optionally followed by a coda consonant (c). The only kind of syllable that does not follow this i[m]n[c] template seems to be a lone nasal consonant.[1]



ID	CONCEPT	DOCULECT	TOKENS	COG	IPA	COGID	SCAIDS	SCAID
977	breathe	BanPaw	h ə̆ k + pʰ ə m	1-62	hə̆k+pʰəm	60	173 174	146
978	breathe	ChaYeQing	d o h + pʰ eː m	1-62	doh+pʰeːm	60	173 174	146
972	breathe	ChuDongGua	h ə h + pʰ əː m	1-62	həh+pʰəːm	60	173 174	146
979	breathe	GuangKa	t o h + pʰ ɛː m	1-62	toh+pʰɛːm	60	173 174	146
968	breathe	HtanHsan	k ə n + ʔ ɔ h + pʰ əː m	1-62	kən+ʔɔh+pʰəːm	60	175 173 174	147
967	breathe	KhunHawt	k a n + h ɔ h + pʰ əː m	1-62	kan+hɔh+pʰəːm	60	175 173 174	147
980	breathe	MangBang	t o h + pʰ əː m	1-62	toh+pʰəːm	60	173 174	146
970	breathe	ManLoi	pʰ əː m + l e h	1-62	pʰəːm+leh	60	174 173	148
971	breathe	MengDan	pʰ əː m	1-62	pʰəːm	60	174	149
966	breathe	NamHsan	kʰ r i ʔ + pʰ əː m	1-62	kʰriʔ+pʰəːm	60	173 174	146
981	breathe	NanSang	t o h + pʰ əː m	1-62	toh+pʰəːm	60	173 174	146
976	breathe	NoeLae	h ə̆ k + pʰ ə m	1-62	hə̆k+pʰəm	60	173 174	146
975	breathe	NyaungGone	pʰ ə m	1-62	pʰəm	60	174	149
969	breathe	PangKham	tʰ u j + pʰ əː m	1-62	tʰuj+pʰəːm	60	173 174	146
974	breathe	PongNuea	h ə k + pʰ əː m	1-62	hək+pʰəːm	60	173 174	146
973	breathe	XiangZhaiTang	h ə k + pʰ əː m	1-62	hək+pʰəːm	60	173 174	146

Figure 1 EDICTOR display of the concept "breathe" in 16 Palaung dialects. COGID: cognate detection by Deepadung et al. (2015). SCAIDS: partial cognate detection using Sound Class Alignment. SCAID: whole cognate detection using Sound Class Alignment.

Next, the Partial class (List et al., 2016) extending the Lexstat class (List, 2012) from the lingpy.compare Python library was used. This class attempts to automatically detect cognates among morphemes within words in the dataset, instead of allowing only *full* words to register as cognates. The result of partial cognate detection can be seen in the "SCAIDS" column of Figure 1.

In this paper, two methods—Sound Class-Based Phonetic Alignment (SCA) and Lexstat—are used to compare the concepts in languages.

SCA

SCA (List, 2010) is a comparative method that combines a specific sound class and sequence alignment methods to find the similarities among languages. Unlike the simple sound correspondence comparison, which cannot detect similar sounds as cognates, SCA can sort them as similar sounds according to their classification, making language comparison more accurate. For sequence conversion, input words are converted into default sound classes, and the sonority is calculated respectively (List, 2010). In Levenshtein distance, /k/ and /g/ will give the same distance as when /k/ and /b/ are compared. However, in SCA, /k/ and /g/ will be classified as in the same sound class, thus reducing the measured difference between the sounds to zero.

Next, *alignments* are used in a dynamic programming approach in which similar segments or sequences can be arranged in the same position using recursion and a matrix. If a segment in one of the two comparanda would interrupt an otherwise optimal alignment, a blank, represented by a gap character, is used in the other. In this "columnar form," by comparing the sequence, regular sound changes can often be identified by checking the frequency of the sound change. Alignment analysis makes the cognate judgment clearer and provides enough data for the linguist's decision.

[1] According to Mak (2012), syllable structure in Ta-ang dialects can consist of what is known as a pre-syllable and a main syllable. Pre-syllables are more constrained in their possibilities: their initial consonants must come from a more limited set, they can't have medial glides, their nuclear vowel is always reduced to schwa in articulation, and they can't have final consonants. However, the data provided by Deepadung et al. show that in order to consistently apply the concept of pre-syllables across even Ta-ang dialects, they would need to be allowed to have final nasals/glides. It would be an interesting direction for further research in automated Palaung phylogeny to implement the notion of pre-syllable and main syllable in the algorithm.

Automatic phonetic alignment can be divided into two sections: pairwise alignment and multiple alignment analyses. Parameters of pairwise alignment include scoring function, gap function, and alignment mode. The scoring function gives a certain number to the pair, which represents the relatedness of the segments. The gap function operates only when there is a gap in the alignment. Lastly, alignment mode determines which part of the word is compared since it is meaningless to compare the parts that are already not cognates. Multiple alignment analyses basically use pairwise comparison as a first step and organize data until all alignments are formatted in multiple alignments. Current implementations of SCA use multiple alignment analyses.

Lexstat

Lexstat is an automatic cognate detection approach that uses SCA for sequence comparison and then employs a statistical approach known as a *permutation test*. It creates an *expected* distribution of "cognates" based on randomly matching words for distinct concepts in two languages a certain number of times, called the number of *runs*, to establish a baseline. The *attributed* distribution is then computed by comparing words that have the same meaning in the different languages (List, 2012).

Automatic cognate detection is processed through two major steps: pairwise alignment and a tree- or network-based algorithm. Pairwise alignment is used to find the distance between pairs of words ranging from 0 to 1. There are three ways to calculate the distance: normalized edit distance, sound-class-based alignment method, and Lexstat. Lastly, the UPGMA algorithm (Sokal, 1958) is used to return a tree form based on the distance between the languages (List, 2014; List et al., 2018).

In Wu et al. (2020), Hmong-Mien language data sets are used to test the computer-assisted language comparison. Raw data was tokenized so that they can be put in the same meaning slot. Orthography profiles are recommended to increase the consistency of the transcription. Once the data is ready, these cognate sets are aligned phonetically to be compared. Partial cognacy is a common feature of many Southeast Asian languages. After the comparison within the same meaning of languages, words are aligned across the meaning to infer the cognates from them. The algorithm detects the similarity of phonetic segments of the languages, and by collecting these data for concepts, it creates a phylogenetic tree. This tree explains the genetic relationship or closeness among languages and gives insight into the subgrouping. This algorithm is applied to the Deepadung et al. (2015) dataset.

EDICTOR is a web-based tool created to help historical linguists to create, edit, analyze, and publish etymological datasets (List, 2017). In the EDICTOR display, three distinct methods are shown: cogid, scaids, and scaid. Here, cogid means cognate sets as determined by human experts, whereas the other two methods use automated cognate detection. The only difference between scaid and scaids is that scaids uses partial cognate detection. If one looks into the core of the code that calculates the distance of the words, scaids use part.partial-cluster, unlike scaid, which just uses part.cluster function. Two or three scaids are written in the diagram for each word since the word is segmented into multiple parts. The word 'breathe', as illustrated in Figure 1, is interesting because it gives multiple scaid values whereas the scaids values are more consistent. Furthermore, scaid and scaids will give different trees because of this variation.

The algorithm uses these cogid or scaid values to calculate the distance matrix between the languages. The simple Python function get_score first counts the set of words in which the two languages share the same ID. Then this number is divided by the size of the total set of words provided by the dataset. Lastly, this quotient is subtracted from 1 to create the distance between the two languages. By repeating this process, a distance matrix is formed and from this, the phylogenetic tree is generated by UPGMA.

4. Results

A B-cubed score is a calculation that is used to comprehend how well the program detects cognates between dialects as compared to the "gold standard," meaning the cognate judgments done by experts, which in this case are those found in Deepadung et al. (2015). The number ranges from 0 to 1, and a higher B-cubed score represents a better match between cognate sets generated by the program and cognate sets determined by the experts. To start this calculation, one should find for each word the size of the intersection between its cognate set as created by the gold standard and as created by the user's algorithm. Then B-cubed precision P and recall R are calculated by averaging the size of the intersection divided by the size of the cognate set from the user's algorithm and from the gold standard, respectively. By using the formula $2PR/(P+R)$, the B-cubed *F*-score is calculated, and by looking at the number, one can find the accuracy of the program (List, Greenhill, and

Gray, 2017). The *F*-scores of our scacogid algorithm is 0.8896, and the *F*-score of our lexstatcogid algorithm is 0.8805. Both of the programs seem to produce similar results compared to the gold standard. However, since there is barely a difference between these numbers, it is hard to determine which algorithm did a better job of cognate detection.

The detection of *partial* cognates by the SCA method, counting shared scacogids to generate a distance matrix, followed by UPGMA to construct a phylogenetic tree, yielded the following (Figure 2):

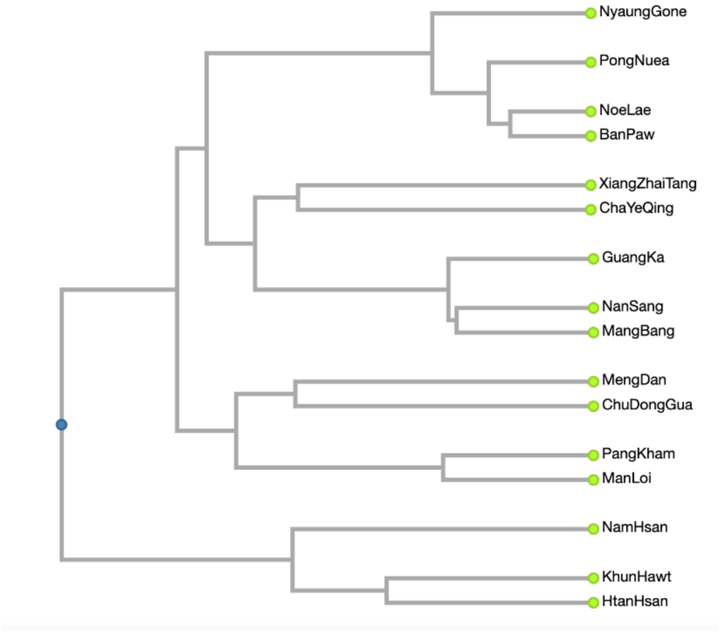


Figure 2 A scacogids tree, using the detection of partial cognates by the SCA method

The detection of *full* cognates by the SCA method, counting shared scacogid to generate a distance matrix, followed by UPGMA to construct a phylogenetic tree, yielded the following (Figure 3):

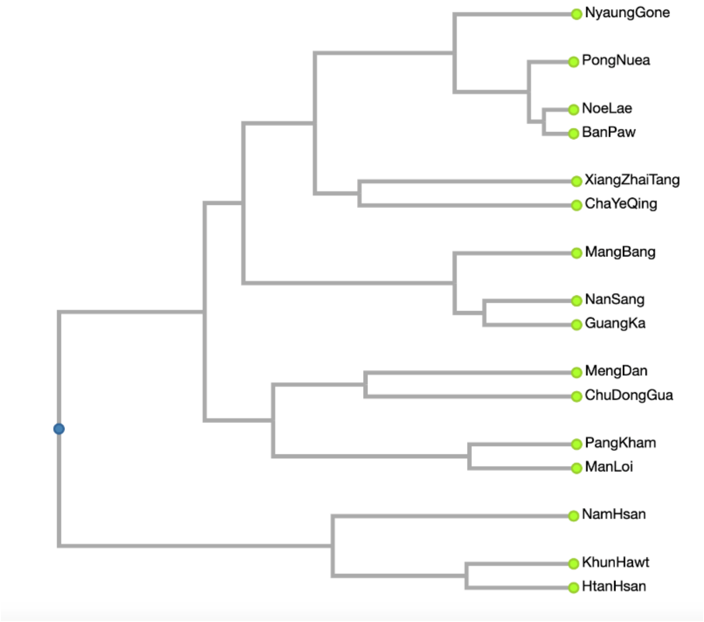


Figure 3 A scacogid tree, using the detection of full cognates by the SCA method

The detection of *partial* cognates by the LexStat method with 10000 runs, counting shared lexstatcogids to generate a distance matrix, followed by UPGMA to construct a phylogenetic tree, yielded the following (Figure 4):

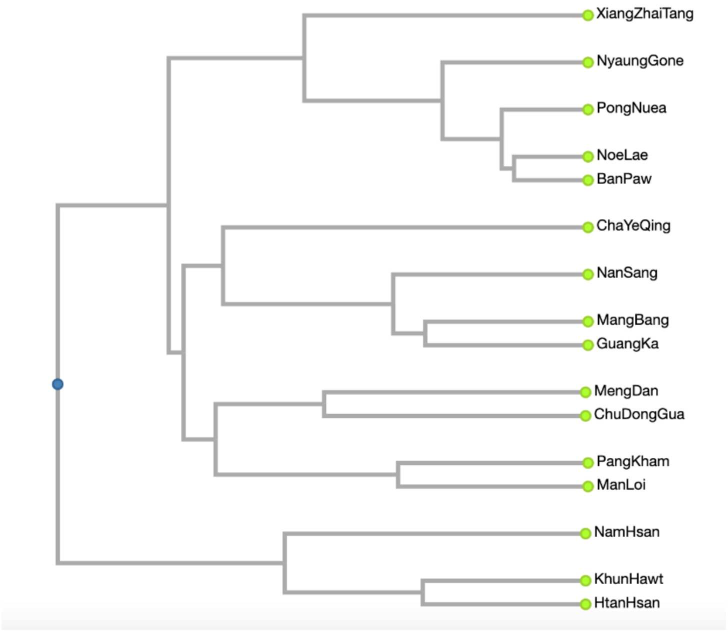


Figure 4 A lexstatcogids tree, using the detection of partial cognates by the LexStat method

The detection of *full* cognates by the LexStat method with 10000 runs to establish the baseline, and counting shared lexstatcogid to generate a distance matrix, followed by UPGMA to construct a phylogenetic tree, yielded the following (Figure 5):

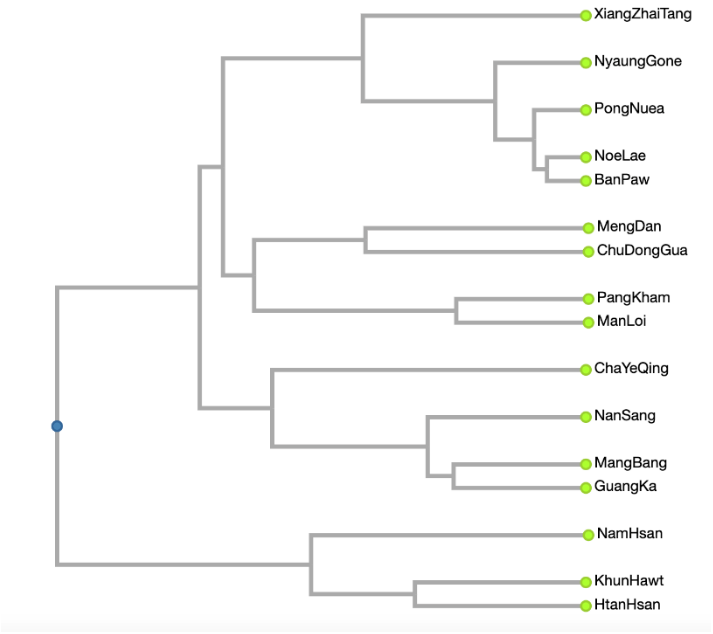


Figure 5 A lexstatcogid tree, using the detection of full cognates by the LexStat method

Counting shared "cog" from original cognate judgments as in Deepadung et al. (2015) to generate a distance matrix, followed by UPGMA to construct a phylogenetic tree, yielded the following (Figure 6):

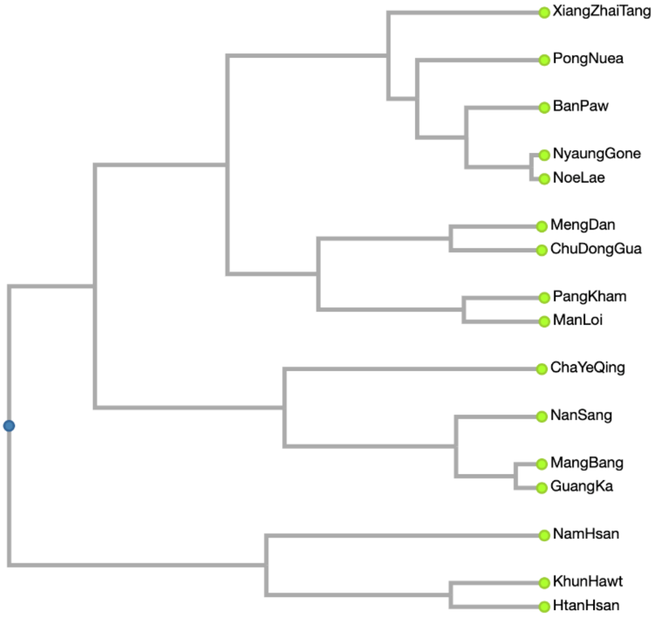


Figure 6 A cogid tree, based on the original cognate judgments as in Deepadung et al. (2015)

Of the five methods described above, lexstatcogid and cogid have shown the most similar results. They both sort out Ta-ang as the furthest dialect among other languages, and other dialects were clustered in the same group. The only disagreement is in the internal structure of the Dara-ang cluster. The lexstatcogid tree has Nyaung Gone as the furthest out of all core Dara-ang dialects and Ban Paw and Noe Lae with the smallest distance. The scacogids tree also has the exact same internal structure of core Dara-ang as lexstatcogid. The difference between the two trees, however, is that the scacogids tree has Cha Ye Qing and Xiang Zhai Tang clustered together against Rumai dialects. Xiang Zhai Tang, which was originally classified with Dara-ang by cogid, is thus separated from the Dara-ang dialects.

In addition, the scacogid and scacogids algorithms create almost the exact same tree. One difference is the placement of Cha Ye Qing and Xiang Zhai Tang. Whereas they are grouped together against Rumai in the scacogids tree, in the scacogid tree, Xiang Zhai Tang is closer to core Dara-ang. The only other difference is in the internal structure of Rumai.

5. Conclusion

Deepadung et al. (2015) used their expert judgments to determine the subgrouping of four Palaung dialect clusters: Ta-ang, Pule, Dara-ang, and Rumai. In this paper, four other methods were used for automatic cognate and phylogeny detection. The results were compared with the judgments of Deepadung et al. (2015). The tree of the LexStat whole cognate method resembled the traditional golden standard the most. The only difference between the lexstatcogid tree and the cogid tree was the internal structure of Dara-ang. The lexstatcogids tree was exactly the same as that for lexstatcogid except for the location of the Pule dialect in regard to Dara-ang.

Whether the method was partial or full, LexStat or SCA, the program was able to cluster 16 dialects into four consistent groups except for two dialects—Xiang Zhai Tang and Cha Ye Qing. The methods also consistently placed Nyaung Gone as the furthest out of core Dara-ang dialects. The internal structure of Dara-ang should be studied thoroughly with more words attested, as with just 100 concepts, the four different methods created slightly different results. Therefore, we conclude that the example of the Palaung dialects showed that automatic cognate and phylogeny detection using algorithms is a viable complement to historical linguistic research, worthy of further development.

Acknowledgments

I want to thank Dr. Johann-Mattis List of Max Planck Institute for the Science of Human History, who helped me with preparing data and guiding the direction of the paper. Special thanks to my mentor Daniel Briggs who supported me in the process of writing this paper.

Supplementary materials

The Python code that was used to generate these results and the tutorial that allows readers to apply it are available at: github.com/Juunlee/Palaung_scripts

References

Anderson, C., Tresoldi, T., Chacon, T., Fehn, A.-M., Walworth, M., Forkel, R., & List, J.-M. (2018). A cross-linguistic database of phonetic transcription systems. *In Yearbook of the Poznan linguistic meeting* (Vol. 4, pp. 21–53).

Baxter, W. H., & Manaster Ramer, A. (2000). Beyond lumping and splitting: probabilistic issues in historical linguistics. *Time depth in historical linguistics*, *1*, 167–188.

Deepadung, S., Buakaw, S., & Rattanapitak, A. (2015). A lexical comparison of the Palaung dialects spoken in China, Myanmar, and Thailand. *Mon-Khmer Studies*, *44*, 19–38.

Dolgopolsky, A. B. (1964). Гипотеза древнейшего родства языковых семей северной евразии с вероятностной точки зрения [A probabilistic hypothesis concerning the oldest relationships among the language families of Northern Eurasia]. *Вопросы языкознания*, *2*, 53–63.

Enfield, N. J. (2005). Areal linguistics and mainland Southeast Asia. *Annu. Rev. Anthropol.*, *34*, 181–206.

Forkel, R., & List, J.-M. (2020). CLDFBench: Give your cross-linguistic data a lift. In *Proceedings of the 12th language resources and evaluation conference* (pp. 6995–7002).

Gray, R. D., Drummond, A. J., & Greenhill, S. J. (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science, 323*(5913), 479–483.

Greenhill, S. J., Drummond, A. J., & Gray, R. D. (2010). How accurate and robust are the phylogenetic estimates of Austronesian language relationships? *PLoS One*, *5*(3).

Greenhill, S. J. (2011). Levenshtein distances fail to identify language relationships accurately. *Computational Linguistics*, *37*(4), 689-698.

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2020, October). *glottolog/glottolog: Glottolog database 4.3.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.4061162 doi: 10.5281/zenodo.4061162

Jäger, G., & List, J.-M. (2016). Statistical and computational elaborations of the classical comparative method.

List, J.-M. (2010). SCA: phonetic alignment based on sound classes. In *New directions in logic, language and computation* (pp. 32–51). Springer.

List, J.-M. (2012). LexStat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 joint workshop of LINGVIS & UNCLH* (pp. 117–125).

List, J.-M. (2014). *Sequence comparison in historical linguistics* (Unpublished doctoral dissertation). Düsseldorf University Press

List, J.-M. (2017). A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In *Proceedings of the software demonstrations of the 15th conference of the European chapter of the Association for Computational Linguistics* (pp. 9–12).

List, J.-M., Cysouw, M., & Forkel, R. (2016). Concepticon: A resource for the linking of concept lists. In *Proceedings of the tenth international conference on language resources and evaluation* (*LREC*'16) (pp. 2393– 2400).

List, J.-M., Greenhill, S. J., & Gray, R. D. (2017). The potential of automatic word comparison for historical linguistics. *PLoS One*, *12*(1), e0170046.

List, J.-M., Lopez, P., & Bapteste, E. (2016, August). Using sequence similarity networks to identify partial cognates in multilingual wordlists. In *Proceedings of the 54th annual meeting of the Association for Computational Linguistics* (*volume 2: Short papers*) (pp. 599–605). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P16-2097 doi: 10.18653/v1/P16-2097

List, J.-M., Walworth, M., Greenhill, S. J., Tresoldi, T., & Forkel, R. (2018). Sequence comparison in computational historical linguistics. *Journal of Language Evolution*, *3*(2), 130–144.

Mak, P. (2012). *Golden Palaung: A grammatical description*. Asia-Pacific Open Access Monographs A-PL 003. Canberra: Asia-Pacific Linguistics (SEAMLES).

Mann, N. (2004). *Mainland Southeast Asia comparative wordlist for lexicostatistic studies*. Chiang Mai, Department of Linguistics, Graduate School, Payap University.

Matisoff, J. A. (1978). *Variational semantics in Tibeto-Burman. the "organic" approach to linguistic comparison*. Institute for the Study of Human Issues.

Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.*, *38*, 1409–1438.

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts. With special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society, 96*(4), 452-463.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics, 21*(2), 121-137.

Wu, M.-S., Schweikhard, N. E., Bodt, T. A., Hill, N. W., & List, J.-M. (2020). Computer-assisted language comparison: State of the art. *Journal of Open Humanities Data, 6*(2), 1–14.