DATA NOTE

Lexical data for the historical comparison of Rgyalrongic languages [version 1; peer review: 1 approved with reservations]

Yunfan Lai1,2, Johann-Mattis List2,3

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

V1 First published: 20 Jun 2023, 3:99
https://doi.org/10.12688/openreseurope.16017.1
Latest published: 20 Jun 2023, 3:99
https://doi.org/10.12688/openreseurope.16017.1

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 present lexical data sets 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 are collected by the first author from 2011 to 2016. Cognate identification is based on the authors' expertise in Rgyalrong historical linguistics through the neogrammarian 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.

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

This article is included in the Languages and Literature gateway.
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: 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: 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 1; peer review: 1 approved with reservations] Open Research Europe 2023, 3:99 https://doi.org/10.12688/openreseurope.16017.1

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

Rgyalrongic languages are mainly spoken in Western Sichuan, China, albeit including a now extinct, medieval language, namely Tangut, attested in today’s Ningxia Province. For most non-native speakers, they are the most difficult branch of languages to learn in the Sino-Tibetan family, as they exhibit a large array of word formation strategies such as inflection and derivation. Their word complexity points to their old age and their value in finding out the origin of Sino-Tibetan. This database aims at gathering lexical information of Rgyalrongic languages, in preparation for future work on historical and evolutionary linguistics.

**Introduction**

Rgyalrongic languages are a Sino-Tibetan branch, mainly spoken in Rngaba Tibetan and Qiang Autonomous Prefecture in Sichuan, China. They are closely related to Lolo-Burmese languages. Apart from their modern varieties, which are mostly endangered or vulnerable, a medieval language, Tangut, is recently recognised to belong to said branch. 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. 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 Autonomous Region, are also Rgyalrongic languages. These “newcomers” are provisionally termed “Peripheral Rgyalrongic” in the following.

Rgyalrongic is one of the most morphologically conservative branches in the Sino-Tibetan family, and certainly the closest to Proto-Sino-Tibetan compared to all other varieties in China. Therefore, understanding the history of Rgyalrongic languages is vital for the study on the evolution of Sino-Tibetan. The phylogeny of Rgyalrongic with 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 order to infer the phylogenetic subgrouping of Rgyalrongic, a lexical dataset with high quality curation is inevitable. This database provides the first annotated resource for the phylogenetic analyses of Rgyalrongic languages. It includes twenty varieties in East, West and Peripheral Rgyalrongic, as shown in the map in Figure 1 and Table 1.

**Methods**

The entire 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 specifically designed and...
Table 1. Languages selected for the database.

<table>
<thead>
<tr>
<th>Languages ID</th>
<th>Language name</th>
<th>Glottolog</th>
<th>Longitude</th>
<th>Latitude</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bantawa</td>
<td>Kiranti Bantawa</td>
<td>bant1281</td>
<td>87.05</td>
<td>27.12</td>
<td>Doornebal (2009)</td>
</tr>
<tr>
<td>Daofu</td>
<td>rGyalrong Daofu</td>
<td>horp1240</td>
<td>101.12</td>
<td>30.98</td>
<td>Huang (1992)</td>
</tr>
<tr>
<td>Japhug</td>
<td>rGyalrong Japhug</td>
<td>japh1234</td>
<td>101.96</td>
<td>32.21</td>
<td>Jacques (2015)</td>
</tr>
<tr>
<td>Tangut</td>
<td>Tangut</td>
<td>tan1334</td>
<td>106.29</td>
<td>38.48</td>
<td>Li (1997)</td>
</tr>
<tr>
<td>WobziKhroskyabs</td>
<td>Khroskyabs Wobzi</td>
<td>eree1240</td>
<td>101.43</td>
<td>31.63</td>
<td>Lai (2017)</td>
</tr>
<tr>
<td>Zhaba</td>
<td>Qiangic Zhaba</td>
<td>zhab1238</td>
<td>101.06</td>
<td>30.54</td>
<td>Huang (1992)</td>
</tr>
<tr>
<td>MaerkangrGyalrong</td>
<td>rGyalrong Maerkang</td>
<td>situ1238</td>
<td>102.30</td>
<td>31.87</td>
<td>Huang (1992)</td>
</tr>
<tr>
<td>MuyaKangding</td>
<td>Muya-kangding</td>
<td>west2417</td>
<td>101.96</td>
<td>30.06</td>
<td>Huang (1992)</td>
</tr>
<tr>
<td>OldBurmese</td>
<td>OldBurmese</td>
<td>oldb1235</td>
<td>96.60</td>
<td>21.46</td>
<td>Dictionary</td>
</tr>
<tr>
<td>QueyuXinlong</td>
<td>Queyu Xinlong</td>
<td>quey1238</td>
<td>100.26</td>
<td>30.98</td>
<td>Huang (1992)</td>
</tr>
<tr>
<td>QueyuPubarong</td>
<td>Queyu Pubarong</td>
<td>quey1238</td>
<td>100.78</td>
<td>30.22</td>
<td>Guan Xuan's fieldwork</td>
</tr>
<tr>
<td>SiyueuwKhroskyabs</td>
<td>Siyuewu Khroskyabs</td>
<td>siya1242</td>
<td>101.42</td>
<td>31.75</td>
<td>Author's fieldwork</td>
</tr>
<tr>
<td>BragbarSitu</td>
<td>Bragbar Situ</td>
<td>situ1238</td>
<td>101.91</td>
<td>31.82</td>
<td>Zhang (2020)</td>
</tr>
<tr>
<td>Geshiza</td>
<td>Geshiza</td>
<td>horp1240</td>
<td>101.65</td>
<td>31.03</td>
<td>Honkasalo (2019)</td>
</tr>
<tr>
<td>GuanyinqiaoKhroskyabs</td>
<td>Guanyin</td>
<td>guan1252</td>
<td>101.66</td>
<td>31.78</td>
<td>Huang (2007)</td>
</tr>
<tr>
<td>NjorogsKhroskyabs</td>
<td>Njorogs Khroskyabs</td>
<td>yelo1242</td>
<td>101.86</td>
<td>31.82</td>
<td>Yin (2007)</td>
</tr>
<tr>
<td>Tshobdun</td>
<td>Tshobdun</td>
<td>tsho1240</td>
<td>101.83</td>
<td>32.21</td>
<td>Sun (2019)</td>
</tr>
<tr>
<td>NgyaltsuZbu</td>
<td>Ngyaltsu Zbu</td>
<td>zbu1234</td>
<td>101.74</td>
<td>32.15</td>
<td>Gong (2018)</td>
</tr>
<tr>
<td>Zlarong</td>
<td>Zlarong</td>
<td>zlar1234</td>
<td>98.09</td>
<td>29.93</td>
<td>Zhao (2019)</td>
</tr>
<tr>
<td>KyomkyoSitu</td>
<td>Kymokyo Situ</td>
<td>situ1238</td>
<td>102.04</td>
<td>31.99</td>
<td>Prins (2017)</td>
</tr>
<tr>
<td>MazurStau</td>
<td>MazurStau</td>
<td>dao1238</td>
<td>101.05</td>
<td>31.04</td>
<td>Gates (2021)</td>
</tr>
</tbody>
</table>

Figure 2. Workflow of building up the Rgyalrongic lexical database.
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.\footnote{21}

**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 has already been 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, without physical contact. An additional source of data was published dictionaries and word lists which were judged as reliable by the authors (see Table 1). Dictionaries and word lists typically contain word forms and translations in Chinese, French or English. Some word lists also provide morphological information and example sentences. 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. 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 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\footnote{22}). Maerkang (Maerkang\textit{Gyalrong}), Bragbar Situ (Bragbar\textit{Situ}) and Kyomkyo Situ (Kyomkyositu) share the glottocode “\textit{situ1238}”, however, they are distinct varieties of Situ with limited intelligibility. The two dialects of Minyag sharing the glottocode “\textit{west2417}”, labelled Muya\textit{Kangding} and Menya\textit{Gao}, are closely related dialects with minor differences. Queyuxinlong and Queyu\textit{Pubarong} (quey\textit{1238}) 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 \textit{et al.}\textsuperscript{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 to 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 \textit{et al.}\textsuperscript{2} is employed as a guideline of our word selection in each language, including 313 concepts linked to Concepticon\textsuperscript{24} which provides a unique identifier to all concepts and thus largely 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 particular, we added concepts having a wide coverage in Rgyalrongic languages which are however not widely distributed from the perspectives of the entire Sino-Tibetan family. For instance, we use the general concept for ‘person, human’ instead of ‘the man (male human)’ used in Sagart \textit{et al.}\textsuperscript{1}, which has been shown to be indicative of language subgrouping by Lai\textsuperscript{25}; we also included ‘girl’, as a significant innovation in West Gyalrongic with an \textit{s}-prefix (compare Stau (West) \textit{s-mi} and Japhug \textit{me} (East)), discussed in Lai \textit{et al.} \textcite{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 as worthwhile to be included 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, 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.cldf.org, \textsuperscript{21,26–28}) and set up an orthography profile\textsuperscript{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 \textit{foot} and German \textit{fuß}, both originating from Proto-Germanic \textit{*fōts}. Cognate forms in daughter languages can be deducted through regular sound rules from the proto-form. In Sino-Tibetan languages, more often than not, we find cognates in word parts rather than in entire words. For instance, Mandarin Chinese \textit{yuē-liäng} ‘moon’ and Taiwanese Southern Min \textit{guēh-níu ê} ‘moon’ only share the first part \textit{yuē} and \textit{guēh}, respectively, as cognates. This
type of cognacy is termed “partial cognacy”. Partial cognates enable us to segment word forms into morphemes, which improves the accuracy of the computation of language subgrouping. It is thus essential to annotate partial cognates, rather than full cognates, in our Rgyalrongic database. Partial cognate identification is conducted manually with the expert knowledge of the authors, using the web-based EDICTOR tool (https://digling.org/edictor, 30,31).

Statistics
The current dataset contains a total of 6335 word forms for 22 distinct language varieties. Word forms correspond to 305 different concepts, and use a total of 413 distinct sounds, with an average inventory size of 72 different sounds per language variety. The word forms have been morphologically segmented, comprising a total of 9116 morphemes. 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.

Quality control
We carefully verified the data to ensure the accuracy and correctness. We use our expert knowledge to review every lexical entry in the database. 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 reuses. Phonemic and morphemic segmentation, as well as cognate judgements are carefully processed with expert knowledge. 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. Although there exist searchable databases such as STEDT32 (https://stedt.berkeley.edu/) and the rGyalrongic Language Database33 (https://htq.minpaku.ac.jp/databases/rGyalrong/), the present database is the first Rgyalrongic lexical database that involves expert data curation and cognate annotation, and the only one that is ready for phylogenetic analyses.

For now, only those morphemes are assigned to the same cognate set which occur in words sharing the same meaning. In the future, we hope to extend this analysis to account for cognates across meaning slots, specifically concentrating also on language-internal partial cognates along the lines of the analysis pioneered in Hill and List21 and further extended in Schweikhard and List34. Having annotated the data in this form, cognacy can also be annotated on the word level35 and computational approaches to phylogenetic reconstruction of Rgyalrongic (and beyond) can be carried out. Thus, the present

<table>
<thead>
<tr>
<th>DOGULECT</th>
<th>CONCEPT</th>
<th>TOKENS</th>
<th>ID-7567</th>
<th>ID-7579</th>
<th>ID-7566</th>
<th>ID-7615</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wobzi</td>
<td>Yesterday</td>
<td>snəɣ</td>
<td>y</td>
<td>snə</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tangut</td>
<td>Yesterday</td>
<td>jinjì:</td>
<td></td>
<td>njì:</td>
<td>ji</td>
<td></td>
</tr>
<tr>
<td>Zhaba</td>
<td>Yesterday</td>
<td>jìne</td>
<td></td>
<td>ne</td>
<td>jì</td>
<td>sə</td>
</tr>
<tr>
<td>Muya</td>
<td>Yesterday</td>
<td>jìse</td>
<td></td>
<td></td>
<td>jì</td>
<td>sì</td>
</tr>
<tr>
<td>Siyuewu</td>
<td>Yesterday</td>
<td>xsnə</td>
<td>x</td>
<td>snə</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zlarong</td>
<td>Yesterday</td>
<td>jisi</td>
<td></td>
<td></td>
<td>ji</td>
<td>si</td>
</tr>
</tbody>
</table>

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). Languages differ in the choices of the two parts. Annotating the two parts separately allows us to visualise and analyse the internal morphology of the forms for more accurate inference of phylogeny.
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/.

Archived source code and data at time of publication: https://doi.org/10.5281/zenodo.7866796

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 to our questions. We also thank Mei-Shin Wu for her help in refining the data.

References

   Reference Source
   Reference Source
   Published Abstract | Publisher Full Text | Free Full Text
   Publisher Full Text
   Published Abstract | Publisher Full Text
   Reference Source
    Reference Source
    Reference Source
    Reference Source
   Reference Source
   Reference Source
   Publisher Full Text
   Publisher Full Text
   Publisher Full Text
   Publisher Full Text
   Publisher Full Text
   PubMed Abstract | Publisher Full Text | Free Full Text
   Publisher Full Text
   Publisher Full Text
   Reference Source
   Publisher Full Text
   Publisher Full Text
   Reference Source
   Reference Source
   Reference Source
   Publisher Full Text
Open Peer Review

Current Peer Review Status: ?

Version 1

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
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), Lavrun (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»
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

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/2021EHESS0054), or a direct link to the PDF, though the latter option is likely to prove less ‘time-proof’ (https://www.theses.fr/2021EHESS0054/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.