Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

Digitizing Legacy Lexical Data of Muishaung for Computer-Assisted Language Comparison

Kellen Parker van Dam
Chair for Multilingual Computational Linguistics
University of Passau

This study describes the process of digitizing legacy materials into a computer-readable format for the purposes of computational typology and computer-assisted historical reconstruction. It presents a comparative wordlist that is made available in the formats recommended by the Cross-Linguistic Data Formats initiative.

1 Needham's "*A collection of a few Môshâng Nâgâ words*"

The original publication, *A collection of a few Môshâng Nâgâ words* (Needham 1897) provides the first written account of Muishaung (Glottolog: mosa1240), a Tibeto-Burman language spoken Arunachal Pradesh, India as well as neighbouring regions of Sagaing Region, Myanmar. Data were collected during a trip made by Needham to the Muishaung area in 1881. Today Muishaung is estimated to have around 2000 speakers in India, with an uncertain population size in Myanmar. This is one of a handful of texts written by Needham at the time. The others include similar descriptions of Tai Khamti and Singpho.

Needham's text includes 264 concepts — although with some minor misunderstandings around which concepts were actually being provided by speakers — along with basic grammatical notes on topics such as gender, verbal morphology and the pronominal system. Aside from being the earliest known account of any Tangsa-Nocte variety, Needham's text provides valuable insights into historical sound change in the region; Muishaung is one of over two-dozen closely related language varieties within Tangsa-Nocte, but significantly it is the most phonologically divergent of the group. For example, it has innovative dental stops, including a split between /n/ and /n̪/ not seen

Computer-Assisted Language Comparison in Practice
Volume 8, Number 1
URL: https://calc.hypotheses.org/8624 DOI: 10.15475/calcip.2025.2.1

Published on 23/07/2025

elsewhere in the group, analogous to the /t/ and /t̪/ pair as reflexes of *t and *ð respectively. It also shows a split between /g/ and /ɣ/ as reflexes of *ɣ, and finally, Muishaung has been undergoing a process of vowel fracture not seen elsewhere among the closely related varieties. Access to historical texts such as Needham's allow us to better understand the timing of such changes and the possible mechanisms behind them. For this reason, proper analysis of such descriptions is important. By developing computer-readable versions of such data sets, we are better able to include them in comparative work.

With somewhat inconsistent orthographic conventions (many of which are never described by the author) along with some clear mistakes in the data which were elicited, the full value of Needham's text is not apparent without a more in-depth investigation. This was done in van Dam & Mossang (2025), which analyzed the entirety of the text. This study undertook careful investigation into the terms given through comparison to modern-day forms of the concepts and reconstructed proto-forms. The full study is available as an open-access publication in the *Journal of Asian and African Studies* published by the Tokyo University of Foreign Studies.

2 Data Availability as a CLDF Data Set

In the process of analyzing the text, a digital version of the lexical data along with terms found in the grammatical notes was created in a flat tabular format. These were transcribed exactly as they occurred in the original text, including the use of circumflexes for marking distinctions in vowel quality and an underlined ⟨n⟩ for vowel nasalization. Needham's transcription was then converted to IPA and paired with modern-day pronunciations for cases in which cognates are attested today. In those instances where no modern-day cognate was found, the term which replaced the form in Needham's time was given.

In an additional step, the data from the flat tabular format were converted to the formats recommended by the Cross-Linguistic Data Formats (CLDF) initiative (Forkel et al. 2018, https://cldf.clld.org), using the workflow for the handling of comparative wordlists developed for the Lexibank repository (List et al. 2022, Blum et al. 2025, https://lexibank.clld.org).

The CLDF dataset also includes all comments given in the original text, of which where were a few, along with notes by the authors of the 2025 study indicated cases where Needham may have elicited a term other than what was intended. For example in eliciting 'flea', the term given was actually one for 'cat', perhaps the result of gesturing toward a flea-infested feline at the time of elicitation.

An example of the Forms table, with some columns removed here for the sake of saving space, as seen in Table 1.

Local_ID	Form	Segments	Comment	Source
MuishaungNeedham-1_above-1	rʌŋ	r ʌ ŋ	Shâng´gê is distant from Môshâng about 20 miles, much less as the crow flies. J.N.	Needham1897
MuishaungModern-1_above-1	rɐuŋ$_2$	r ɐu ŋ $_2$/^{231}		VanDam2025
MuishaungNeedham-2_acid-1	ɑ.hiˀ	ɑ + h i ˀ/ʔ		Needham1897
MuishaungModern-2_acid-1	ə$_0$hi$_2$	ə $_0$/^{0} h i $_2$/^{231}		VanDam2025
MuishaungNeedham-3_all-1	wʌ.tɒŋ	w ʌ + t ɒ ŋ		Needham1897
MuishaungModern-3_all-1	βə$_0$tɐuŋ$_2$	β ə $_0$/^{0} t ɐu ŋ $_2$/^{231}		VanDam2025

Table 1: Form table of the CLDF dataset.

An orthography profile was also created, and all concepts were mapped to the corresponding CONCEPTICON IDs (List et al 2025) where applicable. Language varieties are also linked to their corresponding Glottocodes (Hammarström et al 2025) to facilitate language identification.

3 Next Steps

In addition to forms given for Muishaung, the original text also includes a number of words from the Shecyü variety under the name Shâng´gê, Needham's representation of the common exonym Shangke (Glottolog: sank1250). Occasionally forms are also given for Singpho, a distantly related Tibeto-Burman language, as well as Tai Hkamti, a Kra-Dai variety. In addition to Needham's Muishaung description, he also published texts on Singpho and Tai Hkamti, and was thus knowledgable about both. He regularly included reference to these languages in cases where he felt the term in Muishaung was borrowed from one or the other, although in some cases with Singpho he was simply identifying cognate terms.

In a future version of the data set, these forms will be fully encoded with both Needham's orthographic representation and its corresponding IPA form, as well as the modern-day equivalents. This is intended to be published as an update to the current dataset in order to further support computer-assisted cross-linguistic comparative work.

References

Blum, Frederic, Carlos Barrientos, Johannes Englisch, Robert Forkel, Simon Greenhill, Christoph Rzymski, and Johann-Mattis List (2025): Lexibank 2: pre-computed features for large-scale lexical data [version 2; peer review: 3 approved]. Open Research Europe 5.126. 1-19. https://doi.org/10.12688/openreseurope.20216.2

van Dam, Kellen Parker and Kelim Mossang, Wanglung. (2025a). A Classified Account of J. F. Needham's A Collection of A Few Môshâng Naga Words. In: Journal of Asian and African Studies 2025 (109), 111-145. https://doi.org/10.57275/ilcaajaas.2025.109_111 van Dam, Kellen Parker and Kelim Mossang, Wanglung (2025b). Supplementary materials for van Dam & Kelim Mossang 2025 [Data set, Version 1.0.0]. In Journal of Asian and African Studies (Vol. 109, pp. 111–145). Zenodo. https://doi.org/10.5281/zenodo.14053893

Forkel, Robert, Johann-Mattis List, Simon Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon Kaiping, and Russell D. Gray (2018): Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. Scientific Data 5.180205. 1-10. https://doi.org/10.1038/sdata.2018.205

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank (2025): Glottolog [Dataset, Version 5.2], Leipzig: Max Planck Institute for Evolutionary Anthropology. https://glottolog.org

List, Johann Mattis, Annika Tjuka, Frederic Blum, Alžběta Kučerová, Carlos Barrientos, Christoph Rzymski, Simon Greenhill, and Robert Forkel (2025): CLLD Concepticon [Data set, Version 3.4.0]. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://concepticon.clld.org

List, Johann-Mattis, Robert Forkel, Simon Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray (2022): Lexibank, A public repository of standardized wordlists with computed phonological and lexical features. Scientific Data 9.316. 1-31. https://doi.org/10.1038/s41597-022-01432-0

Needham, J. F. (1897). A collection of a few Moshang Naga words. Shillong: Assam Secretariat Printing Office. https://archive.org/details/collectionoffewm00needrich

Supplementary Material
Code and data are curated on GitHub (https://github.com/phonemica/needhammuishaung, Version 1.0.0) and archived with Zenodo (https://doi.org/10.5281/zenodo.14053893).