Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

Creating a Standardized Comparative Wordlist of Newari Varieties

Zhenyang Liu[1], Guillaume Jacques[1], Johann-Mattis List[23]
CRLAO[1], DLCE[2], Chair of Multilingual Computational Linguistics[3]
EHESS Paris[1], MPI-EVA[2], University of Passau[3]

Newari is one of the few ancient Sino-Tibetan languages attested in written texts. Since previous studies on the phylogeny of Sino-Tibetan did not take Newari data into account, we felt it is important to close this gap by providing an up-to-date comparative wordlist of Newari varieties. This wordlist has now been finalized in a first version that has additionally been standardized following the recommendations of the Cross-Linguistic Data Formats initiative.

1 Introduction

There are not many ancient languages in the Sino-Tibetan language family reflected in written sources. Apart from Tibetan and Burmese, dating back no more than 1000 years, there is Chinese and Tangut, but both present their own difficulties due to the non-phonetic nature of their writing systems. Newari is one additional language that has been attested in written sources from around the 12th century of our time. Given that it was not included in the phylogenetic analysis of Sino-Tibetan languages by Sagart et al. (2019), which is still the most complete dataset with respect to the degree by which cognates have been annotated, we considered it important to close the gap by preparing a comparative wordlist of Newari varieties that could be later used in a larger study with more language varieties, which we hope to find time to carry out in the future.

As mentioned, first inscriptions in Newari date back to the 12th century of our time, but earlier Newar words appear in Sanskrit inscriptions in the centuries before. Being spoken by about a million people up to today, the language is one of the larger representatives of the Sino-Tibetan languages. Since its structure meets the criteria for languages selected in the study by Sagart et al., it is also a crucial source worthwhile to be added to additional phylogenetic investigations.

Computer-Assisted Language Comparison in Practice
Volume 6, Number 2, 53–58
URL: https://calc.hypotheses.org/6229/ DOI: 10.15475.calcip.2023.2.1

Published on 26/07/2023

Due to the close geographic contact with speakers of Indic languages, there are numerous loanwords in Newari varieties, which need to be carefully identified and marked as such when trying to compile a comparative wordlist.

The preparation of the Newari comparative wordlist was carried out in two major stages. First, the first author of this study compiled the wordlist in the form of a spreadsheet using an individual annotation format. Second, computational methods were used to convert the data into the tab-separated format used by the LingPy software library (List et al. 2022) that allows to inspect and correct the data with the help of the EDICTOR tool (List 2022). From there, the data was converted to Cross-Linguistic Data Formats (CLDF) as required for inclusion into the Lexibank repository (List et al. 2022).

In the following, we will quickly provide information on the data sources, the initial data preparation, and the conversion to LingPy and CLDF formats.

2 Data Sources

This Newari wordlist comprises four varieties of Newari, including:

- Old Newari: the oldest stage of the language, predating the 16th century CE.
- Classical Newari: 16th-18th centuries CE.
- Kathmandu: the modern dialect with the most native speakers.
- Dolakha dialect: a rather conservative modern dialect with some features that are absent even in Old Newari.

The vocabulary of Old Newari and Classical Newari is collected from two dictionaries, the Dictionary of Classical Newari (DCN, Nepal Bhasa Dictionary Committee 2000) and the Dictionary of the Classical Newari by Jørgensen (1936). Both dictionaries contain data from the two earlierst periods. The DCN is the result of two decades' work by a dozen scholars with over 30,000 words from 96 documents, while Jørgensen's dictionary contains more than 6,000 words from eighteen manuscripts.

For the Kathmandu dialect, all the data come from Kölver and Shresthacarya's Dictionary of Contemporary Newari (1994). The dictionary is based on extensive collections which were assembled over quite a number of years chiefly in Kathmandu and in Pāṅgā, a village between Kathmandu and Kīrtipur.

Information on Dolakha was taken from A Grammar of Dolakha Newar by Genetti (2007), an exceptional reference grammar with a concise wordlist at the end. However, due to the limited content, some words such as "dirty" are absent and could therefore not be elicited for the current comparative wordlist.

3 Initial Data Preparation

The initial preparation of the data consisted in a spreadsheet that was specifically organized in such a way that cognate words would be listed in the same row, with the information for individual language varieties of Newari listed on three columns per variety. For each variety, the first column indicated the word as it was found in the source, the second column provided a transcription in IPA, and the third column provided the reference, including the page in the book where the word had been found.

The very first column of the spreadsheet listed the concept (using the concepts employed by Sagart et al. 2019). The second column provided a tentative root form for the cognate set, or more exactly, the basic form for hte most prominent word in the cognate set. This showed to be particularly useful for verbs, because the final consonants of a lot of verbs are "blurred" synchronically due to the suffixes attached directly to the root. For example, mvāt- 'to be alive' has as infinitive mvāya but the form mvātasā meaning 'if lived'. The final consonants of some verbs only resurface in specific forms. Therefore, it is necessary to determine the final consonant of the verb by examining the entire paradigm, as this is crucial for its comparison with other languages. At the same time, several verbs weren't well-attested, like ṅāye 'to walk' in Old Newari, the infinitive form being the sole occurence.

One additional column at the end of the spreadsheet was used to mark borrowings. Here, great care was taken to identify all borrowings from Indic languages, for which also source forms are provided, and the last column was used for individual notes.

Until the 19th century, virtually all manuscripts were written in nepālākṣara, which is quite similar to the devanāgarī script. The dictionaries used employ the International Alphabet of Sanskrit Transliteration (IAST) for transcription, except for the usage of visarga. A sequence like Vḥ denotes a long vowel resulting primarily from compensatory lengthening. Since scripts for denoting long i and u already exist, this usage of visarga only appears with a and ā. In Old Newari there are only a few examples for this usage, but it's frequently seen in Kathmandu, since a huge amount of apocope and compensatory lengthening took place.

The IPA transcriptions for the words in Kathmandu and Dolakha are given according to Genetti (1988 and 2007) and Kölver (1994). As for the older stage of the language, the synchronic phonology of Classical Newari is already addressed in Jørgensen's grammar of Classical Newari, as well as in Otter (2021).

The situation of Old Newari is a little more tricky, as a grammar of Old Newari is still an urgent desideratum. However, having been working on its synchronic phonology for a while, our first author employed internal reconstructions as well as comparison with other Sino-Tibetan languages for the IPA transcription here. For instance, although the retroflex stops are thought to be inexistent in Classical Newari, they should be distinct

phonemes in Old Newari, based on cognates with other Sino-Tibetan languages and Dolakha, as Dolakha still has them as distinct phonemes. As a result, they were marked as such in the IPA transcription.

4 Conversion to EDICTOR and CLDF

Having prepared the original spreadsheet, only minimal modifications had to be made in order to prepare the conversion script with which we turned the data into the formats proposed by the Cross-Linguistic Data Initiative (Forkel et al. 2018). Since details of these conversion scripts have been discussed before, we won't discuss the details here, but refer interested readers to the GitHub repository, where the code can be found (https://github.com/lexibank/liunewari).

What should be mentioned is that our conversion code revealed some small problems in the original table, such as some concepts that were spelled differently in the spreadsheet compared to the original version of the concept list underlying Sagart et al. (2019) in the Concepticon (List et al. 2023). We also found that there was no need to use orthography profiles (Moran and Cysouw 2018) for the conversion of the IPA transcriptions to the standardized version of the IPA in the Cross-Linguistic Data Formats reference catalogue (List et al. 2022), since there were only a couple of sounds that caused problems, such as the representation of [j] as y. As a result, only a couple of replacement statements were added to the conversion script.

The resulting dataset now contains 837 distinct lexemes, distributed over 4 language varieties and 180 concepts, with a total of 65 different sound segments and 45 segments on average per variety. Apart form the CLDF version, we offer also a TSV that can be directly edited in the EDICTOR tool (https://digling.org/edictor).

5 Outlook

We hope that the comparative wordlist of Newari varieties presented here can fill a gap in phylogenetic studies on Sino-Tibetan languages. Given that it is integrated with the major standards and reference catalogues produced as part of the CLDF initiative, it should be easy to integrate the data with different sources in future work.

The dataset is curated on GitHub (https://github.com/lexibank/liunewari) and has been archived in Version 1.0.0 with Zenodo (https://doi.org/10.5281/zenodo.8169353).

References

List, Johann-Mattis and Anderson, Cormac and Tresoldi, Tiago and Forkel, Robert (2021): Cross-Linguistic Transcription Systems. Version 2.1.0. Jena:Max Planck Institute for the Science of Human History.

Nepal Bhasa Dictionary (2000): A Dictionary of Classical Newari. Compiled from Manuscript Sources. Kathmandu: Cwasā pāsā.

List, Johann-Mattis and Tjuka, Annika and van Zantwijk, Mathilda and Blum, Frederic and Barrientos Ugarte, Carlos and Rzymski, Christoph and Greenhill, Simon J. and Robert Forkel (2023): CLLD Concepticon [Dataset, Version 3.1.0]. Leipzig:Max Planck Institute for Evolutionary Anthropology. https://concepticon.clld.org

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

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

Carol Genetti (1988): A contrastive study of the Dolakhali and Katmandou Newari dialects. Cahiers de linguistique -- Asie orientale 17.2. 161–191.

Carol Genetti (2007): A Grammar of Dolakha Newar. Berlin and New York: Mouton de Gruyter.

Jørgensen, Hans (1936): A Dictionary of the Classical Newari. Copenhagen.

Kölver, Ulrike and Shresthacarya, Iswarananda (1994): A Dictionary of Contemporary Newari. Newari -- English. Bonn:VGH Wissenschaftsverlag.

List, Johann-Mattis and Forkel, Robert (2022): LingPy. A Python library for quantitative tasks in historical linguistics [Software Library, Version 2.6.9]. Leipzig:Max Planck Institute for Evolutionary Anthropology. https://lingpy.org

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

Moran, Steven and Cysouw, Michael (2018): The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles. Berlin:Language Science Press.

Otter, Felix (2021): A Course in Reading Classical Newari: Selections from the Vetālapañcaviṃśati. Heidelberg and Berlin: CrossAsia.

Sagart, Laurent and Jacques, Guillaume and Lai, Yunfan and Ryder, Robin and Thouzeau, Valentin and Greenhill, Simon J. and List, Johann-Mattis (2019): Dated language phylogenies shed light on the ancestry of Sino-Tibetan. Proceedings of the National Academy of Science of the United States of America 116. 10317-10322.