Check for updates

DATA NOTE

# REVISED Lexical data for the historical comparison of Rgyalrongic languages

[version 2; peer review: 3 approved]

Yunfan Lai [1,2], Johann-Mattis List [2,3]

[1]Trinity Centre for Asian Studies, The University of Dublin Trinity College, Dublin, Leinster, D02 PN40, Ireland
[2]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, 04103, Germany
[3]Multilingual Computational Linguistics, Universitat Passau, Passau, Bavaria, 94032, Germany

**v2** **First published:** 20 Jun 2023, **3**:99
https://doi.org/10.12688/openreseurope.16017.1
**Latest published:** 18 Oct 2023, **3**:99
https://doi.org/10.12688/openreseurope.16017.2

## Abstract

As one of the most morphologically conservative branches of the Sino-Tibetan language family, most of the Rgyalrongic languages are still understudied and poorly understood, not to mention their vulnerable or endangered status. It is therefore important for available data of these languages to be made accessible. The lexical data sets the authors have assembled provide comparative word lists of 20 modern and medieval Rgyalrongic languages, consisting of word lists from fieldwork carried out by the first author and other colleagues as well as published word lists by other authors. In particular, data of the two Khroskyabs varieties were collected by the first author from 2011 to 2016. Cognate identification is based on the authors' expertise in Rgyalrong historical linguistics through application of the comparative method. We curated the data by conducting phonemic segmentation and partial cognate annotation. The data sets can be used by historical linguists interested in the etymology and the phylogeny of the languages in question, and they can use them to answer questions regarding individual word histories or the subgrouping of languages in this important branch of Sino-Tibetan.

## Plain language summary

Rgyalrongic languages are mainly spoken in Western Sichuan, China, though Tangut, an extinct mediaeval language, was attested in today's Ningxia province and its surrounding regions from 1036 to 1502 AD. They are the most difficult branch of languages to learn in the Sino-Tibetan family, as they exhibit complex word formation strategies such as inflection and derivation. Their word complexity points to their historical depth in the language group and their value in exploring Sino-Tibetan language history. The database considered in this article aims at gathering lexical information of Rgyalrongic languages, as a tool for research in the field of historical and

**Open Peer Review**

**Approval Status** ✓ ✓ ✓

| | 1 | 2 | 3 |
|---|---|---|---|
| **version 2** (revision) 18 Oct 2023 | ✓ view | ✓ view | |
| **version 1** 20 Jun 2023 | ? view | ? view | ✓ view |

1. **Alexis Michaud**, LACITO - Langues et civilisations à tradition orale, Paris, France
2. **Mark Alves**, Montgomery College, Rockville, USA
3. **Norihiko Hayashi**, Kobe City University of Foreign Studies, Kobe, Japan

Any reports and responses or comments on the article can be found at the end of the article.

evolutionary linguistics.

**Keywords**
lexical data; historical linguistics; language phylogeny; Rgyalrongic; Sino-Tibetan; language subgrouping; partial cognate annotation; endangered languages




This article is included in the Horizon 2020 gateway.




This article is included in the Horizon Europe gateway.



This article is included in the Languages and Literature gateway.




This article is included in the European Research Council (ERC) gateway.



This article is included in the Linguistic Diversity collection.

**Corresponding author:** Yunfan Lai (khroskyabs@gmail.com)

**Author roles: Lai Y**: Conceptualization, Data Curation, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **List JM**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Software, Supervision, Validation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This research was funded by the European Union under the Horizon 2020 research and innovation program (ERC Starting Grant CALC, grant agreement number 715618, DOI: \url{https://doi.org/10.3030/715618}, awarded to JML) and under the Horizon Europe research and innovation program (ERC Consolidator Grant ProduSemy, grant agreement number 101044282, DOI: \url{https://doi.org/10.3030/101044282}, awarded to JML), and by the Irish Research Council under the SFI-IRC Pathway Programme (Project id: 21/path-a/9374, Gyalrongic unveiled: Languages, Heritage, Ancestry, awarded to Yunfan Lai). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2023 Lai Y and List JM. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Lai Y and List JM. **Lexical data for the historical comparison of Rgyalrongic languages [version 2; peer review: 3 approved]** Open Research Europe 2023, **3**:99 https://doi.org/10.12688/openreseurope.16017.2

**First published:** 20 Jun 2023, **3**:99 https://doi.org/10.12688/openreseurope.16017.1

**REVISED** **Amendments from Version 1**

We have revised our report according to the reviewers' comments, focusing on two key areas:

1) Enhancing Report Clarity:

In response to feedback from reviewers, we have improved the clarity and style of our report. Our efforts include refining language and style to ensure the report meets high standards of clarity. Key actions include:

• Methodology Enhancement: We've substantially improved the methodology section for cognate annotation by incorporating reliable, well-referenced sources, replacing vague reliance on "expert knowledge."

• Improved Examples: We replaced the inadequate Chinese example on partial cognate with a more representative one sourced directly from our database.

• Clearer Explanations: To enhance comprehension, we have provided more coherent and lucid explanations of our statistical analyses.

• Syntax Standardization: Language names within our database have been standardized for uniformity and easy reference.

• Visual Aid: We have added a new map to visually depict the geographical distribution of languages within our database.

2) Optimizing GitHub Repository:

To enhance user experience and reduce confusion, we have taken several actions regarding our GitHub repository:

• Data Reliability: We have completed and corrected the sources for all wordlists, ensuring data reliability.

• File Management: We meticulously removed outdated and extraneous files to streamline the repository and prevent user confusion.

• User-Friendly Interface: Users can now interact with data and cognate judgments easily through a convenient Edictor link, designed for precision in exploring the database's structure.

• Comprehensive Guide: We have refined the Readme file to serve as a comprehensive guide, providing valuable insights and instructions for users.

In summary, our revisions cover both the substantive and practical aspects of our work. Further efforts will also be made to provide users with an improved, user-friendly experience.

**Any further responses from the reviewers can be found at the end of the article**

## Introduction

Rgyalrongic languages form a Sino-Tibetan branch and are mainly spoken in Rngaba Tibetan and Qiang Autonomous Prefecture in Sichuan, China[1,2]. They belong to the Qiangic sub-branch of Burmo-Qiangic and are thus more, yet still remotely related to Lolo-Burmese languages than to other branches of Sino-Tibetan[3]. Apart from their modern varieties, which are mostly endangered or vulnerable, the extinct Tangut language has been recently recognized as a Rgyalrongic language[4]. Rgyalrongic languages are traditionally divided in two sub-branches, the east sub-branch and the west sub-branch. East Rgyalrongic is comprised of four main languages: Situ, Zbu, Japhug and Tshobdun, and West Rgyalrongic of three further sub-branches, Khroskyabs, Stau (aka. Daofu) and Tangut. Recent phylogenetic studies. However, show that Zhaba is also clustered in Rgyalrongic[3]. Thus, we have considered that language varieties closely linked to Zhaba, such as Queyu and Minyag (aka. Muya, Menya) and Zlarong spoken in the Tibetan Autonmous Region, to be Rgyalrongic languages in our study. These new additions to the Rgyalrongic group are provisionally termed "Peripheral Rgyalrongic" in the following.

Rgyalrongic is one of the most morphologically conservative branches in the Sino-Tibetan family, and it has a complex and, in our view, highly conservative morphological system that may give hints as to ancient features in the Sino-Tibetan language family. Therefore, understanding the history of Rgyalrongic languages is vital for the study of the evolution of Sino-Tibetan. Phlogenetic research on Rgyalrongic to provide dating information is thus an essential step towards this goal. Lexical data is the most accessible means to approach language phylogeny and has been proven to show accurate results in both Sino-Tibetan and other language families[3,5]. In order to infer the phylogenetic subgrouping of Rgyalrongic, a lexical dataset with high quality curation is indispensable. This database provides the first annotated resource for the phylogenetic analysis of Rgyalrongic languages. It contains lexical data from twenty varieties in East, West and Peripheral Rgyalrongic, as shown in the map in Figure 1 and Table 1.

## Methods

The workflow to build our database is illustrated in Figure 2. We started with the collection of raw data, collected from original fieldwork and from existing word lists. We then organised our raw data into a designed and curated word list. In the third step, we conducted data standardisation conforming to standards outlined by the Cross-Linguistic Data Formats Initiative. Finally, we identified and annotated cognate sets for individual morphemes, also known as *partial cognates*[6].

### Major sources of the dataset

The major sources of the dataset include original fieldwork from one of the authors of this study (YFL) and various colleagues who generously shared their lexical data, ie. word lists. The author's original field data involves two varieties of Khroskyabs, Siyuewu and Wobzi. All the vocabulary needed for this dataset was collected collected before 2017, prior to any of the research projects acknowledged in this paper. Fieldwork involved verbal exchanges with native speakers, requesting pronunciations of words and expressions.[1] An additional source of data was published dictionaries and word lists which were judged as reliable by the authors (see Table 1). These dictionaries and word lists typically contain word forms and translations in Chinese, French or English. Some of the

[1] We requested the native speakers' consent and conducted our survey only under their consent. The investigations did not involve any physical contact, thus did not raise relevant ethical issues.

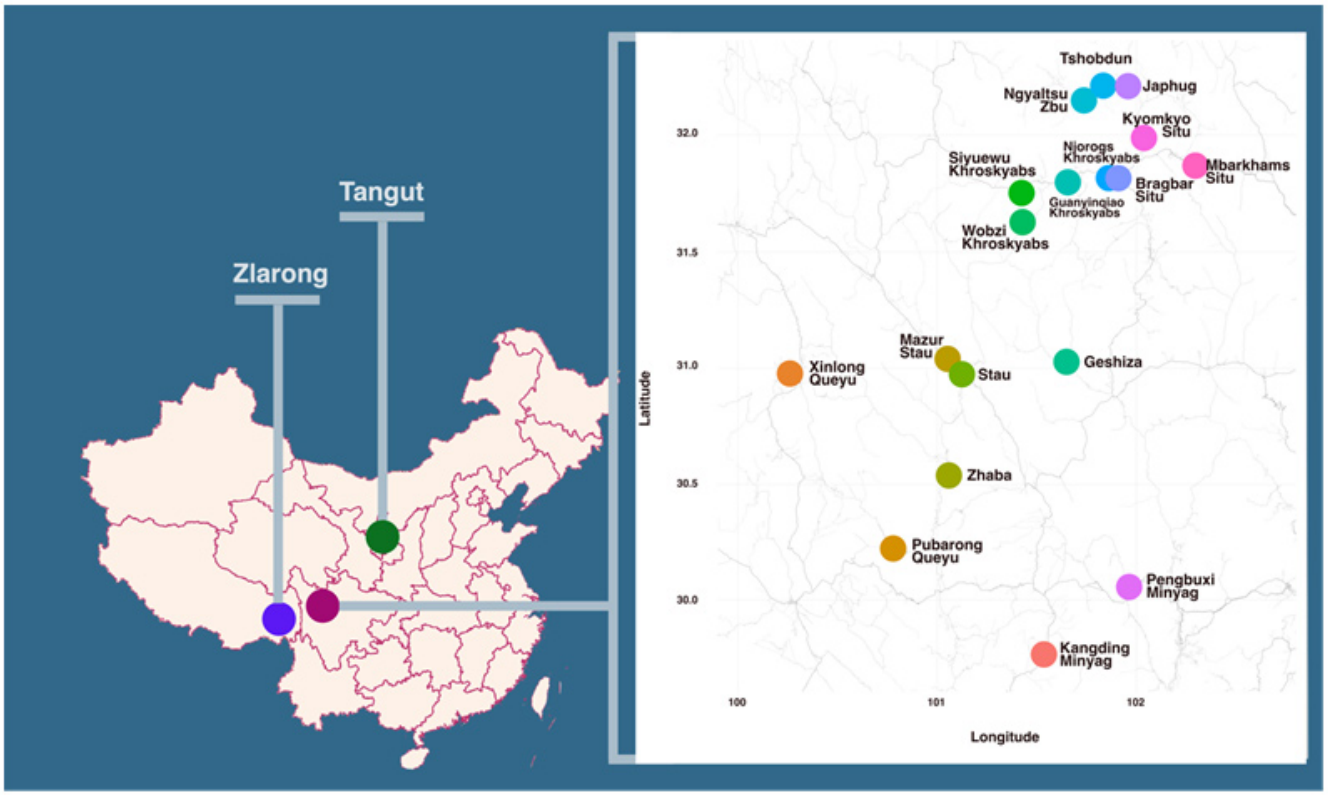


**Figure 1. Geographical distribution of the languages in the database.** Bantawa and Old Burmese, which are not Rgyalrongic languages but used as outgroups for phylogenetic analysis, are not shown on this map.

word lists also provide morphological information and example sentences. Reliability is assessed through two aspects: i) internal phonological consistency of the source data and ii) external regularity of sound correspondences with the comparative method. The authors check if phonemes are correctly identified and that allophones conditioned by phonological environments and morphological alternations are adequately represented in the original sources. In addition, the authors checked if cognate forms in the sources exhibit regular correspondences or correspondences that can potentially be explained through alternations or analogy. Languages in our dataset are listed with their sources, their Glottocodes[7] and approximate coordinates in Table 1. There are several cases where two or three languages share the same Glottocode (for the general idea behind Glottocodes, see Forkel and Hammarström[8]). Maerkang (MaerkangrGyalrong), Bragbar Situ (BragbarSitu) and Kyomkyo Situ (Kyomkyositu) share the glottocode "situ1238". However, they are distinct varieties of Situ with limited intelligibility. The two dialects of Minyag sharing the glottocode "west2417", labelled MuyaKangding and MenyaGao, are closely related dialects with minor differences. In contrast, Queyuxinlong and QueyuPubarong (quey1238) are closely related dialects with significant differences in phonology and vocabulary.

Apart from Rgyalrongic languages, we have included two outgroup Sino-Tibetan languages for the accuracy of phylogenetic inference: Bantawa and Old Burmese. The outgroup is used as a reference point to locate and root the ingroup (Rgyalrongic languages). Bantawa belongs to the Kiranti branch mainly spoken in Nepal. Old Burmese was an ancient Lolo-Burmese language attested between 12th and 16th century in present day Myanmar. These two languages are remotely related to Rgyalrongic. According to Sagart *et al.*[3], Kiranti languages branched off from other Sino-Tibetan subgroups approximately 5500 years from present, and Lolo-Burmese separated from Rgyalrongic some 4300 years from present. These two languages are suitable for outgroups in the present study, as Bantawa is sufficiently remote from Rgyalrongic, and Old Burmese has a clear date of attestation and can be used for the calibration of dating.

## Data presentation

An extended concept list based on the one used in Sagart *et al.*[3] is employed as a guideline of our word selection in each language, including 313 concepts linked to Concepticon[9] which provides a unique identifier to all concepts and thus facilitates language documentation and historical comparison of lexicon. The concept list used is specially designed for Sino-Tibetan languages. Therefore, it is most suitable as the starting point of the present dataset. According to our data quality and coverage, we made minor modifications to that concept list by adding and deleting some of the concepts. In

**Table 1. Languages selected for the database.**

| Languages ID | Language name | Glottolog | Longitude | Latitude | Source |
|---|---|---|---|---|---|
| Bantawa | Kiranti Bantawa | bant1281 | 87.05 | 27.12 | Doornebal (2009)[10] |
| Daofu | rGyalrong Daofu | horp1240 | 101.12 | 30.98 | Huang (1992)[11] |
| Japhug | rGyalrong Japhug | japh1234 | 101.96 | 32.21 | Jacques (2015)[12] |
| Tangut | Tangut | tan1334 | 106.29 | 38.48 | Li (1997)[13] |
| WobziKhroskyabs | Khroskyabs Wobzi | eree1240 | 101.43 | 31.63 | Lai (2017)[14] |
| Zhaba | Qiangic Zhaba | zhab1238 | 101.06 | 30.54 | Huang (1992)[11] |
| MaerkangrGyalrong | rGyalrong Maerkang | situ1238 | 102.30 | 31.87 | Huang (1992)[11] |
| MuyaKangding | Muya-kangding | west2417 | 101.96 | 30.06 | Huang (1992)[11] |
| MenyaGao | Menya-Gao | west2417 | 101.53 | 29.80 | Gao (2015)[15] |
| OldBurmese | OldBurmese | oldb1235 | 96.60 | 21.46 | Dictionary |
| QueyuXinlong | Queyu Xinlong | quey1238 | 100.26 | 30.98 | Huang (1992)[11] |
| QueyuPubarong | Queyu Pubarong | quey1238 | 100.78 | 30.22 | Guan Xuan's fieldwork |
| SiyuewuKhroskyabs | Siyuewu Khroskyabs | siya1242 | 101.42 | 31.75 | Author's fieldwork |
| BragbarSitu | Bragbar Situ | situ1238 | 101.91 | 31.82 | Zhang (2020)[16] |
| Geshiza | Geshiza | horp1240 | 101.65 | 31.03 | Honkasalo (2019)[17] |
| GuanyinqiaoKhroskyabs | Guanyinqiao Khroskyabs | guan1252 | 101.66 | 31.78 | Huang (2007)[18] |
| NjorogsKhroskyabs | Njorogs Khroskyabs | yelo1242 | 101.86 | 31.82 | Yin (2007)[19] |
| Tshobdun | Tshobdun | tsho1240 | 101.83 | 32.21 | Sun (2019)[20] |
| NgyaltsuZbu | Ngyaltsu Zbu | zbua1234 | 101.74 | 32.15 | Gong (2018)[21] |
| Zlarong | Zlarong | zlar1234 | 98.09 | 29.93 | Zhao (2019)[22] |
| KyomkyoSitu | Kymokyo Situ | situ1238 | 102.04 | 31.99 | Prins (2017)[23] |
| MazurStau | MazurStau | daof1238 | 101.05 | 31.04 | Gates (2021)[24] |

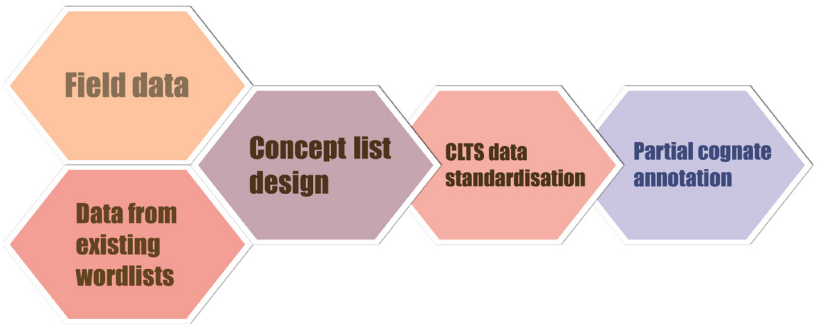


**Figure 2. Workflow of building up the Rgyalrongic lexical database.**

particular, we added concepts having a wide coverage in Rgyalrongic languages which are not widely distributed in other branches of Sino-Tibetan. For instance, we use the general concept for 'person, human' instead of 'the man (male human)' used in Sagart *et al.*[3], which has been shown to be indicative of language subgrouping by Lai[25]; we also included 'girl', as a

significant innovation in West Rgyalrongic with an *s*-prefix (compare Stau (West) *s-mi* and Japhug *me* (East)), discussed in Lai *et al.* [4, 177]. In addition, concepts such as 'knife', 'work' and 'sit' and so on also exhibit similar types of innovations across Rgyalrongic languages. We therefore consider them worth including in the dataset.

### Data standardisation

After collecting the raw word list of each language, we conducted a standardisation process of the data, because the original phonetic transcriptions may differ from each other, and some may not adhere strictly to the rules of the International Phonetic Alphabet. The revised transcriptions are based on the transcription conventions in Cross-Linguistic Data Formats reference catalog (CLTS, https://clts.clld.org, 6,26–28) and set up an orthography profile[29] that helped us automatically convert all transcriptions according to our standard. The standardised data aims specifically at the computation of language phylogeny.

### Partial cognate annotation

Cognates are words or part of words in different languages that share the same origin, such as English *foot* and German *Fuß*, both originating from Proto-Germanic **fōts*. Cognate forms in daughter languages can be deduced through regular sound rules from the proto-form. In Sino-Tibetan languages, more often than not, we find cognates in word parts in addition to those in entire words. As is shown in Figure 3, words for 'yesterday' across Rgyalrongic languages involves compounds with a part meaning 'past' and another meaning 'day'. There are two forms with distinct origins for 'past', one with a velar consonant (*x-* or *ɣ-*), and the other with a palatal consonant (*j-*); similarly, there are two etymologically unrelated forms for 'day', one with the nasal initial *sn-* or *n-* and the other with only *s-*. Different Rgyalrongic languages combine different partial cognates to form the word for 'yesterday'. Siyuewu Khroskyabs combines the velar *x-* for 'past' and the nasal *sn-* for 'day': *x-snə́*, Zhaba combines the palatal *ji:* for 'past' and the nasal *n-* to form *ji:-nə*; while Zlarong has the palatal *ji* for 'past' and the sibilant *si* for 'day': *ji-si*. Thus, Zhaba shares the palatal part for 'past' with Zlarong, and the nasal part for 'day' with Siyuewu Khroskyabs, while Siyuewu Khroskyabs shares no element with Zhaba. The identification of partial cognates enable us to segment full cognate forms into cognate morphemes, which improves the accuracy of the computation of language subgrouping along with full cognate identification. It is thus essential to annotate partial cognates, rather than full cognates, in our Rgyalrongic database. Partial cognate identification is conducted manually with the knowledge of the authors, using the web-based EDICTOR tool (https://digling.org/edictor, 30,31).

### Statistics

The current dataset contains a total of 6,335 word forms for 22 distinct language varieties, including 20 Rgyalrongic languages and two outgroup languages, namely Old Burmese and Bantawa. Word forms correspond to 305 different concepts, and use a total of 413 distinct speech sounds (i.e. consonants and vowels), with an average inventory of 72 different sounds per language variety. The word forms have been morphologically segmented, comprising a total of 9,116 morphemes. These morphemes have been assigned to 3,109 cognate sets. Of these cognate sets, 1,665 are unique sets with forms that exist in only one language.

| DOCULECT | CONCEPT | TOKENS | ID-7567 | ID-7579 | ID-7566 | ID-7615 |
|---|---|---|---|---|---|---|
| Wobzi | Yesterday | snəɣ | ɣ | snə | | |
| Tangut | Yesterday | jɨnjɨː | | njɨː | jɨ | |
| Zhaba | Yesterday | jɪnə | | nə | jɪ | |
| Muya | Yesterday | jisə | | | ji | sə |
| Siyuewu | Yesterday | xsnə | x | snə | | |
| Zlarong | Yesterday | jisi | | | ji | si |

**Figure 3. Partial cognate annotation: The concept 'yesterday' in Rgyalrongic is a compound of 'past' (cognates ID-7566 or ID-7567) and 'day' (cognates ID-7579 or ID-7615).** The word forms differ in terms of cognacy of different morphemes. Annotating the two parts separately allows us to visualise and analyse the internal morphology of the forms for more accurate data used for computational phylogenetic analysis.

### Quality control

We carefully verified the data to ensure the accuracy. We use our knowledge established through fieldwork and cross-linguistic comparison to review every lexical entry in the database. We searched for typos, misinterpreted phonemes, wrong entries and other issues in the sources. Whenever in doubt, we would contact the authors of the original sources for confirmation and correction.

Using an orthography profile, the transcriptions were converted according to a unified standard for potential reuse. Phonemic and morphemic segmentation, as well as cognate judgments, are carefully processed based on regular sound correspondences, phonological patterns of borrowings, educated guesses, as well as published cognate analyses such as 32–38. See Figure 3.

## Discussion and conclusion

Rgyalrongic languages are one of the most essential keys to the reconstruction of Proto-Sino-Tibetan as well as to the subgrouping of this language family[39,40]. Although there exist searchable databases such as STEDT[41] (https://stedt.berkeley.edu/) and the rGyalrongic Language Database[42] (https://htq.minpaku.ac.jp/databases/rGyalrong/), the present database is the first Rgyalrongic lexical database that involves data curation with historical linguistic considerations and cognate annotation, and the only one that is ready for phylogenetic analyses.

For now, only those morphemes assigned to the same cognate set occur in words sharing the same meaning. In the future, we hope to extend this analysis to account for cognates with the same meaning, specifically concentrating on language-internal partial cognates along the lines of the analysis pioneered in Hill and List[6] and further extended in Schweikhard and List[43]. Having annotated the data in this form, cognacy can also be annotated at the word level[44] and computational approaches to phylogenetic reconstruction of Rgyalrongic (and beyond) can be carried out. Thus, the present contribution may serve as the very base of future phylogeny of one of the most conservative sub-branches of Sino-Tibetan.

## Data and software availability

Data and Software available from: https://github.com/lexibank/lairgyalrong/releases/tag/v0.2.

Archived source code and data at time of publication: https://doi.org/10.5281/zenodo.8383011

License: Creative Commons Attribution 4.0 International license (CC-BY 4.0)

### Acknowledgements

Our sincere gratitude goes to Gao Yang, Jesse Gates, Gong Xun, Guan Xuan, Sami Honkasalo, Guillaume Jacques, Zhang Shuya and Zhao Haoliang for their generosity of contribution of data and detailed explanations in answer to our questions. We also thank Mei-Shin Wu for her help in refining the data.

## References

1. Sun JTS: **Stem alternations in Puxi verb inflection: Toward validating the rGyalrongic subgroup in Qiangic.** *Lang Linguistics.* 2000; **1**(2): 211–232.
Reference Source
2. Sun JTS: **Parallelisms in the verb morphology of Sidaba rGyalrong and Lavrung in rGyalrongic.** *Lang Linguistics.* 2000; **1**(1): 161–190.
Reference Source
3. Sagart L, Jacques G, Lai Y, *et al.*: **Dated language phylogenies shed light on the ancestry of Sino-Tibetan.** *Proc Natl Acad Sci U S A.* 2019; **116**(21): 10317–10322.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Lai Y, Gong X, Gates JP, *et al.*: **Tangut as a West Gyalrongic language.** *Folia Linguistica Historica.* 2020; **41**(1): 171–203.
Publisher Full Text
5. Gray RD, Atkinson QD: **Language-tree divergence times support the Anatolian theory of Indo-European origin.** *Nature.* 2003; **426**(6965): 435–439.
PubMed Abstract | Publisher Full Text
6. Hill NW, List JM: **Challenges of annotation and analysis in computer-assisted language comparison: A case study on burmish languages.** *Yearbook of the Poznan Linguistic Meeting.* 2017; **3**(1): 47–76.
Publisher Full Text
7. Hammarström H, Forkel R, Haspelmath M, *et al.*: **Glottolog [Dataset, Version 4.7].** Zenodo. Geneva, 2023.
Publisher Full Text
8. Forkel R, Hammarström H: **Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information.** *Semantic Web.* 2022; **13**(6): 917–924.
Publisher Full Text
9. List JM, Tjuka A, van Zantwijk M, *et al.*: **Concepticon [Dataset, Version 3.1].** Zenodo, Geneva, 2023.
Publisher Full Text
10. Doornenbal M: **A Grammar of Bantawa: Grammar, paradigm tables, glossary and texts of a Rai language of Eastern Nepal.** PhD thesis, Leiden University, 2009.
Reference Source
11. Huang B, Dai Q: **A Tibeto-Burman Lexicon.** Beijing: Central Minzu Univerisity, 1992.
12. Jacques G: **Dictionnaire Japhug-Chinois-Français.** Paris: Projet HimalCo, 2015.
13. Li F: **Xia-Han zidian [Tangut-Chinese dictionary].** Beijing: China Social Sciences Press, 1997.
14. Lai Y: **Grammaire du khroskyabs de Wobzi.** Paris: Université Sorbonne Nouvelle (Paris 3) dissertation, 2017.
Reference Source
15. Gao Y: **Description de la langue menya: phonologie et syntaxe.** PhD thesis, École des hautes études en sciences sociales, 2015.
Reference Source
16. Zhang S: **Le rgyalrong situ de Brag-bar et sa contribution à la typologie de l'expression des relations spatiales: l'orientation et le mouvement associé.** PhD thesis, Institut National des Langues et Civilisations Orientales, 2020.
Reference Source
17. Honkasalo S: **A grammar of Eastern Geshiza.** PhD thesis, University of Helsinki, 2019.
Reference Source
18. Huang B: ***Lawurongyu yanjiu* [Study on the Lavrung language].** Beijing: Nationalities Press, 2007.
Reference Source
19. Yin W: ***Yelong Lawurongyu Yanjiu* [Study on the 'Jorogs Lavrung language].** Beijing: Nationalities Press, 2007.
20. Sun JTS: **Tshobdun Rgyalrong Spoken Texts With a Grammatical Introduction.** Taipei: Academia Sinica, 2007.

21. Gong X: **Le rgyalrong zbu, une langue tibéto-birmane de Cine du Sud-ouest. Une etude descriptive, typologique et comparative.** PhD thesis, Paris: Institut National des Langues et Civilisations Orientales, 2018.
**Reference Source**

22. Zhao H: **A sketch of zlarong the newly discovered language: its phonology, lexicon, morphology and genealogy.** Guanghou: Sun Yat-Sen University master's dissertation, 2019.

23. Prins M: **A Grammar of rGyalrong, Ji˘aomùzú (Kyom-kyo) Dialects: A Wen of Relations.** Brill, 2017.

24. Gates JP: **Grammaire du stau de Mazur.** Paris: Écoles des Hautes Études en Sciences Sociales dissertation, 2021.
**Reference Source**

25. Lai Y: **Preinitial denasalisation and palatal fortition in Khroskyabs and the Gyalrongic word for 'man'.** *Linguistics of the Tibeto-Burman Area.* 2022; **45**(2): 211–229.
**Publisher Full Text**

26. Forkel R, List JM, Greenhill SJ, *et al.*: **Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics.** *Sci Data.* 2018; **5**(1): 180205.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

27. List JM, Hill NW, Foster CJ: **Towards a standardized annotation of rhyme judgments in Chinese historical phonology (and beyond).** *Journal of Language Relationship.* 2019; **17**(1–2): 26–43.
**Publisher Full Text**

28. Anderson C, Tresoldi T, Chacon T, *et al.*: **A cross-linguistic database of phonetic transcription systems.** *Yearbook of the Poznan Linguistic Meeting.* 2018; **4**(1): 21–53.
**Reference Source**

29. Moran S, Cysouw M: **The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles.** Language Science Press, 2018; 145.
**Reference Source**

30. List JM: **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.* 2017; 9–12.
**Publisher Full Text**

31. List JM: **EDICTOR. A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets [Software Tool, Version 2.0.0].** Zenodo, Geneva, 2021.
**Publisher Full Text**

32. Jacques G: **Phonologie et morphologie du japhug (rGyalrong).** Paris: Université Diderot - Paris 7 dissertation, 2004.
**Reference Source**

33. Jacques G: **Esquisse de phonologie et de morphologie historique du tangoute.** Brill, 2017.
**Publisher Full Text**

34. Zhang S, Jacques G, Lai Y: **A study of cognates between Gyalrong languages and Old Chinese.** *Journal of language relationship.* 2019; **17**(1–2): 73–92.
**Publisher Full Text**

35. Lai Y: **When internal reconstruction goes further: proposing the vowel system of Pre-Khroskyabs through examining bound state apophony.** *Folia Linguistica Historica.* 2022; **56**(s43–s1): 213–261.
**Publisher Full Text**

36. Lai Y: **Variations étymologiques sur l'étymon sino-tibétain 'étau, pinces'.** *Bulletin de la Société de Linguistique de Paris.* 2022; **117**(1): 297–312.
**Publisher Full Text**

37. Lai Y: **On plosive-nasal correspondences and alternations in Gyalrongic and their possible solutions.** *Cahiers de Linguistique Asie Orientale.* 2023; **52**(1): 1–39.
**Publisher Full Text**

38. Lai Y: **Lenition alternation in West Gyalrongic and its implications for Southeast Asian panchronic phonology.** *Diachronica.* 2023; **40**(3): 341–383.
**Publisher Full Text**

39. Jacques G: **Agreement morphology: the case of Rgyalrongic and Kiranti.** *Language and Linguistics.* 2012; **13**(1): 83–116.
**Reference Source**

40. Hill NW: **The Historical Phonology of Tibetan, Burmese, and Chinese.** Cambridge: Cambridge University Press. 2019.
**Publisher Full Text**

41. Richard S, John B: **The Sino-Tibetan Etymological Dictionary and Thesaurus: STEDT project data resources and protocols.** University of California, Berkeley, 2000.
**Reference Source**

42. Nagano Y, Prins M: **rGyalrongic Languages Database.** National Museum of Ethnology, Osaka, 2013.
**Reference Source**

43. Schweikhard NE, List JM: **Developing an annotation framework for word formation processes in comparative linguistics.** *SKASE Journal of Theoretical Linguistics.* 2020; **17**(1): 2–26.
**Reference Source**

44. Wu MS, List JM: **Annotating cognates in phylogenetic studies of southeast asian languages.** *Language Dynamics and Change.* 2023; 1–37.
**Publisher Full Text**

# Open Peer Review

## Current Peer Review Status: ✔ ✔ ✔

**Version 2**

Reviewer Report 30 October 2023

https://doi.org/10.21956/openreseurope.18034.r35685

© **2023 Alves M.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Mark Alves**
Montgomery College, Rockville, Maryland, USA

I have read over the authors' amended version and checked with my original comments. It seems suitable to me at this point for "Approved" status.

***Competing Interests:*** No competing interests were disclosed.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 18 October 2023

https://doi.org/10.21956/openreseurope.18034.r35683

© **2023 Michaud A.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Alexis Michaud**
LACITO - Langues et civilisations à tradition orale, Paris, France

Version 2 addresses the concerns brought up, making good progress in terms of clarity of the online data set.

***Competing Interests:*** No competing interests were disclosed.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

**Version 1**

Reviewer Report 31 July 2023

https://doi.org/10.21956/openreseurope.17295.r33646

© **2023 Hayashi N.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Norihiko Hayashi**
[1] Kobe City University of Foreign Studies, Kobe, Japan
[2] Kobe City University of Foreign Studies, Kobe, Japan

First of all, I would like to express my excitement over the recent release of a new database for the rGyalrongic (‘Rgyalrongic’ spelled in this paper) languages. However, I must clarify that I am not an expert in the rGyalrongic languages; rather, my specialization lies in Lolo-Burmese languages. Moreover, I have limited familiarity with statistical methods. Therefore, please be kindly notified that my review may incorporate some perspectives from classical historical linguistics.

As stated in the paper, this database is a compilation of lexical data from nearly 20 languages classified as rGyalrongic. The authors themselves conducted research on two of the languages, while the remaining data either underwent careful analysis by experts in the respective languages or were extracted from dictionaries with detailed word annotations. The finalized database has been published on their website with the cooperation of these experts.

The rGyalrongic languages, as the authors point out, exhibit an exceptionally morphologically complex lexicons among the Tibeto-Burman languages. This complexity not only makes it challenging to explore their historical development but also hinders the investigation of historical relationships among these languages. Until now, historical linguists familiar with these languages have relied on comparative linguistics methods and their own specialized judgment to classify the historical development and phylogeny based on morphological analysis of each language's vocabulary. However, this approach naturally has its limitations.
This paper introduces a system that mitigates assumptions about the historical development of lexicons in each language by incorporating new technologies, search techniques, and even statistical methods. It unveils new possibilities for comparative studies. To achieve this, the paper aims to create a database with maximum scientific transparency, relying on a scrupulous selection of data sets and reliable information provided by experts proficient in each language.

While it has its merits, it also has its drawbacks. To begin with, the attempt is still in its infancy in terms of its overall goal, so one can only hope for further development. The data for rGyalrongic languages is still very limitedly employed. Of course, I comprehend that the principal objective of this project is to meticulously curate and present data in collaboration with experts. However, it is still important to increase the amount of data regarding the future. The author himself mentions that a database of the rGyalrongic language, supervised by Nagano & Prins (2013), is available on the website, where the basic vocabulary and examples of sentences of 81 rGyalrongic languages are uploaded with their sound files. Indeed, similar to the STEDT database, it is essentially a

collection of data lacking morphological analysis. However, in the future, with the cooperation of the supervisors, it would be possible to use the data from Nagano & Prins (2013) to conduct a more precise analysis of historical change.

It should be noted that five of the languages in the dataset (Daofu, Zhaba, MaerkangrGyalrong, MuyaKangding, and QueyuXinlong) are taken from Huang (1992). Huang (1992) was a groundbreaking work at the time, a comparative lexicon of primary sources of Tibeto-Burman languages in China, which is very useful for historical and comparative studies. Unfortunately, Huang (1992) contains many printing errors, and the researcher, Prof. Huang Bufan, is deceased, thus we cannot ask her about her research. There are researchers, such as Satoko Shirai and Huang Yang on Zhaba (Gong 2007 is also a reference grammar of Zhaba), and Takumi Ikeda on Muya. For Situ rGyalrong, Yasuhiko Nagano also recently published a grammar in English (Nagano 2022). With the cooperation of these researchers who are still active in the field, there should be room for additional data and comparative studies in the future.

The paper presents the partial cognate annotation by exemplifying a word ‘yesterday’ in Figure 3, which this paper considers is a compound of ‘day’ and ‘past’. This figure leads us to think that /ɣ/ in the Wobzi word /snə**ɣ**/ corresponds to /x/ in the Siyuewu word /**x**snə/ and both /ɣ/ and /x/ denote the concept ‘past’. This is really inspiring, and if we search the same word in Nagano & Prins 2013 database, there are many cases similar to {pə/ ma/ wa}+{snə}, like /pəʃe’ʃni/ (Jinchuan Jimu Zhouchan), /ma sn̥i lo/ (Maerkang Soman), /mbasni’lo/ (Hongyuan Shuajinsi Selong), /wə$^{3}$ʂo$^{3}$ʃnə$^{5}$/ (Jinchuan Manai Genza), etc. The initial syllable of these words might indicate 'past,' which should be noted in the Rangtang Puxi Siyaowucun word /**ɣ**snə/, having the same meaning as 'yesterday'.

Additionally, Nagano (2022) also provides a few intriguing expressions for ‘yesterday’, warranting a comparative analysis. Collaborating with experts and integrating data from other databases could significantly enhance the development of this project.

Regarding the dataset of this project, while Bantawa and Old Burmese, languages that are quite distant historically from rGyalrongic languages, are included, Written Tibetan and Amdo Tibetan are not included in this dataset. Of course, this may be because there are many cases of Tibetan languages providing loanwords to rGyalrongic, but this situation may require some explanation. As for Old Burmese, it certainly has a phonetic writing system and the date of use is clear, so it is possible that the historical relationship between the two languages can be put on an absolute time scale. However, if the Bantawa data are to be combined, it seems to me that more Qiangic languages should be introduced, including Pumi and various Qiang dialects.

At any rate, as mentioned earlier, this database will continue to develop, and it is expected that the data will be updated in the future. At the moment, this database still needs more user-friendly interfaces which are already presented as in the STEDT and Nagano & Prince’s databases. However, at the same time, if researchers of Tibeto-Burman languages apply the methodology of this project and develop datasets for each subgroup, it will be possible to discover new problems of historical change and to solve existing problems. At the very least, if the Lolo-Burmese dataset, which is the reviewer's specialty, is developed in a similar manner, it will allow for a more scientific analysis of the relationships between languages within Lolo-Burmese (which may modify the framework of Matisoff 1972, 2003, and Bradley 1979). It will also lead to a more detailed elucidation of the relationship between rGyalrongic and Lolo-Burmese, which was discussed here, than has been the case so far. We are convinced that this project deserves further attention.

**References**
1. Bradley D: Proto-Loloish. *Malmö: Curzon Press*. 1979.
2. Gong Q: Zhaba-yu yanjiu. [Study of the Zhaba Language. (in Chinese)]. *Beijing: Minzu chubanshe*. 2007.
3. Zangmian yuzu yuyan cihui. [A Tibeto-Burman Lexicon. (in Chinese)]. *Beijing: Zhongyang minzu xueyuan chubansh*. 1992.
4. Matisoff JA*Berkeley: University of California Press*. Handbook of Proto-Tibeto-Burman.
5. Matisoff JA: Handbook of Proto-Tibeto-Burman. *Berkeley: University of California Press*. 2003.
6. Nagano Y: rGyalrong: A Comprehensive Grammar. *Tokyo: Fukyosha*. 2022.
7. Nagano Y, Prins M (eds.): The rGyalrongic Languages Database. 2013. Reference Source
8. STEDT database Reference Source

**Is the rationale for creating the dataset(s) clearly described?**
Yes

**Are the protocols appropriate and is the work technically sound?**
Yes

**Are sufficient details of methods and materials provided to allow replication by others?**
Yes

**Are the datasets clearly presented in a useable and accessible format?**
Yes

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** My specialties are Tibeto-Burman Linguistics, especially Lolo-Burmese languages. I have also interests in documenting Tai-Kadai languages in Mainland Southeast Asia.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 18 Oct 2023

**Yunfan Lai**

- While it has its merits, it also has its drawbacks. To begin with, the attempt is still in its infancy in terms of its overall goal, so one can only hope for further development. The data for rGyalrongic languages is still very limitedly employed. Of course, I comprehend that the principal objective of this project is to meticulously curate and present data in collaboration with experts. However, it is still important to increase the amount of data regarding the future. The author himself mentions that a database of the rGyalrongic language, supervised by Nagano & Prins (2013), is available on the website, where the basic vocabulary and examples of sentences of 81 rGyalrongic languages are uploaded with their sound files. Indeed, similar to the STEDT database, it is essentially a collection of data lacking morphological analysis.

However, in the future, with the cooperation of the supervisors, it would be possible to use the data from Nagano & Prins (2013) to conduct a more precise analysis of historical change.

- We are gradually expanding the data size with recent fieldwork under Yunfan Lai's project. Yunfan's project is now including two new varieties, Yaoji and Bragsteng into the database. They will be added as soon as data curation is done. Yunfan is also supervising an undergraduate student who works on a new West Gyalrongic variety. At least three new varieties with first hand data will be added to the database shortly. It is indeed a good idea, as the reviewer suggests, to explore more the Nagano and Prins database to see how we can collaborate.

- It should be noted that five of the languages in the dataset (Daofu, Zhaba, MaerkangrGyalrong, MuyaKangding, and QueyuXinlong) are taken from Huang (1992). Huang (1992) was a groundbreaking work at the time, a comparative lexicon of primary sources of Tibeto-Burman languages in China, which is very useful for historical and comparative studies. Unfortunately, Huang (1992) contains many printing errors, and the researcher, Prof. Huang Bufan, is deceased, thus we cannot ask her about her research. There are researchers, such as Satoko Shirai and Huang Yang on Zhaba (Gong 2007 is also a reference grammar of Zhaba), and Takumi Ikeda on Muya. For Situ rGyalrong, Yasuhiko Nagano also recently published a grammar in English (Nagano 2022). With the cooperation of these researchers who are still active in the field, there should be room for additional data and comparative studies in the future.
  - We are well aware that Huang's data are not free from errors. We are going to replace a part of her data with new, recent data. Yes, we will ask Shirai Satoko and Huang Yang for Zhaba data. We will include Cogtse data investigated by Lin You-Jing and Bhola data by Nagano.

- The paper presents the partial cognate annotation by exemplifying a word 'yesterday' in Figure 3, which this paper considers is a compound of 'day' and 'past'. This figure leads us to think that /ɣ/ in the Wobzi word /snəɣ/ corresponds to /x/ in the Siyuewu word /xsnə/ and both /ɣ/ and /x/ denote the concept 'past'. This is really inspiring, and if we search the same word in Nagano & Prins 2013 database, there are many cases similar to {pə/ ma/ wa}+{snə}, like /pəʃe'ʃni/ (Jinchuan Jimu Zhouchan), /ma sn̡i lo/ (Maerkang Soman), /mbasni'lo/ (Hongyuan Shuajinsi Selong), /wə3ʂo3ʃnə5/ (Jinchuan Manai Genza), etc. The initial syllable of these words might indicate 'past,' which should be noted in the Rangtang Puxi Siyaowucun word /ɣsnə/, having the same meaning as 'yesterday'. Additionally, Nagano (2022) also provides a few intriguing expressions for 'yesterday', warranting a comparative analysis. Collaborating with experts and integrating data from other databases could significantly enhance the development of this project.
  - The reviewer's comments are really interesting. In our recent fieldwork, we have cognates (such as pəʃesni) mentioned by the reviewer, so our new first hand data will certainly increase the diversity of the entire work. One thing to point out: Rangtang Siyaowucun (in Nagano and Prin's database) is exactly Siyuewu in our database, and we are confident that our notation is much more

accurate, because it is from Yunfan Lai's fieldwork from 2014 onwards.

- Regarding the dataset of this project, while Bantawa and Old Burmese, languages that are quite distant historically from rGyalrongic languages, are included, Written Tibetan and Amdo Tibetan are not included in this dataset. Of course, this may be because there are many cases of Tibetan languages providing loanwords to rGyalrongic, but this situation may require some explanation. As for Old Burmese, it certainly has a phonetic writing system and the date of use is clear, so it is possible that the historical relationship between the two languages can be put on an absolute time scale. However, if the Bantawa data are to be combined, it seems to me that more Qiangic languages should be introduced, including Pumi and various Qiang dialects.
    - For the purpose of phylogenetic inference, we need to select one or two obvious outgroups, which are known to be remotely related, or totally unrelated, to the branch under analysis. Many studies select only one, but here we have included two, Bantawa and Old Burmese. We did not select them randomly, but on purpose. Bantawa is remotely related to Rgyalrongic, which help us to identify the Rgyalrongic branch from a Sino-Tibetan/Trans-Himalayan perspective. Old Burmese are more closely related to Rgyalrongic, but not Rgyalrongic per se, therefore, it can let us know whether our Rgyalrongic is accurately computed without merging with Old Burmese.

- At any rate, as mentioned earlier, this database will continue to develop, and it is expected that the data will be updated in the future. At the moment, this database still needs more user-friendly interfaces which are already presented as in the STEDT and Nagano & Prince's databases. However, at the same time, if researchers of Tibeto-Burman languages apply the methodology of this project and develop datasets for each subgroup, it will be possible to discover new problems of historical change and to solve existing problems. At the very least, if the Lolo-Burmese dataset, which is the reviewer's specialty, is developed in a similar manner, it will allow for a more scientific analysis of the relationships between languages within Lolo-Burmese (which may modify the framework of Matisoff 1972, 2003, and Bradley 1979). It will also lead to a more detailed elucidation of the relationship between rGyalrongic and Lolo-Burmese, which was discussed here, than has been the case so far. We are convinced that this project deserves further attention.
    - As in the response to Reviewer 1, we are going to provide the Edictor access to readers and users, so that everybody can review our cognate annotation and identification. Future work will focus on making something like the database of Nagana and Prins (2011).

***Competing Interests:*** No competing interests were disclosed.

Reviewer Report 27 July 2023

https://doi.org/10.21956/openreseurope.17295.r33643

**© 2023 Alves M.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

**Mark Alves**
[1] Montgomery College, Rockville, Maryland, USA
[2] Montgomery College, Rockville, Maryland, USA

SUMMARY: This data report presents the status of ongoing historical linguistic research on the Rgyalrongic language group within Sino-Tibetan. It has involved assembling lexical data, identifying cognates, and preparing the data to use in phylogenetic study. The data includes thousands of words from 20 languages (mostly modern language, but the long-dead Tangut language) from a variety of previous studies. The goal is to make this data available for others with similar goals to use and, potentially, re-evaluate/re-test.

METHODOLOGY: The approach in assembling the data is proper: selection and organization of lexical data of as many doculects of Rgyalrongic as possible, identification of cognates / partial cognates, semantic and phonetic standardization (to a reasonable degree), and preparing in files useable with software that can do phylogenetic work.

There is one methodological matter I think worth giving attention to. While the assessment of cognates certainly must be done by researchers with relevant training and experience in working with the languages, such that they have trustworthy intuition about cognate status (your repeated use of the word "expert" in the article, which I think should be reduced), it would also be useful to acknowledge that phonological patterns were part of your mental criteria. You cannot and should not cover Rgyalrongic historical phonology in this short report, but a sentence or two mentioning that your intuition as researchers familiar with the languages involved noting recurring patterns (e.g., instances of consistency, but patterns of differences in consonants or vowels, etc.), along with a reference of a couple key references addressing some of these patterns, would help clarify the cognate identification process beyond "trustworthy intuition".

DATA AND SOFTWARE: While all the data is readily available, some guidance / tips on using the data would be helpful to some potential users.

SUGGESTIONS

Below is a list of suggestions for editing, wording, and clarity of details. Some are optional, some are necessary, and some are strongly recommended.

- Regarding "The present lexical data sets," consider "The lexical data sets the authors have assembled."

- "are collected...from 2011 to 2016" > "were collected...from 2011 to 2016"

- In "expertise in Rgyalrong historical linguistics through the neogrammarian comparative method", the phrase "through the neogrammarian comparative method" is unclear in relation to the preceding main clause. I suggest either "through application of the neogrammarian comparative method" or perhaps "expertise in research on Rgyalrongic

historical linguistics through the neogrammarian comparative method" (I don't think "neogrammarian" is needed).

- segmantation > segmentation
- "albeit including" doesn't connect to the previous sentence smoothly. Consider something like "though Tangut, an extinct medieval language, was spoken in Ningxia Province."
- "For most non-native speakers, they are the most difficult branch of languages to learn..." is odd. Non-native speakers of all Sino-Tibetan languages? Just remove "for most non-native speakers".
- "...they exhibit a large array of word formation strategies such as inflection and derivation..." is not logical since for "a large array", you can only list inflectional and derivational morphology (but perhaps reduplication). I think you mean that the inflectional and derivational morphology has a complex range of morphophonological patterns, combined with complex morphosyntax.
- Instead of "old age," perhaps "historical depth in the language group".
- Instead of "finding out the origin of Sino-Tibetan", perhaps "exploring Sino-Tibetan language history."
- Regarding "This database aims...", since this is an article, not a database, you need to write something like "The database considered in this article" or something like that.
- Instead of "Rgyalrongic languages are a Sino-Tibetan branch, mainly spoken...", perhaps "Rgyalrongic languages form a Sino-Tibetan branch and are mainly spoken..."
- "They are closely related to Lolo-Burmese languages" is minimal information, especially for non-specialists. Consider "They belong to the Qiangic sub-branch of Burmo-Qiangic and are thus closely related to Lolo-Burmese languages."
- "...a medieval language, Tangut, is recently recognised to belong to said branch" > "...the extinct Tangut language has been recently recognized as a Rgyalrongic language" or "has been recently determined to belong to this sub-branch".
- "...languages, Situ, ..." > "...languages: Situ, ..."
- fuß > Fuß
- The sentence "Thus, it is very likely that language varieties closely linked to Zhaba, such as Queyu and Minyag (aka. Muya, Menya) and Zlarong spoken in the Tibetan **Autonmous (> Autonomous)** Region, are also Rgyalrongic languages" is confusing. Since you include these languages in the study, "it is very likely" sounds too tentative, so consider "we have considered....to be Rgyalrongic languages in our study".
- In the phrase "These “newcomers”...", "newcomers" is somewhat ambiguous. Perhaps "new

additions to the Rgyalrongic group"?

- The statement "certainly the closest to Proto-Sino-Tibetan compared to all other varieties in China" is problematic. First, while there are many reconstructions of aspects of Sino-Tibetan morphology, these are not (to my knowledge) agreed upon to a major degree, which makes your claim premature. I agree that Rgyalrongic's complex morphological system suggest historical depth, and undoubtedly sheds light on Proto-Sino-Tibetan. But statements such as "certainly the closest...compared to all..." is a statement that would need substantial support to make, which is not the goal of this report. A statement such as "and it has a complex and, in our view, highly conservative morphological system that may give hints as to ancient features in the Sino-Tibetan language family", or instead of "may give", "has given" if you have a publication to cite with such support.

- "for the study on" > "for the study of"

- Instead of "The phylogeny of Rgyalrongic with dating information" ("with dating information" is awkward), consider "Phlogenetic research on Rgyalrongic to provide chronological information."

- For "has been proven to show accurate results3,5", consider "has been proven to show accurate results in both Sino-Tibetan and other language families3,5" since that is what the two references show.

- "high quality curation is inevitable" > "high quality curation is essential"

- "the phylogenetic analyses of Rgyalrongic languages" > "the phylogenetic analysis of Rgyalrongic languages"

- Instead of "It includes", consider "It contains lexical data from".

- Instead of "The entire workflow", consider "The workflow".

- Instead of "a specifically designed and curated," consider "a designed and curated".

- "has already been collected before 2017" < "was collected before 2017"

- "Dictionaries and word lists" > "These dictionaries and word lists"

- "Some word lists" > "Some of the word lists"

- Reorganize the following for standard presentation of items: "Reliability is assessed through two aspects: i) internal phonological consistency of the source data. The authors check if phonemes are correctly identified and that allophones conditioned by phonological environments and morphological alternations are adequately represented in the original sources. ii) external regularity of sound correspondences with the comparative method." > "Reliability is assessed through two aspects: i) internal phonological consistency of the source data and ii) external regularity of sound correspondences with the comparative method. The authors check if phonemes are correctly identified and that allophones

conditioned by phonological environments and morphological alternations are adequately represented in the original sources. In addition,..."

- "The authors check if..." > "The authors checked if..." (Parallel to the past tense in the subsequent sentence)

- In "...minor differences. Queyuxinlong and QueyuPubarong", consider "minor differences. In contrast, Queyuxinlong and QueyuPubarong" for evident contrast emphasis.

- Be careful: "These two languages are remotely related to Rgyalrongic" of which one is Old Burmese in Lolo-Burmese appears to contrast with your statement that Lolo-Burmese is closely related. You might need to replace "closely" in the previous instance since a time depth of 4,300 years suggests more linguistic distance, albeit within a common branch.

- "...sufficiently remote to..." > "...sufficiently remote from..."

- Instead of "largely facilitates", consider "facilitates".

- Suggestion: ", however," > ". However," and ", therefore" > ". Therefore,"

- Instead of "which are however not widely distributed from the perspectives of the entire Sino-Tibetan family", consider "which are not widely distributed in other branches of Sino-Tibetan."

- "included ‘girl’, as" > included ‘girl’ as"

- "languages, we therefore" > "languages. We therefore"

- "consider them as worthwhile to be included in the dataset" > "consider them worth including in the dataset"

- "from each other, some may" > "from each other, and some may"

- "can be deducted" > "can be deduced"

- The sentence "For instance, Mandarin Chinese yuè-liang ‘moon’ and Taiwanese Southern Min guēh-niûn ‘moon’ only share the first part yuè and guēh, respectively, as cognates" is not ideal as Sinitic is typologically divergent, and this example of bisyllabic compounds doesn't--in my view--sufficiently highlight the challenges in cognacy identification in polysyllabic languages in other branches of Sino-Tibetan. It would be best to show an example of this complex situation from Sino-Tibetan languages in other branches, and that have affixation of different sources, but cognate roots, not compounding. However, if you do keep this example, it would help to clearly show that they share 月, but that liang is for 亮 while Taiwanese niu is for 娘 (with the kind of certainty than cannot be offered for languages outside of Sinitic), and are thus distinct etyma.

- The sentence "Partial cognates enable us to segment word forms into morphemes, which improves the accuracy of the computation of language subgrouping" is not quite right.

First, the cognates don't allow segmentation: identification of partial cognates does this. Next, the cause of improved accuracy of computational subgrouping of the languages is the improved accuracy of cognate identification by marking both full and partial cognates. Consider rewording this.

- Question regarding "...annotate partial cognates, rather than full cognates...", does this mean there are no full cognates in the data? That may be possible if all words are morphologically marked with affixes. If so, make that clear. If not, "rather than" needs to be fixed.

- Perhaps there are different styles, but I generally use commas with numbers such as "6335" > "6,335" except when they are years. If this is standard for this publication, that's fine. Otherwise, please check throughout the paper.

- Regarding "...22 distinct language varieties", your abstract mentions 20. Please clarify.

- The phrase "413 distinct sounds" is ambiguous. I assume it mean "speech sounds" (not phonemes since that would require phonemic assessment). You might add "(i.e., consonants, vowels, and tones)" (if tones are among the sounds).

- "average inventory size of 72 different sounds" > Consider "average inventory of 72 distinct sounds". sounds"

- Regarding "The current dataset contains a total of 6335 word forms for 22 distinct language varieties. Word forms correspond to 305 different concepts," does this mean that 6,335 word forms fit into just 305 meanings? It's possible, but I would expect a larger number unless the wordlists are all short, as fieldwork lists sometimes are. Or do you mean that you focused on identification of words in 305 concepts? Please make this clear.

- Regarding "These morphemes have been assigned to 3109 cognate sets. Of these cognate sets, 1665 are unique, which means that we could not identify related words in other languages," this seems to say that 1,665 cognates sets do not have cognates in other languages, which is contradictory. Please rewrite to clarify.

- Regarding "We carefully verified the data to ensure the accuracy and correctness," the last part is general. First, "accuracy" and "correctness" are synonyms, so just use "accuracy", and then add briefly accuracy of what specifically about the data? That is was entered correctly from the sources, and/or other aspects? Clarifying this can increase confidence.

- I don't recommend overusing "expert, so for "We use our expert knowledge to review every lexical entry in the database" > "We reviewed every lexical entry in the database." However, reviewed them for what specifically? What kinds of problems were or aspects were you concerned with? You can just say "We checked the data," but that's uninformative.

- "potential reuses" > "potential reuse" (noncount noun)

- "judg**e**ments are" > "judgments, are"

- Again "...are carefully processed with expert knowledge", perhaps just "...were processed based on our experience in working with languages in the region" or something like that.
- Regarding "Languages differ in the choices of the two parts.", please clarify. Perhaps "The word forms differ in terms of cognacy of different morphemes"???
- Regarding "...for more accurate inference of phylogeny," this is data that is computationally analyzed, so not inferenced, right? Perhaps "...for more accurate data used for computational phylogenetic analysis."
- Regarding "Rgyalrongic languages are one of the most essential keys to the reconstruction of Proto-Sino-Tibetan as well as to the subgrouping of this language family," you must have at least a couple of references for this claim. If so, add them. If not, you need to hedge this statement, such as "We believe the Rgyalrongic languages have great potential value in reconstruction...".
- Again, "involves expert data curation and cognate annotation", do you need the word "expert"? One use in the article is sufficient, and this sentence is already clear without it.
- "phylogenetic analyses" > "phylogenetic analyses" (The noncount meaning is best here)
- Regarding "For now, only those morphemes are assigned to the same cognate set which occur in words sharing the same meaning.", this is confusing. Perhaps "For now, only those morphemes assigned to the same cognate set occur in words sharing the same meaning." Is that what you mean?
- Regarding "...to account for cognates across meaning slots", by "across meaning slots," do you mean near synonyms and/or instances of semantic extension, or something else? Please clarify.
- "concentrating also on" > "concentrating on" (Is "also" needed"?)
- "on the word level" > "at the word level"

MY RESEARCH BACKGROUND

My research focus is on Austroasiatic, with Vietnamese as the center of this, but study of Vietnamese language history in particular requires much information about Chinese and Kra-Dai language history, and Southeast Asian language history broadly, all part of my broader research agenda. This involves sifting lexical data in digital databases for etymological study, naturally including cognate identification and interdisciplinary evidence of language contact. Also in Austroasiatic historical linguistics, I must deal with neighboring Tibeto-Burman languages and frequently consult STEDT (the Sino-Tibetan Etymological Dictionary)). However, while I have seen studies of Rgyalrongic languages, I have little specific knowledge of of them, though as I am interested in the ongoing research into Sino-Tibetan (aka Trans-Himalayan) language history, this brief report is of interest, both in terms of the potential use in Sino-Tibetan historical linguistics and in the methodology the authors used and challenges they face in identifying cognates.

**Is the rationale for creating the dataset(s) clearly described?**

Yes

**Are the protocols appropriate and is the work technically sound?**
Yes

**Are sufficient details of methods and materials provided to allow replication by others?**
Partly

**Are the datasets clearly presented in a useable and accessible format?**
Yes

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Historical linguistics (phonology, syntax, etymology, dialectology, language contact, inter-disciplinary approaches) of Vietnamese, Austroasiatic, and Southeast Asia broadly

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 27 Jul 2023

**Yunfan Lai**

Dear Mark, thank you very much for your comments. They are very helpful indeed. We are still waiting for the third and probably final review, then we will do our best to revise the paper. Best, Yunfan

***Competing Interests:*** No competing interests were disclosed.

Author Response 18 Oct 2023

**Yunfan Lai**

- There is one methodological matter I think worth giving attention to. While the assessment of cognates certainly must be done by researchers with relevant training and experience in working with the languages, such that they have trustworthy intuition about cognate status (your repeated use of the word "expert" in the article, which I think should be reduced), it would also be useful to acknowledge that phonological patterns were part of your mental criteria. You cannot and should not cover Rgyalrongic historical phonology in this short report, but a sentence or two mentioning that your intuition as researchers familiar with the languages involved noting recurring patterns (e.g., instances of consistency, but patterns of differences in consonants or vowels, etc.), along with a reference of a couple key references addressing some of these patterns, would help clarify the cognate identification process beyond "trustworthy intuition".
    - This is a very good suggestion. In the first place, we have reduced the word “expert”. In the second place, we have mentioned some criteria, including

"trustworthy intuition" or "educated guess" in the revised version. Recently, I have several new publications discussing sound correspondences among Rgyalrongic languages, I have also mentioned them when revising the paper.

- Regarding "The present lexical data sets," consider "The lexical data sets the authors have assembled."
  - I have changed the wording accordingly.

- "are collected...from 2011 to 2016" > "were collected...from 2011 to 2016"
  - I have changed the wording accordingly.

- In "expertise in Rgyalrong historical linguistics through the neogrammarian comparative method", the phrase "through the neogrammarian comparative method" is unclear in relation to the preceding main clause. I suggest either "through application of the neogrammarian comparative method" or perhaps "expertise in research on Rgyalrongic historical linguistics through the neogrammarian comparative method" (I don't think "neogrammarian" is needed).
  - I have changed the wording accordingly.

- segmantation > segmentation
  - I have changed the wording accordingly.

- "albeit including" doesn't connect to the previous sentence smoothly. Consider something like "though Tangut, an extinct medieval language, was spoken in Ningxia Province."
  - I have changed the wording accordingly.

- "For most non-native speakers, they are the most difficult branch of languages to learn..." is odd. Non-native speakers of all Sino-Tibetan languages? Just remove "for most non-native speakers".
  - I have changed the wording accordingly.

- "...they exhibit a large array of word formation strategies such as inflection and derivation..." is not logical since for "a large array", you can only list inflectional and derivational morphology (but perhaps reduplication). I think you mean that the inflectional and derivational morphology has a complex range of morphophonological patterns, combined with complex morphosyntax.
  - The reviewer is right about this point. I have selected a better wording.

- Instead of "old age," perhaps "historical depth in the language group".
  - I have changed the wording accordingly.

- Instead of "finding out the origin of Sino-Tibetan", perhaps "exploring Sino-Tibetan language history."
  - I have changed the wording accordingly.

- Regarding "This database aims...", since this is an article, not a database, you need to write something like "The database considered in this article" or something like that.
  - I have changed the wording accordingly.
- Instead of "Rgyalrongic languages are a Sino-Tibetan branch, mainly spoken...", perhaps "Rgyalrongic languages form a Sino-Tibetan branch and are mainly spoken..."
  - I have changed the wording accordingly.
- "They are closely related to Lolo-Burmese languages" is minimal information, especially for non-specialists. Consider "They belong to the Qiangic sub-branch of Burmo-Qiangic and are thus closely related to Lolo-Burmese languages."
  - I have changed the wording accordingly.
- "...a medieval language, Tangut, is recently recognised to belong to said branch" > "...the extinct Tangut language has been recently recognized as a Rgyalrongic language" or "has been recently determined to belong to this sub-branch".
  - I have changed the wording accordingly.
- "...languages, Situ, ..." > "...languages: Situ, ..."
  - I have changed the wording accordingly.
- fuß > Fuß
  - I have also noticed this error. I have corrected it.
- The sentence "Thus, it is very likely that language varieties closely linked to Zhaba, such as Queyu and Minyag (aka. Muya, Menya) and Zlarong spoken in the Tibetan Autonmous (> Autonomous) Region, are also Rgyalrongic languages" is confusing. Since you include these languages in the study, "it is very likely" sounds too tentative, so consider "we have considered....to be Rgyalrongic languages in our study".
  - We fully agree with the reviewer.
- In the phrase "These “newcomers”...", "newcomers" is somewhat ambiguous. Perhaps "new additions to the Rgyalrongic group"?
  - We also prefer the suggestion of the reviewer.
- The statement "certainly the closest to Proto-Sino-Tibetan compared to all other varieties in China" is problematic. First, while there are many reconstructions of aspects of Sino-Tibetan morphology, these are not (to my knowledge) agreed upon to a major degree, which makes your claim premature. I agree that Rgyalrongic's complex morphological system suggest historical depth, and undoubtedly sheds light on Proto-Sino-Tibetan. But statements such as "certainly the closest...compared to all..." is a statement that would need substantial support to make, which is not the

goal of this report. A statement such as "and it has a complex and, in our view, highly conservative morphological system that may give hints as to ancient features in the Sino-Tibetan language family", or instead of "may give", "has given" if you have a publication to cite with such support.
    - The reviewer's doubt is reasonable. Our claim is indeed premature. We have revised this passage in order to make it more scientifically viable.

- "for the study on" > "for the study of"
    - I have changed the wording accordingly.

- Instead of "The phylogeny of Rgyalrongic with dating information" ("with dating information" is awkward), consider "Phlogenetic research on Rgyalrongic to provide chronological information."
    - I doubt that "chronological" is the appropriate term. Bayesian phylogeny gives exact estimations of dates, instead of chronology, of sub-branches. I have made this more comprehensible.

- For "has been proven to show accurate results3,5", consider "has been proven to show accurate results in both Sino-Tibetan and other language families3,5" since that is what the two references show.
    - I have changed the wording accordingly.

- "high quality curation is inevitable" > "high quality curation is essential"
    - I have changed the wording accordingly.

- "the phylogenetic analyses of Rgyalrongic languages" > "the phylogenetic analysis of Rgyalrongic languages"
    - I have changed the wording accordingly.

- Instead of "It includes", consider "It contains lexical data from".
    - I have changed the wording accordingly.

- Instead of "The entire workflow", consider "The workflow".
    - I have changed the wording accordingly.

- Instead of "a specifically designed and curated," consider "a designed and curated".
    - I have changed the wording accordingly.

- "has already been collected before 2017" < "was collected before 2017"
    - I have changed the wording accordingly.

- "Dictionaries and word lists" > "These dictionaries and word lists"
    - I have changed the wording accordingly.

- "Some word lists" > "Some of the word lists"
  - I have changed the wording accordingly.

- Reorganize the following for standard presentation of items: "Reliability is assessed through two aspects: i) internal phonological consistency of the source data. The authors check if phonemes are correctly identified and that allophones conditioned by phonological environments and morphological alternations are adequately represented in the original sources. ii) external regularity of sound correspondences with the comparative method." > "Reliability is assessed through two aspects: i) internal phonological consistency of the source data and ii) external regularity of sound correspondences with the comparative method. The authors check if phonemes are correctly identified and that allophones conditioned by phonological environments and morphological alternations are adequately represented in the original sources. In addition,..."
  - I have changed the wording accordingly.

- "The authors check if..." > "The authors checked if..." (Parallel to the past tense in the subsequent sentence)
  - I have changed the wording accordingly.

- In "...minor differences. Queyuxinlong and QueyuPubarong", consider "minor differences. In contrast, Queyuxinlong and QueyuPubarong" for evident contrast emphasis.
  - I have changed the wording accordingly.

- Be careful: "These two languages are remotely related to Rgyalrongic" of which one is Old  Burmese in Lolo-Burmese appears to contrast with your statement that Lolo-Burmese is closely related. You might need to replace "closely" in the previous instance since a time depth of 4,300 years suggests more linguistic distance, albeit within a common branch.
  - Yes, this is a problem. I have fixed this.

- "...sufficiently remote to..." > "...sufficiently remote from..."
  - I have changed the wording accordingly.

- Instead of "largely facilitates", consider "facilitates".
  - I have changed the wording accordingly.

- Suggestion: ", however," > ". However," and ", therefore" > ". Therefore,"
  - I have changed the wording accordingly.

- Instead of "which are however not widely distributed from the perspectives of the entire Sino-Tibetan family", consider "which are not widely distributed in other

branches of Sino-Tibetan."
    - I have changed the wording accordingly.

- "included ‘girl’, as" > included ‘girl’ as"
    - I have changed the wording accordingly.

- "languages, we therefore" > "languages. We therefore"
    - I have changed the wording accordingly.

- "consider them as worthwhile to be included in the dataset" > "consider them worth including in the dataset"
    - I have changed the wording accordingly.

- "from each other, some may" > "from each other, and some may"
    - I have changed the wording accordingly.

- "can be deducted" > "can be deduced"
    - I have changed the wording accordingly.

- The sentence "For instance, Mandarin Chinese yuè-liang ‘moon’ and Taiwanese Southern Min guēh-niûn ‘moon’ only share the first part yuè and guēh, respectively, as cognates" is not ideal as Sinitic is typologically divergent, and this example of bisyllabic compounds doesn't--in my view--sufficiently highlight the challenges in cognacy identification in polysyllabic languages in other branches of Sino-Tibetan. It would be best to show an example of this complex situation from Sino-Tibetan languages in other branches, and that have affixation of different sources, but cognate roots, not compounding. However, if you do keep this example, it would help to clearly show that they share 月, but that liang is for 亮 while Taiwanese niu is for 娘 (with the kind of certainty than cannot be offered for languages outside of Sinitic), and are thus distinct etyma.
    - I have removed this example and replaced it with the example for yesterday in Figure 3, which is also in our database.

- The sentence "Partial cognates enable us to segment word forms into morphemes, which improves the accuracy of the computation of language subgrouping" is not quite right. First, the cognates don't allow segmentation: identification of partial cognates does this. Next, the cause of improved accuracy of computational subgrouping of the languages is the improved accuracy of cognate identification by marking both full and partial cognates. Consider rewording this.
    - Yes. The emphasis on partial cognates is that it is the recent innovative technique developped by our team. I have reworded this to make it more accurate.

- Question regarding "...annotate partial cognates, rather than full cognates...", does this mean there are no full cognates in the data? That may be possible if all words are

morphologically marked with affixes. If so, make that clear. If not, "rather than" needs to be fixed.
    - There are indeed full cognates in this data. I have fixed this point.

- Perhaps there are different styles, but I generally use commas with numbers such as "6335" > "6,335" except when they are years. If this is standard for this publication, that's fine. Otherwise, please check throughout the paper.
    - I have corrected it.

- Regarding "...22 distinct language varieties", your abstract mentions 20. Please clarify.
    - There are two out-group languages, Bantawa and Old Burmese, apart from the 20 Rgyalrongic languages. I have clarified this point.

- The phrase "413 distinct sounds" is ambiguous. I assume it mean "speech sounds" (not phonemes since that would require phonemic assessment). You might add "(i.e., consonants, vowels, and tones)" (if tones are among the sounds).
    - Yes. I have added them in the revised version.
- "average inventory size of 72 different sounds" > Consider "average inventory of 72 distinct sounds". sounds"
    - I have corrected this.

- Regarding "The current dataset contains a total of 6335 word forms for 22 distinct language varieties. Word forms correspond to 305 different concepts," does this mean that 6,335 word forms fit into just 305 meanings? It's possible, but I would expect a larger number unless the wordlists are all short, as fieldwork lists sometimes are. Or do you mean that you focused on identification of words in 305 concepts? Please make this clear.
    - The 6,335 word forms fit into the 305 concepts (meanings) in 22 languages, which I think is rather a reasonable number and distribution, since we cannot guarantee that every language has a word form for each of the 305 concepts. I have made this point more explicit in the revised version.

- Regarding "These morphemes have been assigned to 3109 cognate sets. Of these cognate sets, 1665 are unique, which means that we could not identify related words in other languages," this seems to say that 1,665 cognates sets do not have cognates in other languages, which is contradictory. Please rewrite to clarify.
    - This is related to our definition of cognate sets. It is possible that a cognate set contains only one form from one language. In this case, this cognate set is unique. Unique cognate sets can be due to innovations of individual languages, or rare retentions from the proto-language.
- Regarding "We carefully verified the data to ensure the accuracy and correctness," the last part is general. First, "accuracy" and "correctness" are synonyms, so just use "accuracy", and then add briefly accuracy of what specifically about the data? That is was entered correctly from the sources, and/or other aspects? Clarifying this can increase confidence.
    - I agree that this point needs more clarifications. I have added a passage on this

issue.

- I don't recommend overusing "expert, so for "We use our expert knowledge to review every lexical entry in the database" > "We reviewed every lexical entry in the database." However, reviewed them for what specifically? What kinds of problems were or aspects were you concerned with? You can just say "We checked the data," but that's uninformative.
  - I have changed the wording accordingly.

- "potential reuses" > "potential reuse" (noncount noun)
  - I have corrected this.

- "judgements are" > "judgments, are"
  - I have corrected this.

- Again "...are carefully processed with expert knowledge", perhaps just "...were processed based on our experience in working with languages in the region" or something like that.
  - I have changed the wording with more information.

- Regarding "Languages differ in the choices of the two parts.", please clarify. Perhaps "The word forms differ in terms of cognacy of different morphemes"???
  - Yes. This is not clear enough. I have changed the wording.

- Regarding "...for more accurate inference of phylogeny," this is data that is computationally analyzed, so not inferenced, right? Perhaps "...for more accurate data used for computational phylogenetic analysis."
  - The reviewer is right about this point. I have revised the paper accordingly.

- Regarding "Rgyalrongic languages are one of the most essential keys to the reconstruction of Proto-Sino-Tibetan as well as to the subgrouping of this language family," you must have at least a couple of references for this claim. If so, add them. If not, you need to hedge this statement, such as "We believe the Rgyalrongic languages have great potential value in reconstruction...".
  - I have added a couple of references.

- Again, "involves expert data curation and cognate annotation", do you need the word "expert"? One use in the article is sufficient, and this sentence is already clear without it.
  - The reviewer is indeed right to point out my overuse of “expert”.

- "phylogenetic analyses" > "phylogenetic analyses" (The noncount meaning is best here)
  - I have corrected the word.

- Regarding "For now, only those morphemes are assigned to the same cognate set which occur in words sharing the same meaning.", this is confusing. Perhaps "For now, only those morphemes assigned to the same cognate set occur in words sharing the same meaning." Is that what you mean?
  - Yes, this is exactly what I mean.

- Regarding "...to account for cognates across meaning slots", by "across meaning slots," do you mean near synonyms and/or instances of semantic extension, or something else? Please clarify.
  - I mean cognates with different meanings: for instance, sqî in Khroskyabs means ‘female siblings’, its cognate in Bragbar, ɕkʰéj, means ‘female friend’. We will develop a workflow to deal with this kind of cross-semantic cognates.

- "concentrating also on" > "concentrating on" (Is "also" needed"?)
  - I have corrected this.
- "on the word level" > "at the word level"
  - I have corrected this.

***Competing Interests:*** No competing interests were disclosed.

Reviewer Report 07 July 2023

https://doi.org/10.21956/openreseurope.17295.r32973

**© 2023 Michaud A.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

**Alexis Michaud**
[1] LACITO - Langues et civilisations à tradition orale, Paris, France
[2] LACITO - Langues et civilisations à tradition orale, Paris, France

**Summary of the submission:**
This database gathers lexical information of Rgyalrongic languages to facilitate future research on historical and evolutionary linguistics. It includes word lists from fieldwork and published sources, providing data for phylogenetic analysis of the Sino-Tibetan family.

An important disclaimer is that this reviewer has no first-hand knowledge of any Rgyalrongic language, and only contributes to historical-comparative work as part of a team that includes professional historical linguists. I guess that it is relevant to further clarify that my main research focus is on the Naish subgroup of Sino-Tibetan (Trans-Himalayan) – the Naxi, Na (Mosuo) and Laze languages –, and that the topic of Rgyalrongic reconstruction is of great interest to me, as

progress in Rgyalrongic historical linguistics provides a gradually improved basis for establishing cognate sets between Rgyalrongic and Naish.

The data note «Lexical data for the historical comparison of Rgyalrongic languages» introduces the current version of a database which, as can be looked up from the stats of the GitHub repository, has been set up in early 2020 and has undergone significant reworking in the Spring of 2022. Publication on the Open Research Europe platform will hopefully help clarify to a wide audience the usefulness of the database in its present state, as well as draw attention to a number of forward-looking features in its architecture. The database is a cutting-edge research tool for the comparative study of Rgyalrongic in its Sino-Tibetan context. The database does not aim to aggregate *lexical data* for historical-linguistic purposes, as done e.g. in the Sino-Tibetan Etymological Dictionary and Thesaurus (STEDT) project at the University of California at Berkeley. The database under review only contains 305 hand-picked cognate sets, currently amounting to a total of 6,335 word forms (for 22 distinct language varieties). By contrast, the STEDT lexical file is more than an order of magnitude larger, containing over 376,000 words in about 200 languages and dialects. Looking for the word 'nose' in STEDT (https://stedt.berkeley.edu/~stedt-cgi/rootcanal.pl/gnis?t=nose), one gets a set of no less than 1,241 records, with twelve separate etyma. The entries are arranged by language subgroups, without cognacy judgments. There are 129 entries for six Rgyalrongic languages. These languages are (adding, in brackets, the corresponding names in the work under review): Ganzi Danba Geshenzha (Geshizha), Daofu/Ergong (Daofu/Mazur), Lavrung (Khroskyabs), Minyag (Muya/Menya), and several dialects of Rgyalrong. The database under review is much smaller in terms of number of items (again, by nearly an order of magnitude), but has better language coverage (covering about twice as many languages) and, crucially, only returns one item for the intended word: the cognate decided upon manually by the expert in charge of the database (the corresponding author of the data note, LAI Yunfan). There are thus fundamental differences between the STEDT database and the database under review. The STEDT, complete with an online user interface, is a functional and versatile large-scale tool that offers ample opportunity for 'data crunching', including peeking at items that are semantically related ('to blow one's nose', 'nose flaps'...). By contrast, the database under review is a diamond point: it is small by design, intended as a building-block for state-of-the-art explorations in computer-assisted historical linguistics. Use of standardized formats allows it to be plugged into a range of computational pipelines, to be used in association with other datasets as required for a specific research purpose.

From the point of view of research methods, the database under review is up to the highest standards from several perspectives: that of Open Science (the data note itself is published under a permissive CC BY license), that of Sino-Tibetan historical linguistics (as the data curation process is exemplary), and that of interdisciplinary work associating linguistics and computer science. Such lovingly handcrafted, computation-friendly cognate sets allow for a seamless integration of the time-honoured methods of historical linguistics with computational approaches. It is to be hoped that other datasets that are key to progress in Sino-Tibetan historical linguistics, such as Old Chinese reconstructions in the Baxter-Sagart system (currently hosted at a custom website, and thus not looking very 'future-proof': https://ocbaxtersagart.lsait.lsa.umich.edu/), will find their way to (i) archives ensuring long-term conservation and (ii) hubs routinely used by computational linguists and computer scientists, such as GitHub, GitLab, Bitbucket and such – the landscape here changes rapidly, but the investment of publishing datasets on these platforms is well worth the effort, as it greatly facilitates interdisciplinary collaborations.

On a slightly critical note, the data note presenting the database does not appear to me to do full justice to the database's design and applications. In the «Plain language summary», the authors' description of the database as work «in preparation for future work on historical and evolutionary linguistics» strikes me as unnecessarily modest. Perspectives of uses of the database are pushed back into a somewhat vague future, as if the database were intended as food for thought for linguists living at some point along the (theoretical) line of digital eternity. True, posterity may be grateful and appreciative, but the same very general point could be made about any other data set that gets curated and archived. In real life, not that many of the datasets in digital archives such as Zenodo may eventually be used in future, and still fewer may prove useful in research. By contrast, the database of rGyalrongic languages is not only a functional tool for research: it could be argued to be, by now, a tool whose usefulness has been tried and tested. The tool was being used at the same time as it was being set up. In computational terms, one could say that public release of the database is a transition from active 'alpha-testing' to 'beta-testing' by an open-ended list of end-users. Moreover, new end-users could potentially also serve as contributors in the mid run, for other cognate sets that may include a different choice of languages and/or concepts. The authors of the database are part of the team of authors of an influential 2019 article arguing that «Dated language phylogenies shed light on the ancestry of Sino-Tibetan»; there are clear shared features between the cognate sets used in the 2019 article and in the database under review, such as use of the second author's Concepticon. Thus, the database under review is part of a growing body of data and tools that possesses demonstrated value to the field of Sino-Tibetan diachronic research. I would therefore suggest rephrasing the relevant passage of the «Plain language summary» from «in preparation for future work on historical and evolutionary linguistics» to «as a tool for research in the field of historical and evolutionary linguistics».

Concerning matters of form, the data note could do with an additional round of text editing to ensure best readability for authors outside the first circle of specialists of this area of Sino-Tibetan. There are 20 languages in Figure 1, and 22 in Table 1. It would be useful to provide clarifications in the caption to Figure 1. Quick fix: *Figure 1. Geographical distribution of the **Rgyalrongic** languages in the database*. Full explicitness would be a service to some readers: adding a clarification to the effect that '(Bantawa and Old Burmese, which are not Rgyalrongic languages, are not shown on this map.)' Some redundancy would help here. The profusion of proper names specific to the area, and the presence of concepts specific to historical linguistics, are perfectly natural in work of this nature, and are essentially inevitable. But their combination tends to put off members of an extended readership of linguists, when one's best efforts at imbibing the area-specific information in an article leave one baffled. So, efforts at clarity are most needed, to accompany readers accustomed to softer linguistic landscapes. (Decisions at the authors' choice.) The issue extends to language names and language identifiers. There is room for improvement here, and a small amount of further work in this space would entail large benefits for the legibility of the article – and arguably, for the usability of the database for a not-too-narrow public. Currently, language names do not have a unified syntax. Taking «Menya-Gao» as an example, where «Menya» is the language name, and «Gao» the family name of the author who collected the original data: «Menya-Gao» is OK as an identifier (for computational purposes), but not too great as a language name. The syntax is, to say the least, unusual. It could make sense if the authors wanted to lay emphasis on the specific 'doculects' used in their study, and to give credit to the authors who provided the indispensable basis for the database, and who bear responsibility for data quality. If so, the syntax would yield: Menya Gao, Muya Huang, Situ Zhang, Japhug Jacques, Stau Gates, Khroskyabs Lai, Rma Sims, Zbu Gong, and so on. If the authors wish to float this new practice, it should be applied consistently. But since information on the source of data is provided in the

table, it would make excellent sense to go by common practices in English-language publications, and use a syntax such as:

In detail, «Kangding Muya» would be somewhat under-specific, since «Menya-Gao» is spoken in Kangding, too, and the main text describes «Muya» and «Menya» as alternative names for Minyag: «Minyag (aka. Muya, Menya)», seemingly in the same way as «Sino-Tibetan» and «Trans-Himalayan» are different labels for the same language family, or, at the language level, «Pumi» and «Prinmi», «Na» and «Mosuo», «Standard Mandarin» and «Putonghua», «Shixing» and «Xumi», etc. But crucially, there *is* a dialect difference between the scopes of the two labels as used in the database: Figure 1 has Muya and Menya as distinct data points. Changes to the language names in Table 1 do not imply any modifications to the suite of computer files & scripts: only a change to the table and to the main text of the presentation article under review here.

The repository where the dataset is made available could do with quick additional UX improvements, too: improving the experience of less advanced users – basically, linguists with an awareness of the usefulness of computer scripts but without regular practice of scripting. To begin with, a path to the main reference files should be indicated clearly on the repo's landing page, and perhaps repeated in some of the subfolders, along the lines of 'This folder contains . In case you are looking for the database in CSV format, please refer to .' Thus, in the presentation text (the article under review) the language 'MazurStau' appears in Figure 1 (as 'Mazur') and in Table 1. The forms.csv file, which (at the time of this review) had been updated on April 26th, 2023, has the MazurStau forms. But this reviewer has not been able to locate any 'MazurStau' data in the list of cognates (wordlist-cognates.tsv) in the 'analysis' folder. For instance: for 'nose', the .tsv file has 21 entries; the Stau form /sni/ (Gates 2021:50) is conspicuously cognate, and would make for a complete cognate set of 22 varieties. Is the .tsv file, last updated on April 20th, 2020, an old file, not relevant to the database in the present state? If so, some cleanup of the GitHub repo or additional explanations would be a service to potential users, so they don't get led down a garden path when following the enticing lead of an 'analysis' folder name. Providing such folders with a short README file would constitute sufficient guidance, and would be well worth the effort – pending the release of future newfangled tools on the GitHub web interface that allow for easy navigation to key files: those that are updated most often, that are referenced at key points in scripts, and other such criteria.

Minor points:
Introduction: «In order to infer the phylogenetic subgrouping of Rgyalrongic, a lexical dataset with high quality curation is inevitable»: inevitable > indispensible

In the discussion of elicitation, the mention «without physical contact» is puzzling. What are the intended implications? Does it have to do with the differences in legal frameworks for experiments involving human subjects depending on the protocol, with/without physical contact? If the mention is intended as a legal disclaimer, the statement should be made in a relevant section or footnote, not in the main text, where it acts as a distractor. I would suggest making a reference to a publication that sets out the essentials of linguistic fieldwork, with any qualifications and additions the authors may wish to make. Thus, interested readers would get a useful pointer to a resource explaining about linguistic fieldwork, and broaching the all-important human topic of the investigator's relationship with consultants.

p. 5 German *fuß*: why not capitalize the F, thus: *Fuß*?

Acknowledgements: and detailed explanations to our questions > and detailed explanations in answer to our questions

References: the URL provided for
Gates JP: *Grammaire du stau de Mazur*. Paris: Écoles des Hautes Études en Sciences Sociales dissertation, 2021.
is unhelpful: it is a link to an online announcement of the PhD defense, not a link to the document itself. Since the dissertation is available online, the authors should cite an institutional repository: either the dissertation's entry on theses.fr (https://theses.fr/2021EHES0054), or a direct link to the PDF, though the latter option is likely to prove less 'time-proof' ( https://www.theses.fr/2021EHES0054/abes/Gates_Jesse_these_2020.pdf).

**Is the rationale for creating the dataset(s) clearly described?**
Yes

**Are the protocols appropriate and is the work technically sound?**
Yes

**Are sufficient details of methods and materials provided to allow replication by others?**
Partly

**Are the datasets clearly presented in a useable and accessible format?**
Yes

***Competing Interests:*** As stated within the main text of the review: (i) an important disclaimer is that I have no first-hand knowledge of any Rgyalrongic language, and only contribute to historical-comparative work as part of a team that includes professional historical linguists; (ii) my main research focus is on the Naish subgroup of Sino-Tibetan (Trans-Himalayan) – the Naxi, Na (Mosuo) and Laze languages –, and the topic of Rgyalrongic reconstruction is of great interest to me, as progress in Rgyalrongic historical linguistics provides a gradually improved basis for establishing cognate sets between Rgyalrongic and Naish. So I stand to benefit from progress of research on Rgyalrongic languages. I hope that this does not bias my assessment, but to what extent that constitutes a conflict of interest is for others to ascertain.

***Reviewer Expertise:*** Naish subgroup of Sino-Tibetan (Naxi, Na/Mosuo, Laze); phonetics/phonology; language documentation and conservation; Computational Language Documentation

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 08 Jul 2023
**Yunfan Lai**

Dear Alexis, Thank you very much for your detailed comments. We will improve the paper accordingly after we have received one more review.

***Competing Interests:*** No competing interests were disclosed.

Author Response 18 Oct 2023

**Yunfan Lai**

- Differences between our database and the STEDT
  - Another crucial difference between our database and the STEDT is that the first author, Lai Yunfan, is a specialist in Gyalrongic languages, and thus capable of detecting doubtful forms/mistakes in first and second hand sources. Once Lai Yunfan finds a potential error in the sources, he would immediately turn to the authors of the sources or native speakers to ask for verification. This is something the STEDT cannot do: they took nearly everything directly from the sources without significant consideration on their accuracy.

- On a slightly critical note, the data note presenting the database does not appear to me to do full justice to the database's design and applications. In the «Plain language summary», the authors' description of the database as work «in preparation for future work on historical and evolutionary linguistics» strikes me as unnecessarily modest [...] Thus, the database under review is part of a growing body of data and tools that possesses demonstrated value to the field of Sino-Tibetan diachronic research. I would therefore suggest rephrasing the relevant passage of the «Plain language summary» from «in preparation for future work on historical and evolutionary linguistics» to «as a tool for research in the field of historical and evolutionary linguistics».
  - We sincerely thank the reviewer's confidence in our database. We find that the reviewer's points are indeed useful and encouraging. We will change the wording in our plain language summary accordingly.

- Concerning matters of form, the data note could do with an additional round of text editing to ensure best readability for authors outside the first circle of specialists of this area of Sino-Tibetan. There are 20 languages in Figure 1, and 22 in Table 1. It would be useful to provide clarifications in the caption to Figure 1. Quick fix: Figure 1. Geographical distribution of the Rgyalrongic languages in the database. Full explicitness would be a service to some readers: adding a clarification to the effect that '(Bantawa and Old Burmese, which are not Rgyalrongic languages, are not shown on this map.)'
  - We have ameliorated the explicitness of the descriptions throughout the paper.

- The issue extends to language names and language identifiers. There is room for improvement here, and a small amount of further work in this space would entail large benefits for the legibility of the article – and arguably, for the usability of the database for a not-too-narrow public. Currently, language names do not have a

unified syntax. Taking «Menya-Gao» as an example, where «Menya» is the language name, and «Gao» the family name of the author who collected the original data: «Menya-Gao» is OK as an identifier (for computational purposes), but not too great as a language name [...]
    - The reviewer here suggests a crucial revision. We totally agree with it and made the syntax of language names consistent.

- The repository where the dataset is made available could do with quick additional UX improvements, too: improving the experience of less advanced users – basically, linguists with an awareness of the usefulness of computer scripts but without regular practice of scripting [...]
    - This is indeed a very good suggestion. We will surely try our best to improve the user-friendly-ness. For the time being, we will provide a stable Edictor link accessible for readers and users. Edictor is easy to use with a brief instruction, has a good UX design, and is constantly improving with additional features.

- The forms.csv file, which (at the time of this review) had been updated on April 26th, 2023, has the MazurStau forms. But this reviewer has not been able to locate any 'MazurStau' data in the list of cognates (wordlist-cognates.tsv) in the 'analysis' folder. For instance: for 'nose', the .tsv file has 21 entries; the Stau form /sni/ (Gates 2021:50) is conspicuously cognate, and would make for a complete cognate set of 22 varieties. Is the .tsv file, last updated on April 20th, 2020, an old file, not relevant to the database in the present state? If so, some cleanup of the GitHub repo or additional explanations would be a service to potential users, so they don't get led down a garden path when following the enticing lead of an 'analysis' folder name.
    - This is indeed an important point. The most recent .tsv files are actually in the folder "hacks", where we have a Python script add_stau.py to add data of Mazur Stau. The resulting .tsv file is rgyalrong-cogid.tsv which includes Mazur Stau. The file wordlist-cognates.tsv is not relevant anymore. We have done a clean-up in our repository.

- Introduction: «In order to infer the phylogenetic subgrouping of Rgyalrongic, a lexical dataset with high quality curation is inevitable»: inevitable > indispensible
    - We have corrected it.

- In the discussion of elicitation, the mention «without physical contact» is puzzling. What are the intended implications? Does it have to do with the differences in legal frameworks for experiments involving human subjects depending on the protocol, with/without physical contact? If the mention is intended as a legal disclaimer, the statement should be made in a relevant section or footnote, not in the main text, where it acts as a distractor. I would suggest making a reference to a publication that sets out the essentials of linguistic fieldwork, with any qualifications and additions the authors may wish to make. Thus, interested readers would get a useful pointer to a resource explaining about linguistic fieldwork, and broaching the all-important human topic of the investigator's relationship with consultants.
    - This is related to the ethical regulations of our grants. I have made this more

explicitly explained to readers. The disclaimers are now in Footnote 1 in the revised version.

- p. 5 German fuß: why not capitalize the F, thus: Fuß?
  - We have capitalized it.

- Acknowledgements: and detailed explanations to our questions > and detailed explanations in answer to our questions
  - We have corrected it.

***Competing Interests:*** No competing interests were disclosed.