himalayan linguistics

A free refereed web journal and archive devoted to the study of the languages of the Himalayas

# Himalayan Linguistics

*Results from the Linguistic Survey of Sikkim*

**Samopriya Basu[1] & Mark Turin[2]**
[1]Carleton University, [2]The University of British Columbia

**ABSTRACT**

This article re-examines a 2005-2006 school-based survey of language use across Sikkim, a multilingual Himalayan state in northeastern India. 16,527 students in classes VIII–XII from 105 schools answered questions about languages used with grandparents, parents, and siblings, and which they considered to be their "mother tongue". Two patterns emerge from our reanalysis. Heritage languages are still used with elders, showing family-based maintenance. Yet everyday conversations among younger people increasingly shift to other languages—often those used in school and public life—especially in conversations between siblings. About one in nine parent pairs (11.24%) no longer use any of their parents' languages, and about one in eight sibling groups (12.97%) use none of their parents' languages. The "mother tongue" that students officially reported through the survey often differs from what they speak, highlighting a gap between linguistic identity and language practice. There are also regional variations: North Sikkim shows the strongest continuity; East and South show faster language shift; and the West sits somewhere between. Although school-based sampling and simple matching limit what we can deduce about language use, the directionality of language shift is clear: even though the parent generation know the languages of the grandparent generation and use them to communicate across the two generations, in many cases, they use other language(s) to communicate within their own generation. This is a significant finding, and the same pattern continues to the next generation. Going forward, we recommend strengthening vernacular teaching, expanding spaces for heritage languages, and repeating this survey each decade to track change across schools and communities, as well as to inform language policy.

**KEYWORDS**

Sikkim; India; linguistic survey; census; language shift

This is a contribution from *Himalayan Linguistics, Vol. 25(1): 1-18.*
ISSN 1544-7502

*Himalayan Linguistics, Vol. 25(1)*
ISSN 1544-7502

# *Results from the Linguistic Survey of Sikkim*

**Samopriya Basu[1] & Mark Turin[2]**
[1]Carleton University, [2]The University of British Columbia

## 1 Introduction

This study offers a statistical analysis of the Linguistic Survey of Sikkim (LSS), 2005-2006. The landlocked state of Sikkim, India's least populous and second smallest, has a geopolitical significance far beyond its size. Bounded to the north and northeast by the Tibet Autonomous Region of the People's Republic of China, to the west by Nepal, to the southeast by Bhutan, and to the south by the Darjeeling district of the Indian state of West Bengal, Sikkim occupies an important ecological niche along one of the oldest Himalayan trade routes. Much of the state's north and west is perpetually covered by snow and is dominated by the Kanchenjunga massif, India's highest and the world's third highest mountain.

Turin (2011) reports that in 2001, shortly before the survey, the population of Sikkim was a little over 540,000 residents. A slight majority of the population self-identified as Hindu in the 2011 Census of India, and a significant majority of the rest identified as Buddhist. Other religious groups make up smaller percentages. The 2011 census further returned Nepali as Sikkim's majority language, while Bhutia, Lepcha, Limbu—all indigenous languages of the state—amounted to a little under 30% when clustered with other minoritized languages. Other major languages represented in the state are Hindi, Bengali and Bhojpuri, languages that have travelled with migrants from neighboring Indian states. Since the Sikkim Official Language Act of 1977, Nepali, Bhutia and Lepcha have been recognized as co-official languages of the state alongside English. Gurung, Limbu, Magar, Mukhia, Newar, Rai, Sherpa and Tamang are now recognized as additional official languages (Tourism and Civil Aviation Department, Government of Sikkim 2020). For further background context on Sikkim, see Turin (2008). For summaries of the language context of Sikkim as revealed in surveys and censuses prior to the 2005-2006 Linguistic Survey of Sikkim, see Turin (2011).

Structured as a short form census, the linguistic survey that we discuss in this contribution was administered in secondary schools throughout the state of Sikkim with the support of school leadership and the local government. The first phase was spearheaded by linguistic anthropologist Mark Turin together with local colleagues under the auspices and leadership of the Namgyal Institute of Tibetology and in close partnership with the Department of Human Resource Development (formerly Education) of the Government of Sikkim. A first analysis of the resulting data was published in Turin (2011), in which the author highlighted trends of ongoing language shift among the autochthonous linguistic communities of the state as revealed in the findings of the survey, in addition to contextualizing the observed patterns within Sikkim's recent political history and the sociolinguistics of the wider Himalayan region (see Sonntag & Turin 2019). The

2011 paper offered a thorough discussion and disentangling of the terminology surrounding language use in Sikkim (and South Asia as a whole) which might be overwhelming to readers new to linguistic research in the area and lead to misinterpretation, e.g., the distinction between what language(s) someone may identify as their mother tongue vis-à-vis what they actually speak. While the earlier published work sought to tease out some sociolinguistic trends based on the data and aimed to make wider ethnographic points, the rich dataset has remained under-analyzed until now.

The careful statistical analysis of the dataset that we have undertaken in this article has helped to surface many additional tendencies and correlations, and we discuss our process and our findings below. In particular, we shed light on the distribution of patterns of language shift across different regions of Sikkim.[1] The rest of this article is organized as follows. In §2, we introduce the survey dataset and provide key background information on how the original survey was conducted. This is followed by a transparent exposition of the curatorial decisions we have made and how some of the data had to be reframed for the analysis. §3 is the central content of our paper: §3.1 describes the main statistical tool used, viz., Goodman–Kruskal $\tau$, §3.2 details the results of the numerical analysis and comments on patterns of language shift observed, while §3.3 computes frequencies for number of languages lost at each generational transition across the entire surveyed population. In §4, we inspect the data at a more granular level by considering schools in each district separately. In this section, we uncover patterns that speak to both the variation in language dynamics across districts (explained by unique geographical and social factors), and also variations within districts. Finally, in §5, we conclude with some structured reflections on the implications of the preceding analyses together with some preliminary thoughts on future areas of research that might be productive to explore.

## 2 Dataset characteristics and data curation

The original linguistic survey field team travelled to each of Sikkim's four districts to visit schools and administrative offices to seek a better understanding of the complex linguistic reality of the state. From October 2005 to November 2006, surveyors visited 105 government and private schools and asked the higher classes (VIII-XII) to complete a survey of 29 questions on language use. In total, 16,527 survey forms were completed by 8,662 female (52%) and 7,803 male (47%) students in a period between 15 November 2005 and 21 November 2006. The resulting dataset totals 479,283 completed fields (number of questions multiplied by the number of returns). These survey forms were then numbered, photographed and entered into a FileMakerPro database. The entire database was anonymized and returned to the Government of Sikkim and the Namgyal Institute of Tibetology once data verification and rechecking was complete. Included in the survey were questions on which language(s) the respondent speaks with his or her parents, grandparents and siblings; which language(s) a respondent's kin speak with one another; how many languages the respondent could speak and write, and which ones; questions on the different domains and registers of language use (songs, lists, letters, numbers, TV), and which language the respondent identifies as his or her mother tongue.

[1] For the purposes of this paper, we look only at the language data collected in the original survey and not at individual ethnic self-identification on which the contentious issue of defining a Sikkimese resident and the legal consequences thereof rests. For an exploration of the history and legalities of this topic, see Vandenhelsken (2020).

The survey team were able to visit most secondary and senior secondary schools in Sikkim's four districts. As of October 2002, Sikkim was home to 1,949 schools, 1,478 of which were government establishments while the remaining 471 were private institutions. This number has now changed, as some schools have closed, others have been amalgamated and new schools have opened. The distinction between public and private is significant because private schools are not required to abide by the government curriculum that includes a provision for vernacular instruction in the mother tongue. The educational pyramid at the time of the survey can be broken down into 978 Pre-Primary, 297 Lower Primary, 390 Primary, 153 Junior High Schools, 90 Secondary Schools and 41 Senior Secondary Schools. It should also be noted that Gangtok's schools are a particularly diverse range of institutions, from government schools providing free education to exclusive elite schools that are recognized across India. In addition, and in common with other rapidly urbanizing areas across India, Gangtok's schools are also home to many of Sikkim's ethnolinguistic groups as families migrate to the city for education, healthcare and employment.

A snapshot of a few rows and columns of the dataset after some editing (see below for a discussion of our editorial decisions) is provided in Appendix A below. The data set is arranged in a spreadsheet of 16,503 rows corresponding to individual students in classes VIII through XII across several schools in Sikkim. Data collected include personal biographical variables such as name (later anonymized), gender and age (at the time of survey), what school they were attending, alongside several sociolinguistic variables documenting what languages they and their families use in various contexts. These latter variables are the object of the present study with the goal of uncovering patterns of language use and change in usage (if any). Note that some of the sociolinguistic variables may have as many as four enumerated languages corresponding to each individual. For example, in the first data-row of Appendix A, we see that both Nepali and Bhutia are listed as languages that parents (of the student) speak together.

As is often the case with broad census datasets, a degree of curation of the raw data is required to make the set amenable to analysis using statistical software. We used R in the RStudio environment (R Core Team 2021; Posit Team 2025) for our data-cleaning and analysis. The first challenge was to ensure uniformity in transcription of language names. This included rectifying obvious spelling mistakes or minor deviations from standard spellings entered by survey respondents, e.g., "Npeali" or "Napali" for the language officially spelled "Nepali", or "Limboo" for "Limbu". The source of the error could be either at the level of the respondents themselves, or might have been introduced during data-entry (through typing errors) owing to the size and scope of the task involved.

In addition, many languages are known—officially or unofficially—by more than one spelling (or even more than one name, as discussed below) in the published literature. In some cases, there is no single standard that is equally recognized by all, as in "Newar", "Newa" and "Newari", or "Bangla" and "Bengali". Thankfully, it is fairly easy to correct for such discrepancies, even if it is a little time-consuming when undertaken manually, which was necessary given the sensitivity, complexity and scope of the dataset.

A second and more serious challenge relates to those instances where a language may not have a single standard name, or at least one that is well-known or well-attested among speaker communities. The language referred to as "Bhutia" is a case in point. Respondents to the survey used three distinct names for this language (each with spelling variations): "Sikkimese", "Bhutia" and "Denjongke". "Sikkimese" is the general and common English name for this language, "Bhutia" the Indo-Aryan exonym (Turin 2002) and "Denjongke" one of the autoglossonyms used

(Yliniemi 2021), that is, the term used by the native and heritage speech community (or some portion thereof) to refer to their language.[2] Other examples are "Limbu" which often appears as "Subba" and "Lepcha" which alternates as "Rongaring". Standardizing these names requires domain-specific knowledge of the linguistic make-up of the region, information which was provided by one of the authors.

For the strict purpose of our statistical analyses, the actual choice of name—while sociolinguistically salient and culturally important—was inconsequential as long as it was unique (to whichever varieties we identify as forming a single language) and remained consistent throughout. At the same time, wanting to honor the terminological choices made by the students who completed the survey, we chose names that were currently accepted or recommended in the academic literature and, when it was possible to verify, used or at least not opposed by speaker communities as well (see, for example, Bhutia (2024) for "Bhutia" and van Driem (1993) for "Newar").

Third, we encountered several macro-language names, such as "Rajasthani," "Bihari" and "Kirat" (and "Rai"), which do not index a single language, but rather point to clusters or groupings of languages, some of which are reported separately in the dataset. Thus, the superset "Rajasthani" occurs in some fields and "Marwadi" (using various spellings) in others; similarly "Kirat" in some cases and "Bantawa" in others, and so on. For simplicity, all instances of "Marwadi" were replaced by "Rajasthani" (itself distinct from "Hindi" and "Urdu") and "Bihari" was replaced by "Bhojpuri". However, "Bantawa" and "Sunwar" were kept distinct from the larger reported mass of "Rai" and "Kirat" (which were standardized to "Kirat"). This is because we were inherently interested in their dynamics, as these languages are indigenous to the Himalayan region, which cannot be said for immigrant languages such as Rajasthani and Bihari. This decision has, no doubt, obscured a level of granularity from our analysis at the language-specific level, but we assert that it serves to make our analysis more accurate and accessible. That is to say, while our calibrated analysis now references the macro-language unit "Rajasthani" as a whole rather than specific languages such as Marwadi, Harauti and others, we find this to be preferable to reporting partial figures for individual languages. There is no way for us to ascertain which particular language was being indicated in a response that offered simply "Rajasthani" without further context or detail.

---

[2] We are very grateful to an anonymous peer reviewer who reminded us that Yliniemi (2021) notes the term "Lhoke" as the more common autoglossonym among the speaker-base of the Bhutia language. However, in our dataset, among the hundreds of references to and responses for this language, there is not even a single occurrence of "Lhoke" (or a spelling variant thereof). A cursory review Google Ngram for the use of "Lhoke" and "Denjongke" reveals that throughout the 1900s and until the early 2010s, "Lhoke" was indeed more common than "Denjongke", at least in the corpus examined for this statistic. Unfortunately, we could not search in Ngram for the Tibetan script items ཇོང་སྐད *lho-skad* and འབྲས་ལྗོངས་སྐད *ḥbras-ljoṅs-skad*, which most speakers would likely have used to refer to the language. The increase in popularity of the term "Denjongke" from the mid-2010s may be correlated to Dr. Yliniemi's publications on the language, and the subsequent popularization of the term in academic discourse. The absence of "Lhoke" in our dataset from 2005-2006 remains a little surprising and we can only speculate on the reasons for this. One possibility is that in the self-reported responses by the students surveyed, "Lhoke" was considered to be a community-internal term and not one that would be easily understood or welcome in an English language survey, just as a Japanese speaker would likely not use *nìhóńgó* or a German speaker *Deutsch* when referring to their language in (an) English (context). Indeed, even "Denjongke" occurs rarely, "Sikkimese" and "Bhutia" being by far the preferred terms of choice for the language in our dataset. Likewise, "Manipuri" is preferred in responses over "Meithei", this latter term only used by one student parenthesized as "Manipuri (Maitai)".

The procedure we undertook to implement the above curatorial decision and others are briefly outlined below. (For a complete list of language varieties reported in the survey, see Appendix B.) At the outset, all major name-uniformizing away from non-standard names (but not specific spelling variations) was undertaken using a simple document-wide replacement command. This included changing all instances of "Sikkimese" to "Bhutia", "Bangla" to "Bengali", "Shrestha" to "Newar", "Subba" to "Limbu", etc., as well as the few aforementioned cases of clustering entries together into macro-languages.

Next, it was necessary to separate out instances of multiple entries within a single cell to multiple cells in order to apply a spelling standardizer and to compute association measures. When a respondent offered multiple languages in response to a question, we reformatted the string so that each cell in our spreadsheet would contain only one language. For example, in response to what language(s) they speak with their grandparents, some respondents included up to four different languages. These we reformatted as four distinct cells, with each cell containing only a single language entry. When computing association measures between two different variables, if any one of the languages in one variable agrees with that in another variable, we counted the result as a positive match. This procedure is explained in more detail below in the section where we report the results of our analysis.

After this step, an automatic spelling corrector was used to replace aberrant spellings with the closest chosen standard within our list, as long as the Levenshtein distance (van der Loo 2014) between the current form and one of the standard forms was small enough. By necessity, we had to make this threshold fairly small (< 3), because the Levenshtein distance counts the number of edits to match two strings from left to right. Thus, the string "Bhojpuri" will be closer by the Levenshtein metric to "Bhutanese" than to "Voajpuri" even though the latter is a spelling variant, and "Bhutanese", although apparently similar at a surface level, is a completely different and unrelated language. Any remaining discrepancies in the data had to be fixed manually. For example, all instances of "Voajpuri" and other *v*-starting spelling variants of "Bhojpuri" were replaced by "Bihari".

## 3 Results

One of our research goals was to assess whether the language-data reported by the students remained consistent across generations—that is, if there is evidence of any language shift. Since languages spoken by three consecutive generations (grandparents, parents, student-respondents) were recorded in the survey, along with languages spoken in inter-generational conversation (which language(s) parents speak with grandparents, which language(s) the respondents speak with parents, which language(s) the respondents speak with grandparents), this task was achievable using association measures that compared each of the variables. More specifically, we computed the Goodman–Kruskal $\tau$ measure of association in R with the function GKtauDataframe(·) from the GoodmanKruskal package (Pearson 2020) for pairs of variables in the dataset.

### *3.1 The Goodman–Kruskal τ association statistic*

At this point, we offer a heuristically-informed discussion of the main statistical association measure we use in this paper, the Goodman–Kruskal $\tau$, and its interpretation in the setting of this

analysis. The Goodman–Kruskal $\tau$ measures how closely paired observations on two variables $X$ and $Y$ with distinct classes or categories $X_1$, $X_2$, …, $X_m$ and $Y_1$, $Y_2$, …, $Y_\nu$, respectively, agree with one another by checking if the categorization of the dependent variable $Y$ can be predicted from knowledge of an observation's belonging to a category of the independent variable $X$. In our case, for example, $X$ can be languages reported as being spoken by the grandparents with one another and $Y$ the languages the students themselves speak with their siblings. The categories are then individual languages as represented in the dataset. We thus define the $\tau$-association by how knowing what a student's grandparents speak allows us to infer what they themselves speak. If the variables are highly associated ($\tau$-value close to 1), it means the prediction is fairly accurate. If they are not ($\tau$-value close to zero), it means that $Y$ cannot be well-predicted from X, and thus, there have been some unpredictable changes to the languages spoken.[3] To fix notation, consider the following contingency table:

| | | ***X*-categories** | | | | **Totals** |
|---|---|---|---|---|---|---|
| | | $X_1$ | $X_2$ | ⋯ | $X_m$ | |
| ***Y*-categories** | $Y_1$ | $n_{11}$ | $n_{12}$ | ⋯ | $n_{1m}$ | $n_{1+}$ |
| | $Y_2$ | $n_{21}$ | $n_{22}$ | ⋯ | $n_{2m}$ | $n_{2+}$ |
| | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋮ |
| | $Y_\nu$ | $n_{\nu 1}$ | $n_{\nu 1}$ | ⋯ | $n_{\nu m}$ | $n_{\nu+}$ |
| **Totals** | | $n_{+1}$ | $n_{+1}$ | ⋯ | $n_{+m}$ | $n$ |

Table 1. Outline of a contingency table

Here, we have a sampled population of $n$ individuals of whom $n_{ji}$ belong to categories $X_i$ and $Y_j$. For instance, in our illustration above, this will mean that $n_{ji}$ students in the dataset speak language $Y_j$ (e.g., Nepali) with their siblings while their grandparents speak language $X_i$ (e.g., Bhutia) between themselves. There are $m$ categories of $X$ and $\nu$ of $Y$. The items $n_{j+}$ denote the sum of frequencies in row $j$, i.e., the total number of individuals in category $Y_j$. Likewise, the items $n_{+i}$ denote the sum of frequencies in column i, i.e., the total number of individuals in category $X_i$.

The Goodman–Kruskal $\tau$ of $Y$ on $X$ essentially computes the ratio of the difference between the probability of misclassification into $Y$ and the probability of misclassification into $Y$ with knowledge of $X$ to the probability of misclassification into $Y$, where the task of classifying an observation into a $Y$-category is done so that the $Y$-column totals on the right-margin of the table are maintained. For a precise mathematical description, we refer to the original paper of Goodman and Kruskal (1954) or the slightly less rigorous treatment by Reynolds (1977). For our analysis, we don't require an exact formula as the tedious computation of all the $\tau$-s reported is undertaken in R. It can be shown that the measure is necessarily bounded between 0 and 1.

The association measure $\tau$ is not symmetrical. In other words, if $X$ were considered the dependent variable and $Y$ the independent, one does not expect to see the same value for $\tau$ of $X$ on $Y$ as $\tau$ of $Y$ on $X$. This, of course, is perfectly in line with the real-world interpretation of our dataset. For the variables we analyze, in service of understanding language shift in the region, there is a clear chronological gradient to causality, as in, it is more meaningful to try to predict a current

[3] By "unpredictable" we mean only from the knowledge of the data as contained in the contingency table. Of course, one may have external sources that inform us how the implied change has occurred, e.g., a lingua franca like Nepali or Hindi gaining ground at the expense of Indigenous languages.

generation's language habits from that of an older generation in order to gauge the sociolinguistic setting rather than the other way around. That said, mathematically, there is nothing preventing us from running the measure in the opposite direction and that can be a helpful diagnostic for spotting changes in certain extreme cases.

Consider, for example, a situation where there is a shift from every distinct *X* category to a single *Y* category. In that case, despite changes, the $\tau$ of *Y* on *X* is 1 because we have perfect prediction (there is nothing to predict as there is just a single Y category). However, the $\tau$ of *X* on *Y* will be low here as there is no way to predict the different *X* categories which have been conflated as a single *Y* category. In this case, the inability to clearly back-predict provides a way of identifying major changes in datasets as having a clear directionality.

If *Y* is exactly predictable from *X*, then $\tau = 1$; while if nothing can be said of *Y* from knowledge of *X*, then $\tau = 0$. Between these two extremes, higher values of $\tau$ indicate greater association, i.e., greater predictive capability, and lower values indicate lesser. In the context of our dataset, this has a very natural interpretation. Consider the two variables, grandparents speak together and speak with grandparents, which we place under the labels *X* and *Y* we have been using for illustration. If *X* values of "Nepali" match with most instances of "Nepali" in *Y*, instances of "Bhutia" in *X* match instances of "Bhutia" in Y, and so on, then we will get a high value of $\tau$. One way of interpreting this is to say that the two variables under consideration are strongly associated. Another, and equivalent, interpretation is that knowing what grandparents speak together allows us to predict fairly accurately what our respondents might speak with grandparents.

There is yet another way through which a high value of $\tau$ might be attained. Since the predictive power of *X* on *Y* underlies what is measured, it could be the case that if most instances of "Nepali" in *X* correspond to "Bhutia" in *Y*, and most instances of "Bhutia" in *X* correspond to "Nepali" in *Y* (the values have been switched so to speak), we will still see a strong association even though this is not what we are after. However, for sociological reasons, we do not expect such "language interchange" across generations, and this scenario can safely be set aside as an unlikely mathematical curiosity. In general, language shift often requires a prestige hierarchy, often imposed, making it unidirectional (see Verhaeghe, van Avermaet, & Derluyn (2022), and Turkistani & Almoaily (2025) on explicit case-studies of language-shift to dominant varieties).

Finally, since for each variable, some students reported multiple languages (which is very natural as families are often multilingual, both within and across generations), there was a risk of underestimating cross-variable associations since computer programs will treat "Language 1, Language 2, Language 3" as different from "Language 2, Language 1, Language 3" or "Language 1, Language 2". We resolve this issue by counting a positive match if at least one of the languages reported matches across the columns.

### ***3.2*** *Results*

The following table shows $\tau$ values computed for pairs of variables which can reveal interesting insights about continuity of language use versus language shift across the generations. In particular, we singled out seven variables to compute $\tau$ across pairs among them which we imagine will reveal the most helpful chronological information on language attrition and shift. These are (languages which) grandparents speak together (GST), grandparents speak with parents (GSP), (respondents) speak with grandparents (SG), parents speak together (PST), (respondents)

speak with parents (SP), (respondents) speak with siblings (SS) and (respondents' self-identified) mother tongue (MT). The last variable, mother tongue, is of independent interest as it is well-known that there is no unified consensus across cultures and societies on a definition of what constitutes a "mother tongue" (Turin 2011).

| | **GST** | **GSP** | **SG** | **PST** | **SP** | **SS** | **MT** |
|---|---|---|---|---|---|---|---|
| **GST** | | 0.85 | 0.82 | | | | 0.64 |
| **GSP** | 0.86 | | | 0.64 | | | |
| **SG** | 0.81 | | | | 0.65 | | |
| **PST** | | 0.73 | | | 0.82 | | 0.66 |
| **SP** | | | 0.76 | 0.82 | | 0.65 | |
| **SS** | | | | | 0.59 | | |
| **MT** | 0.57 | | | 0.67 | | | |

Table 2. Goodman–Kruskal $\tau$ associations between selected pairs of variables

Several interesting aspects are revealed in the numbers in the above table. First, it is clear that while the Goodman–Kruskal $\tau$ is not a symmetric measure, in most cases, we do not see too much of a difference between the explicating power of $x$ on $y$ vis-à-vis $y$ on $x$. For instance, $\tau$ of GSP on GST is 0.85, while $\tau$ of GST on GSP is 0.86. However, there are some exceptions to this trend. The value of $\tau$ for what parents speak together (PST) on what grandparents speak to parents (GSP) is 0.64, noticeably lower than the $\tau$ of GSP on PST at 0.73. That is, the predictive power in one direction is somewhat greater than the opposite. Something similar is shown for the pair of variables SP and SG.

Having established these baselines, we now discuss the broad picture painted by these numbers. Overall, there is high association between GST and GSP, meaning that in most cases, the grandparent generation had passed down at least one of their languages to the parent generation. However, the association between GSP and PST is somewhat lower, meaning that even though the parent generation know the languages of the grandparent generation and use them to communicate across the two generations, in many cases, they use some other language(s) to communicate within their generation. This is a significant finding, and the pattern continues to the next generation.

Interestingly, even though there seems to be a shift between the grandparent and parent generations, this is not clearly reflected in the languages the respondents speak to their grandparents. Between SG and GST, the association remains very high. There could be two reasons for this. First, it is possible that the languages of the grandparent generation have been passed down to the respondent generation, but the scope of usage of these languages is restricted to only speaking with the older generation. The alternative explanation is that because in many cases, more than one language is reported by an individual respondent for each variable, and we compute association measures based on at least one match, it is possible that different languages are being counted across generations. As an illustration, suppose that an entry for the grandparent generation lists Bhutia and Nepali as languages spoken in GST, the latter being somewhat of a lingua franca, and only Nepali is found in SG, this still counts as a match for our computation even though a language has been lost across the generations. Unfortunately, this sort of change, though very significant in the study of language shift, is not captured very well by statistical measures looking at

bulk data unless they are weighted differently, which may introduce inadvertent biases on the part of the researchers. In our paper we therefore avoid any weighted analysis as we wish to present what the data represents without further intervention. We return to this point later and offer some partial remedies in the next section by tracking the number of matches across generations. By doing so, it will be clear if more than one language is often passed down.

The association is not particularly strong between what respondents speak with parents (SP) and what they speak to grandparents (SG), which implies that there is indeed evidence that language shift has occurred in some families. Even though grandparent generation languages are known by the respondents and used when speaking to grandparents, they are not in use when speaking to parents, and even less so when speaking with age mates (siblings or friends). Perhaps significantly, the association measures between SP and SS are even lower. That is, between siblings, not even the languages spoken with parents are used, but rather the languages that are taught or used as the medium of instruction in school. This matter deserves some attention because it signifies not one, but two, different instances of language shift (or at least strong preference) across the three generations covered in the study.

Interestingly, the language(s) reported as mother tongue (MT) is/are not strongly associated with either what parents speak together, or what grandparents speak. This result may be somewhat surprising, and can perhaps be attributed to different interpretations of the phrase "mother tongue" in different students' minds. The phenomenon where individuals report a mother tongue they do not actively speak is well-documented in the sociolinguistic and applied linguistics literature (Borossa 1998). This situation often arises due to intergenerational language shift, heritage language attrition, and sociocultural factors influencing language use within immigrant and minority communities (Bonfiglio 2010).

### ***3.3*** *Generational change in multiple language preservation*

While the association measures above clearly indicate trends in language shift and language persistence, it is interesting to know whether more than one language was reported in a family in any of the three generations in the dataset, and then how many of those were likely to be passed down. We chose three of the variables that offer the most compelling evidence for language use within a generation, viz., what grandparents speak together (GST), what parents speak together (PST), and what the respondents speak with their siblings (SS). To gauge how many of the reported languages in GST are in common with PST, we computed the frequency of matches for each respondent. The results are provided in the table below.

| **Number of matches** | **Frequency** |
|---|---|
| 1 | 13,840 |
| 0 | 1,855 |
| 2 | 777 |
| 3 | 30 |
| 4 | 1 |

Table 3. Matches in languages reported between GST and PST

It is no surprise that a single match has the highest frequency. This means that at least one language spoken in the grandparents' generation was passed down to the parents' generation. Of course, if there was only a single language reported in GST, then that would be the only language that could be passed down; but if more than one language was spoken, then this count includes only one of those languages being passed down, but not the others. Two languages being passed down is much rarer—there are only 777 counts—and even fewer for three and four languages. The case of zero languages being passed is of greatest interest for us because it implies that a complete language shift has occurred between generations. The frequency for this occurrence is 1,855 out of a total of 16,503, which is 11.24% of the total. In other words, 11.24% of the parent generation did not retain (or at least use, as reported in the survey) the language(s) of their parents. This is concerning because it indicates that individuals in this 11.24% might not be able to communicate well with their parents, potentially severing intra-familial communication. As is known from instances of inter-generational loss of language elsewhere, this can have profound emotional and psychological consequences for the people concerned. For Tlingit, for instance, Dauenhauer and Dauenhauer (1998) find that one of the most tragic aspects of this language shift is the breakdown in intergenerational communication. Some children cannot talk to their own grandparents except in English, a language that is often foreign to the elders.

Between the respondents' generation and their parents, we find very similar numbers, as shown below.

| Number of matches | Frequency |
|---|---|
| 1 | 13,018 |
| 0 | 2,139 |
| 2 | 1,268 |
| 3 | 76 |
| 4 | 2 |

Table 4. Matches in languages reported between PST and SS

As before, counts of one are the highest, indicating that for the vast majority, at least one language was passed down within a family. However, here too we see that in 2,139 cases out of 16,503—12.97%—there was some degree of complete language replacement. In these households, the respondents' generation use a different language from their parents.

## 4 A geographically segmented analysis

In this section, we explore whether any discernable differences exist in patterns of language shift across the various subregions of Sikkim as captured in the dataset. We take schools—which were recorded for each student—as a helpful proxy for region, as we know the geographical location of each school and students generally attend a school close to their home. Our goal here is to identify whether there are local strongholds of language persistence even when the state-wide pattern has been shown to be a steady cross-generational shift, and if so, where these are located.

This part of our study was designed as follows. Names of schools were recorded in the original dataset together with the number of students surveyed in each school and the region in which the school was located: east, west, north and south Sikkim. We chose a few schools in each

region and computed association statistics for each to compare with the Sikkim-wide statistics. The choice of schools was partly random, but also informed by other factors, for example, excluding schools with fewer surveyed students. This is because more observations from specific sites make statistical measures more reliable (namely, they have greater statistical power), while conclusions drawn from fewer observations are inherently more unstable. In most cases, each selected school had over a hundred students surveyed, and in many instances, a few hundred. As already noted, the rationale was for the statistical measures to be robust, something that can only be guaranteed with larger sample sizes. The one exception to this general rule were schools in North Sikkim. The highly mountainous district of North Sikkim is sparsely populated, home to fewer schools than other parts of the state, and the number of students in the relevant classes numbered fewer than 100 students in each case. Consequently, our sample sizes across schools in North Sikkim are quite small. Nevertheless, we certainly did not want to exclude data from North Sikkim because it was precisely in this more remote northern district that we expected language persistence to be the most pronounced, and this was where our hypothesis could best be tested.

As helpfully noted by an anonymous peer reviewer of this manuscript, it may be the case that our preferential sampling of larger schools could have biased the results somewhat, notwithstanding the anticipated gain in statistical reliability. In particular, larger schools tend to be located in urban or more populated areas, thereby providing a picture that may be specific to those areas and not generally representative of other regions. We consider some of this risk to be offset by considering the four districts separately, and by identifying specific schools that behave conspicuously and differently (e.g., Pelling Government Secondary School, see below). Quite intentionally, we did not pool data within each of the four regions so that sub-regional differences (understood through the proxy of schools) were not smoothed out. After all, immigration dynamics, ethnic composition and other factors influencing language distribution are locally contingent.

The schools selected from the four geographical districts at the time were as follows:[4]

- **North Sikkim**: Mangan Senior Secondary School, Phodong Senior Secondary School, Lingdong Secondary School, Passingdong Secondary School
- **East Sikkim**: Enchey Senior Secondary School, Samdong Senior Secondary School, Tadong Senior Secondary School, Sir Tashi Namgyal Senior Secondary School, Rangpo Secondary School, Tarpin Secondary School
- **South Sikkim**: Namthang Senior Secondary School, Temi Senior Secondary School, Melli Bazaar Secondary School, Namchi New Secondary School
- **West Sikkim**: Pelling Government Senior Secondary School, Soreng Senior Secondary School, Daramdin Secondary School, Kaluk Secondary School, Tharpu Secondary School

We clarify here that the above four regional categories are administrative divisions, and inferences on them, reported below, should be interpreted in terms of their internal geography (remoteness, terrain, proximity to the rest of India) and social dynamics (population density, degree of urbanization, desirability to migrants). Despite the names of the districts, they do not simply

[4] The reorganization of Sikkim's districts was officially announced on December 21, 2021, when the Sikkim Legislative Assembly passed the Sikkim (Re-Organization of Districts) Bill, 2021. This legislation led to the creation of two new districts—Pakyong and Soreng—and the renaming of the four existing districts to better reflect their administrative centers. The new districts Pakyong and Soreng were carved out from parts of East and West Sikkim, respectively.

reflect cardinal directions, although certainly are correlated with them. The four districts are not entirely cardinally arranged, e.g., all of East, South and West Sikkim combined occupy less territory than North Sikkim. Thus, a statement along the lines of "East Sikkim schools show greater evidence of language-shift than North Sikkim" should not be read as referring to strictly cardinal eastern and northern Sikkim, but specifically to the erstwhile districts of East and North Sikkim.

In 2021, the borders of all districts were slightly redrawn, and two new districts were created. While the renamed district of Mangan occupies roughly the same area as the former North Sikkim, it is in the south that much of the major restructuring happened. This may have been administratively convenient because North Sikkim/Mangan is considerably more mountainous and remote, less densely inhabited and farther removed from the rest of India, directly bordering Tibet to its north. This relative remoteness is also reflected in our data: despite its larger geographic area, there were fewer schools in North Sikkim, and each school had fewer students than schools in other regions. Because the survey was carried out prior to this reorganization with data points labeled in accordance with the districts of the time (Turin 2011), we do not apply the newer administrative changes onto our present analysis.

Some notable features arise from our analysis of the data in each of these schools. For reasons of space, not all the figures are reported here, but we provide regional summaries. In reading the results that follow, we note that the state capital Gangtok, the largest urban center of the state, is in East Sikkim. In addition, East, South and West Sikkim directly border the state of West Bengal, and in the case of East Sikkim, also Nepal. These are also sites of emigration and immigration, locations of settlement as well as the movement of military and administrative service people and their families, to and from other parts of India. In contrast, North Sikkim has been relatively shielded from such developments and is home to more remote Indigenous communities residing in more traditional mountainous villages, and where children mostly attend schools within the district. These factors no doubt play a role in the patterns observed in the data outlined below.

Among the chosen schools in West Sikkim, the computed Goodman–Kruskal $\tau$ association measure shows that the language(s) spoken remained roughly the same for the parent- and grandparent-generations, but that there was an ongoing shift in the respondents' generation. Knowledge of ancestral languages, however, has not been lost as these languages are still being used to communicate with parents and grandparents even though the preferred language for speaking to siblings has changed. All instances of the association between what students speak to their siblings and what they speak to their parents are fairly low compared with what they speak to their parents and grandparents.

Of note is Pelling Government Senior Secondary School where language persistence is high, even into the respondents' generation, likely on account of its relatively remote location within the district. Tharpu Secondary School also offers an exception in that fewer respondents appear to be using their ancestral language even when speaking with their parents and grandparents. Here, the association between what parents and grandparents speak together is fairly strong (as for other schools in the region), but the association measure between what the respondents speak to their parents and grandparents is much lower and what they speak amongst themselves is weaker still.

In South Sikkim schools, the language shift trend that was observed in West Sikkim seems to be a generation ahead. The association between languages the parent-generation speak together and those spoken by the grandparent-generation is noticeably low. Interesting to note is that the

association values at the transition between the parent- and respondent-generations are also low—in fact, more so than at the prior generational transition—possibly capturing the same pattern of two levels of language-shift as evidenced in our Sikkim-wide analysis and as outlined in §3.

The East Sikkim statistics mirror the South Sikkim situation very closely. There is a clear and sharp decline in association measures at the grandparent-to-parent-generation transition and another decline at the parent-to-respondent-generation transition. What is different from South Sikkim, however, is that of the six schools analyzed, there is some heterogeneity at the first generational transition. That is to say, in some schools there is a marked low association between the languages that parents speak together and those that grandparents speak together, while for others, this is not so. We suspect this to be a consequence of the district being the main urban center where Sikkim's state capital, Gangtok, is located and where language-shift patterns may be more advanced than in rural locations. The question of the interaction between language use and urban density certainly requires further investigation.

In the four North Sikkim schools examined here, there appears to be very minimal language shift. The $\tau$ association measure for languages being passed down remains high at both generational transitions. For Phodong Senior Secondary School, in particular, there are fairly high $\tau$ values between languages spoken with parents and those spoken with siblings. In other schools in this district, however, we see a noticeable decline in this measure, but still not as high as the decline in schools in other regions.

All in all, our original suspicion that more remote North Sikkim schools will show greater language-endurance than other areas seems to hold true. Furthermore, it also follows that East Sikkim, home to the main urban center of Gangtok, and South Sikkim which directly borders the rest of India, manifest the effects of language shift(s) at a more accelerated rate than North Sikkim. West Sikkim is, statistically speaking, situated in between these two opposing patterns. Here, our decision to not pool the dataset across schools in a given region is validated. Pelling, which is located in the northern reaches of West Sikkim, has clear evidence of slower language-shift, patterning like North Sikkim schools, while Soreng, close to the border with West Bengal, patterns more like South Sikkim with greater shift. If we had amalgamated Pelling, Soreng and other schools in West Sikkim, the granularity of this analysis would have been lost.

## 5 Conclusion

In sum, our reanalysis of the 2005-2006 Linguistic Survey of Sikkim dataset demonstrates a patterned, directional shift in language use across three generations, even as knowledge of ancestral languages persists in specific and constrained domains. After a statistically-informed and rigorous review of the entire dataset and following the harmonization of language labels, Goodman–Kruskal $\tau$ statistics across seven key variables reveal strong continuity from "grandparents speak together" to "grandparents speak with parents", but a marked attenuation at the transition to "parents speak together", with an additional decline from "respondents speak with parents" to "respondents speak with siblings".

High associations between what respondents speak to grandparents and what grandparents speak together likely reflect domain-specific maintenance (rather than comprehensive and uninterrupted intergenerational transmission), while a weak alignment between self-declared mother tongue and actual usage underscores the complex interaction between identity and practice.

Frequency counts of intergenerational matches sharpen this perception further: approximately 11.24% of parent-generation dyads show complete replacement of grandparental languages, and 12.97% of respondent-generation dyads show complete replacement of parental languages, pointing to potentially serious communicative discontinuity and the interruption of intergenerational language transmission within families.

A geographically segmented view adds additional nuance. North Sikkim exhibits the highest levels of continuity—including relatively strong alignment between languages used with parents and with siblings—while East and South Sikkim, shaped by urban density and border-proximate mobility, display more advanced shifts. West Sikkim sits between these poles, with intra-district heterogeneity (e.g., greater persistence in Pelling versus accelerated shift nearer Soreng/Tharpu). Although constrained by the school-based sampling method that was used in the original survey, choices relating to name standardization, and conservative matching rules that favor any language-use overlap across generations over differential weighting accounting for indigeneity, these findings nonetheless provide a robust baseline for policy and community action.

Our findings lead us to assert that the targeted reinforcement of vernacular instruction (especially in private and urban schools) and the deliberate expansion of domains for heritage languages to thrive beyond informal peer networks and classrooms would support language continuity. In addition, we recommend that an enumeration along the lines of original survey be conducted every decade to track changes and document socio-linguistic trajectories within the state over time. In the process, we hope to learn more about the mechanisms that sustain and erode multilingual repertoires in Sikkim's diverse sociolinguistic ecologies.

## REFERENCES

Bhutia, Karma S. 2024. "Understanding and clarifying the term 'Bhutia'". *International Journal of Humanities and Social Science Invention* 13.10: 112-114.

Bonfiglio, Thomas P. 2010. *Mother Tongues and Nations: The Invention of the Native Speaker*. Berlin: DeGruyter. https://doi.org/10.1515/9781934078266

Borossa, Julia. 1998. "Identity, loss, and the mother tongue". *Paragraph* 21.3: 391-402. https://doi.org/10.3366/para.1998.21.3.391

Dauenhauer, Nora Marks; and Dauenhauer, Richard. 1998. "Technical, emotional, and ideological issues in reversing language shift: Examples from Southeast Alaska". In: Grenoble, Lenore A.; and Whaley, Lindsay J. (eds.), *Endangered Languages: Language Loss and Community Response*, 57-98. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139166959.004

van Driem, George. 1993. "The Newar verb in Tibeto-Burman perspective". *Acta Linguistica Hafniensa* 26.1: 23-43. https://doi.org/10.1080/03740463.1993.10415451

Goodman, Leo A.; and Kruskal, William H. 1954. "Measures of association for cross-classifications". *Journal of the American Statistical Association* 49.268: 732-764. https://doi.org/10.2307/2281536

van der Loo, Mark P. J. 2014. "The stringdist package for approximate string matching". *The R Journal* 6.1: 111-122. https://doi.org/10.32614/rj-2014-011

Pearson, Ron. 2020. "GoodmanKruskal: Association analysis for categorical variables". *R package* version 0.0.3., at https://CRAN.R-project.org/package=GoodmanKruskal Accessed Aug 27, 2025.

Posit team. 2025. "RStudio: Integrated development environment for R". *Posit Software, PBC*, Boston, MA, at http://www.posit.co Accessed Aug 27, 2025.

R Core Team. 2021. "R: A language and environment for statistical computing". *R Foundation for Statistical Computing*, Vienna, Austria, at https://www.R-project.org Accessed Aug 27, 2025.

Reynolds, Henry T. 1984. *Analysis of Nominal Data*. Sage Publications. https://doi.org/10.4135/9781412983303

Sonntag, Selma K.; and Turin, Mark (eds.). 2019. *The Politics of Language Contact in the Himalaya*. Cambridge, UK: Open Book Publishers. https://doi.org/10.11647/OBP.0169

Tourism and Civil Aviation Department, Government of Sikkim. 2020. https://sikkimtourism.gov.in/Public/ExperienceSikkim/history Accessed Nov 16, 2025

Turin, Mark. 2002. "Ethnonyms and other-nyms: Linguistic anthropology among the Thangmi of Nepal". *Proceedings of the Ninth Seminar of the IATS, 2000. Volume 9: Territory and Identity in Tibet and the Himalayas*: 253-269. Brill. https://doi.org/10.1163/9789004483101_019

Turin, Mark. 2008. "Sikkim". In: Stearns, Peter N. (ed.), *Oxford Encyclopedia of the Modern World: 1750 to the Present*. Oxford: Oxford University Press. ISBN13: 9780195176322, ISBN10: 0195176324. https://doi.org/10.1093/acref/9780195176322.001.0001

Turin, Mark. 2011. "Results from the linguistic survey of Sikkim: Mother tongues in education". In: Balikci-Denjongpa, Anna; and McKay, Alex (eds.), *Buddhist Himalaya: Studies in Religion, History and Culture*, II, Namgyal Institute of Tibetology: 127-142, *The Sikkim papers*, 978-8192226118.

Turkistani, Sumaiya; and Almoaily, Mohammad. 2023. "Language shift or maintenance? An intergenerational study of the Tibetan community in Saudi Arabia". *International Journal of Language and Literary Studies* 5.3: 301-314. https://doi.org/10.36892/ijlls.v5i3.1407

Vandenhelsken, Mélanie. 2020. "The 1961 Sikkim subject regulation and 'indirect rule' in Sikkim: ancestrality, land property and unequal citizenship". *Asian Ethnicity* 22.2: 254-271. https://doi.org/10.1080/14631369.2020.1801338

Verhaeghe, Floor; van Avermaet, Piet; and Derluyn, Ilse. 2019. "Meanings attached to intergenerational language shift processes in the context of migrant families". *Journal of Ethnic and Migration Studies* 48.1: 308-326. https://doi.org/10.1080/1369183X.2019.1685377

Yliniemi, Juha S. 2021. "A descriptive grammar of Denjongke (Sikkimese Bhutia)". *Himalayan Linguistics*. http://dx.doi.org/10.5070/H920146466

Samopriya Basu
basusamapriya@gmail.com

Mark Turin
mark.turin@ubc.ca

## Appendix A: Snapshot of the dataset

A few rows and columns of the Linguistic Survey of Sikkim dataset (after curation; see §2) are replicated below, showing, in particular, two of the variables used in the present analysis ("mother tongue" and "(languages which) parents speak together"). The original data-file is a Microsoft Excel spreadsheet. This was converted to a CSV spreadsheet for easier import to R for the analysis.

| **Languages spoken** | **Languages written** | **Mother tongue** | **Number of siblings** | **Parents speak together** | | |
|---|---|---|---|---|---|---|
| 3 | 2 | Bhutia | 0 | Nepali | Bhutia | |
| 4 | 3 | Tibetan | 0 | Tibetan | English | |
| 4 | 3 | Bhutia | 0 | Bhutia | Nepali | English |
| 3 | 3 | Tibetan | 0 | Tibetan | | |
| 4 | 3 | Bhutia | 0 | Bhutia | | |
| 1 | 1 | Lepcha | 0 | Lepcha | | |

## Appendix B: List of languages attested in dataset

In this appendix, we list all the languages that appeared in the dataset with name-variants noted. Spelling differences (including mistakes) of a single language-name are not given here as there are far too many to list. Macro-language names like "Arunachali", "Bihari", "Rajasthani" and "Rai" are also listed here separately for convenience. A few names were reported only once (or by one individual) and were impossible for us to identify, even taking spelling mistakes into account (e.g., "Kashai", "Palsi", "Rangchoo" and "Ghindaghi"). It could be these are very local dialectal or village names, perhaps even clan or tribe names, or entirely new language varieties not yet documented (unlikely, but not impossible). Because of their low frequency, these single enumerations don't affect the statistical analysis in any way, and have therefore also not been included below.

| **Language** | **Name variant** | **Additional variant** |
|---|---|---|
| Adi | | |
| Arunachali | | |
| Assamese | | |
| Awadhi | | |
| Bantawa | Kerawa | |
| Bhojpuri | | |
| Bhutia | Sikkimese | Denjongke |
| Bihari | | |
| Bodo | | |
| Braj | | |
| Chamling | | |
| Chinese | | |
| Deori | | |
| Dogri | | |
| Dukpa | | |
| English | | |
| Garhwali | | |
| Garo | | |
| Goalparia (marked by respondent as "dialect" of Bengali) | | |
| Gujarati | | |
| Gurung | | |
| Haryanvi | | |
| Himachali | | |
| Hindi | | |
| Kannada | | |
| Karbi | | |
| Khamti | Tai-Khamti | |
| Khasi | | |

| | | |
|---|---|---|
| Kirat | Rai | |
| Kulung | | |
| Kumaoni | | |
| Lepcha | Rongaring | Lachungpa |
| Limbu | Subba | Yakthungba |
| Magahi | | |
| Magar | | |
| Maithili | Tirhut | |
| Malayalam | | |
| Manipuri | Meithei | |
| Marathi | | |
| Marwari | | |
| Mewari | | |
| Monpa | | |
| Naga | | |
| Nagamese | | |
| Nepali | Gorkhali | Chetri |
| Newar | Pradhan | |
| Nishi | | |
| Oriya | | |
| Padam | | |
| Punjabi | | |
| Rajasthani | | |
| Sanskrit | | |
| Santali | | |
| Sherpa | | |
| Sindhi | | |
| Sumi | | |
| Sunwar | Mukhia | |
| Tamang | Moktan | |
| Tamil | | |
| Tangkhul | | |
| Telugu | | |
| Thami | | |
| Tibetan | | |
| Sharchop | | |
| Urdu | | |
| Uttaranchali | | |
| Yolmo | | |