Challenges to Representing Personal Names and Language Names in Language Archives: Examples from Northeast India

Mary Burke[†]
Department of Information Science
University of North Texas
Denton TX, USA
Mary.Burke@unt.edu

Shobhana Chelliah
Department of Linguistics
University of North Texas
Denton TX, USA
Shobhana.Chelliah@unt.edu

ABSTRACT

Language archives are not only a valuable resource for language communities to tell their stories and to create lasting records of their ways of life, but also for those interested in anthropology, linguistics, agriculture, or art history. This recent emphasis on archiving primary datasets in linguistics has resulted in an abundance of datasets online; however, of the languages of South Asia, only a small percentage are represented in digital language archives or described thoroughly. Though several of these languages are being documented, this material is at risk of being lost or inaccessible without concerted attention paid to long-term preservation. There are several obstacles to documenting and archiving language materials from this area, including political instability and lack of access to infrastructure. This submission reviews one particular challenge to data management relevant to South Asia, which is the complexity of names (of individuals, groups, and languages). We provide examples from Northeast India and recommendations based on experience from CoRSAL (Computational Resource for South Asia).

KEYWORDS

Personal names, language names, name authority control, information organization in language archives

1 Introduction

Language archives are not only a valuable resource for language communities to tell their stories and to create lasting records of their ways of life, but also for those interested in anthropology, linguistics, agriculture, or art history. This recent emphasis on archiving primary datasets in linguistics has resulted in an abundance of datasets online; however, of the languages of South Asia, only a small percentage are represented in digital language archives or described thoroughly [1]. Though several of these languages are being documented, this material is at risk of being lost or inaccessible without concerted attention paid to long-term preservation. There are several obstacles to documenting and

archiving language materials from this area, including political instability and lack of access to infrastructure. This submission reviews one particular challenge to data management relevant to South Asia, which is the complexity of names (of individuals, groups, and languages). In this paper, we provide examples from Northeast India and recommendations based on experience from CoRSAL (Computational Resource for South Asia).

Name authority control is a crucial component of information organization whereby names are represented uniformly across metadata records despite having multiple possible variants. For example, the Library of Congress Name Authority File (LCNAF) cites 'Zin, Hauard, 1922-2010' as a variant of the authorized name form 'Zinn, Howard, 1922-2010'. Implementing name authority control improves the recall and accuracy of search results. This can be done with a public name authority file, such as the LCNAF or the Virtual International Authority File (VIAF), or using a local name authority file (e.g., a local spreadsheet). Few linguists, students, and language community members have records in a public name authority files, so, in the case of language archives, it is more common to use locally-developed name authority files [2].

2 Variation in personal names

Common practice in metadata records is to represent personal names in an inverted structure with the name parts appearing in the order "Family, Given Middle" -- as in, Turing, Alan Mathison. However, for many communities worldwide, personal names include village names, caste names and/or clan names, and come with a host of variations including abbreviations, nicknames, and differing orders of name parts based on convenience or context. Examples of variation in personal names are described in literature on this topic [3, 4, 5]. This section reviews some examples of variation in personal names through examples from Northeast India.

Name structure may change based on personal preferences, but could also be influenced by large-scale factors like political or religious affiliations, and the impact of colonization. [6] describes

how naming conventions of the Meitheis changed as a result of the ruler Pamheiba (1709-48) introducing Hindu traditions to Manipur, and again during the Manipur renaissance to assert ethnic identity. Beyond just Manipur, the effects of colonization can be seen on personal names when a Western convention is imposed on a previously uncodified system.

2.1 Ordering of name parts

[6] reviews some examples from Northeast India which are relevant here. Following the effects of the Hindu influence of the 1700s, in Manipur, Kshatriya caste names were adopted. For women, 'Devi'; for men, 'Singh'. Thus, the typical name structure for the Meithei community would be:

FAMILY NAME - GIVEN NAME - CASTE NAME
Thounaojam - Harimohon - Singh

However, these name parts may appear in any of the following orders:

FAMILY NAME - GIVEN NAME - CASTE NAME
GIVEN NAME - FAMILY NAME - CASTE NAME
FAMILY NAME - GIVEN NAME
GIVEN NAME - FAMILY NAME

Each order is possible, and the choice seems to depend on personal preference given a certain situation or purpose. Thounaojam Harimohon Singh, for example, might sign an informal letter *Harimohon Thounaojam*. He signed the acknowledgments section of his dissertation *Th. Harimohon Singh*. When buying a plane ticket, however, one needs to ensure the name matches what is written on their passport, and this may yield yet another ordering of the name parts.

2.2 Inclusion, omission, and abbreviation of name parts

Another situation is illustrated by nicknames used by the Lamkangs of Manipur. In this community, individuals are sometimes identified by nicknames or birth order names (e.g., first, second, third born). They may also have Lamkang given names and additional Christian names given at the time of baptism.

Take, for example, a young Lamkang speaker recorded as part of the Lamkang project in 2017. In conversation, he introduced himself as 'Koko,' a nickname assigned based on birth order. On the IRB paperwork, he signed his full name, 'William Shilshi.' In preparing the recordings for archiving, the research team had difficulty choosing the 'correct' name to include in metadata records because this speaker was identified only as 'Koko' except for on the formal paperwork.

It is useful to know that these conventions and styles of identification exist so that documenters can properly keep track of individuals' personal names. For example, for Lamkang names, it might be useful to provide a known nickname in parentheses after the given name: *William (Koko) Shilshi*. Both the given name and

the nickname are needed to identify the individual, as many people may have the same nickname.

Another example from the Lamkang community involves individuals taking on additional names for the purpose of affirming identity (e.g., *Swamy Tholung **Ksen***). In this example, *Ksen* is neither a given name nor part of the birth order naming system described in this paper, but was adopted to represent the individual's pride to be Lamkang—the Lamkang refer to themselves as *ksen mi*, 'the red people.'

Alternatively, some individuals who take part in language documentation projects do not provide their full name to the collector of the material, but rather only a first name. This may cause confusion when depositors prepare for archiving because it is not clear whether the individual is distinct from another with the same first name.

Name parts may also be excluded for convenience. For instance, because all Meithei males assumed the caste name '*Singh*', one may choose to omit the '*Singh*' to avoid confusion. Also for convenience, or to save space, the family name may be abbreviated, as seen in the following examples ([6], p. 3):

- *L. Bhima Singh*
- *N. Promodini Devi*
- *W. Ibemu Devi*
- *M. Kirti Singh*
- *Ch. Yashawanta Singh*

It is important to note that the same abbreviation might be used for different family names (e.g., both *Layrikyengbam* and *Leimapokpam* are abbreviated 'L.').

This section has illustrated the complexity of naming systems with two examples from Northeast India: the Meithei community, and the Lamkang community. Still, this is by no means a comprehensive review of every possibility—every group will have their own practices surrounding personal names. The following section briefly reviews the multiple forms of language names.

3 Language names

Much of the early recorded history in Northeast India was created by colonial officials, resulting in inaccurate or incomplete names for languages (called *glossonyms* in [7]) and the groups of people who speak them (*ethnonyms*). [8] notes that a language name may be falsely recorded as the name of the village where it is spoken, or as an outsiders' name for the group (*exonym*). There may be one term used for the group of speakers, the village where the language is spoken, and the language itself (*glossonym*), or multiple names for each.

Further, any one language name is likely to have multiple spellings or pronunciations (*allonyms*) or variants (*allograms*) [7]. Spelling differences can be attributed to Westernizations, or differences in transliterations and transcriptions. Take, for example, pairs like *Tsʹairel* and *Chairel*, or *Hlota* and *Lhota*. The name used for a language may change depending on individual preferences, or over

time. See [7, 8] for extensive discussion of this phenomenon specific to the Tibeto-Burman language family.

Generally, one would rely on a standard to navigate the variants and select the 'correct' name for a language. The primary standards for representing language names are the ISO 639-2 and 639-3 codes, developed by the Summer Institute of Linguistics (now called SIL International), a Bible translation organization. Though this standard is used by the linguistics community at large, it is maintained only by SIL International. Many take issue with the ISO 639-3 standards [9, 10, 11, 12] and, depending on the languages in focus, prefer to use Glottocodes or a regional standard like AUSTLANG instead.

4   Discussion and Recommendations

This paper has briefly demonstrated the ways in which personal names and language names may vary across contexts. To summarize, we offer some basic recommendations to maximize consistency when representing these names in archival collections.

For archival staff:
- When reviewing metadata, ask depositors about personal names
- In the case that two names look similar, it is better to ask for clarification than to assume this is a typo. (e.g., Abdullah and Abdulla may refer to two separate individuals.)

For language documenters/ depositors:

- Understand naming conventions used in the area. What strategies do people use to identify themselves and each other, and what factors affect this?
- Make note of all possible name forms, and make these clear in the consultant or language contributor records (e.g., spreadsheets, SayMore).
- Include pictures of contributors as recommended in [13].
- If abbreviations are common in the naming convention, be aware of the expanded names from which these abbreviations arise.

5   Conclusion

Given that there is such variation, when it comes time for archiving, it is necessary for depositors and archival staff to be familiar with different naming conventions that may apply to the language or region in question. These issues need not prevent the languages of South Asia from the level of digital accessibility that archives provide. With a better understanding of the structure of personal names, we can represent personal names as consistently and accurately as possible.

REFERENCES

[1]   Post, M. 2020. The Eastern Himalayan ethno-linguistic diversity hotspot: Where is it, why is it significant, why is it endangered, and what should "we" do about it? ISSN: 2249-1511.

[2]   Burke, Zavalina, O. L., Phillips, M, Author. Organization of knowledge and information in digital archives of language materials. Journal of Library Metadata. doi: 10.1080/19386389.2020.1908651

[3]   Akinnaso, F. N. 1980. The Sociolinguistic Basis of Yoruba Personal Names. Anthropological Linguistics 22:275—304.

[4]   Abd-el-Jawad, H. 1986. A Linguistic and Sociocultural Study of Personal Names in Jordan. Anthropological Linguistics 28:80—94.

[5]   Hong, B. 1985. Politeness in Chinese: Impersonal Pronouns and Personal Greetings. Anthropological Linguistics 27:204—13.

[6]   Chelliah 2005. Asserting nationhood through personal name choice: The case of the Meithei of northeast India. Anthropological linguistics, 169-216.

[7]   Matisoff, J. A. 1996. General introduction. In Matisoff, J. A., Baron, S. P., & Lowe, J. B. eds. (1996). Languages and dialects of Tibeto-Burman. Sino-Tibetan Etymological Dictionary and Thesaurus Project, Center for Southeast Asia Studies, University of California, Berkeley.

[8]   Marrison, G. E. 1967. The classification of the Naga languages of North East India. (Doctoral dissertation, SOAS, University of London).

[9]   Dobrin, L. M., & Good, J. 2009. Practical language development: Whose mission? Language, 85(3), 619-629.

[10]   Epps, P. et al. 2006. Letter of opposition to adopting Ethnologue's language codes for ISO 639-3. SSILA Bulletin, 246.

[11]   Errington, J. 2008. Linguistics in a colonial world. Oxford: Blackwell.

[12]   Morey, S., Post, M. & Friedman, V. A. 2013. The language codes of ISO 639: A premature, ultimately unobtainable, and possibly damaging standardization. Paper presented at the Workshop Research, records and responsibility: Ten years of the Pacific and Regional Archive for Digital Sources in Endangered Cultures. Melbourne,   University   of   Melbourne https://ses.library.usyd.edu.au/handle/2123/9838

[13]   Everett, D. & Sakel, J. 2012. Linguistic Fieldwork. Cambridge University Press, Cambridge.