Linguistic Marginalization as a Weapon of Economic Oppression: An NLP Analysis of Financial Literacy Outcomes in India

Ihita Ghosh

Supervisor: Dr. Meredith Landman

Department of Slavic Languages, Barnard College of Columbia University

Laidlaw Scholars Research and Leadership Program

August 2025

Abstract

India is currently facing a "linguistic genocide" of minority languages due to the implementation of nationwide hegemonic policies. This has severe repercussions for speakers, as it contributes to cycles of intergenerational poverty. Experts suggest that many of the declining languages are spoken by Scheduled Tribes, a factor that has played a role in their widespread impoverishment. To build on these claims, this study employed AI-based corpus analysis and statistical modeling to investigate the relationship between the availability of financial literacy terms and financial literacy outcomes across different language groups. Additional factors, including language vitality, script type, geographic region, and population density, were also assessed. The results revealed a significant correlation between the presence of financial literacy terms and financial literacy outcomes, with language vitality and script type serving as key explanatory variables. Languages with the lowest percentages of financial literacy terms and financial literacy scores typically belonged to historically marginalized tribal groups. These findings underscore the role of linguistic exclusion in perpetuating poverty and highlight the urgent need for financial literacy education in local languages.

Keywords: Language Accessibility, Financial Literacy, Scheduled Tribes, Marginalization, Poverty, India, Mother-Tongue Education, Script Type, Natural Language Processing, Regression Analysis

1 Introduction

India's rich cultural heritage can be largely attributed to its linguistic diversity, a facet that is viewed as a reflection of the nation's "composite nature" (Montaut, 2005). Among the hundreds of different languages and dialects, 22 are recognized by the Eighth Schedule of the Indian Constitution. This lists Assamese, Bangla, Boro, Dogri, Gujarati, Hindi, Kashmiri, Kannada, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Tamil, Telugu, Sanskrit, Santali, Sindhi, and Urdu as the major languages of the country (Government of India, Ministry of Education, n.d.). However, throughout history, many of these languages, among others, have been marginalized in an attempt to gain control of the population. A key example occurred during the British Colonial period, when local languages were suppressed and English was imposed as the mandatory medium of instruction (Rao, 2008). Supporters of these policies argued that knowledge of English would enable Indians to access British resources in the sciences and liberal arts, thereby allowing individuals with English language proficiency to become more educated in society. These advocates were able to gain greater access to economic opportunities and eventually formed an elite "buffer" class (Rahman et al., 2018). Such financial hierarchies have persisted in modern times, with language serving as a significant contributing factor.

Individuals who did not belong to the English-educated buffer class harbored strong feelings of resentment after experiencing decades of marginalization. This cultivated a landscape for the rise of subnational political entities that called for the "linguistic reorganization" of states (Rao, 2008). Among the proposed options, Hindi and Urdu were advocated the most due to their wide geographic spread and deep cultural ties to North India. However, the Southern states, each with its own distinct culture, vehemently opposed both languages, arguing that neither adequately represented the nation as a whole. Ultimately, no national language was established, though heated debates over the issue persist to this day.

The national language controversy even played a role during the partition of India and Pakistan, which was mainly caused by a religious schism between Hindus and Muslims. The disagreement between Hindi and Urdu as the national language was central to this division. Most Hindus advocated for Hindi, while Muslims advocated for Urdu, which created a strong association between language and religion (Farooqi, 2017). Later on, strife gave way to the establishment of an independent India for Hindus and an independent Pakistan for Muslims, often referred to as the Partition of 1947 (Talbot & Singh, 2009). The partition deepened the association of Urdu with Muslims

and fueled distrust of the language among Indian Hindus (Farooqi, 2017). These negative attitudes toward Urdu have intensified in recent times with the rise of religious nationalism. Subsequently, debates in India have shifted from selecting a language to focusing on whether Hindi should be established as the country's national language.

The Indian government has strongly advocated for establishing Hindi as the national language through initiatives such as adding numerals in Devanagari (the writing system for Hindi) to the national currency, changing highway signs from English to Hindi, and pushing to have Hindi as a required language in schools until the 10th grade (Ranjan, 2021). However, these efforts have been met with resistance, as communities seek to preserve their distinct languages and cultures. In this context, many individuals have started to view Hindi as a political weapon being used to push a homogeneous North Indian Hindu identity across the entire country. Although communities have attempted to resist government efforts, linguistic diversity has been on a sharp decline in recent years, a phenomenon that experts are terming "linguistic genocide" (Ghose et al., 2024). This especially pertains to less dominant languages that are typically spoken by Scheduled Tribes (STs). The marginalization of these minority languages is a significant concern as it largely contributes to identity erasure and reduces the diversity of India's culturally rich heritage as a whole.

Along with identity erasure, linguistic marginalization may also trigger poverty cycles for speakers of oppressed language groups. This occurs because linguistic marginalization excludes languages from the public domain, preventing the establishment of standardized orthographic conventions or writing systems, and thereby limiting the availability of educational materials for future generations (Farooqi, 2017). Without adequate access to written instructional resources, individuals become trapped in intergenerational poverty, with limited opportunities for social mobility (Sujatha, 2002). To break this cycle, members of STs attempt to learn a second language that has greater economic and social access. However, even with these integration efforts, India's rigid hierarchical system makes it difficult for them to continue their educational journeys. High levels of discrimination contribute to many members of tribal communities dropping out of school and resigning themselves to their position in society as *fait accompli* (Mohanty, 2008). The lack of adequate accommodations for STs results in identity crises, deprivation of freedom, and educational failure, all of which contribute to persistent cycles of poverty. Consequently, many tribal communities continue to struggle with financial insecurity in the present day.

2 Purpose of the present study

The purpose of this study is to investigate how the lack of language accessibility contributes to intergenerational poverty by applying natural language processing and statistical analysis to linguistic data. The analysis examines the relationship between two variables: the availability of financial literacy terms in local language materials and the financial literacy outcomes of individuals across different speaker groups. Additional factors, such as linguistic marginalization, lack of standardized writing systems, and rigid hierarchical standards, are also explored to gain a comprehensive understanding. Regional and distributional patterns were also assessed, but did not yield significant results. This study makes a critical contribution to computational linguistics and economics by building on existing analyses and providing quantitative evidence to the discussion of how the lack of language accessibility contributes to intergenerational poverty in India. The study draws on Indian linguistic corpus data, Perplexity AI outputs, census statistics, state-wise financial literacy data, and scholarly literature linking language marginalization to poverty. Finally, the study examines the implications of the current system and advocates for reform.

3 Methodology

3.1 Sample

The sample for this study is the 21 modern languages recognized by the Eighth Schedule of the Indian Constitution (Sanskrit was not considered due to its limited usage). The datasets for the percentage of financial literacy terms per language group were gathered from online public corpora. Table 1 presents information on the corpora and language profiles.

Table 1: Data Overview

Language	**Corpus Source**	**Corpus Word Count**	**Writing System(s)**
Assamese	Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics, n.d.	10,000	Assamese Script
Bengali	Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics, n.d.	10,000	Bengali Script
Boro	Department of Linguistics, K.M. Institute of Hindi and Linguistics, 2024	100,121	Roman Script, Devanagari Script, Bengali/Assamese Script
Dogri	Open-Speech-EkStep, n.d.	492,148	Devanagari Script, Nasta'liq Script
Gujarati	Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics, n.d.	10,000	Gujarati Script

Language	**Corpus Source**	**Corpus Word Count**	**Writing System(s)**
Hindi	Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics, n.d.	10,000	Devanagari Script
Kannada	Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics, n.d.	10,000	Kannada Script
Kashmiri	Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics, n.d.	2,530	Nasta'liq Script, Devanagari Script
Konkani	Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics, n.d.	10,000	Devanagari Script, Kannada Script, Malayalam Script, Nasta'liq Script, Roman Script
Maithili	Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics, n.d.	10,000	Devanagari Script

Language	**Corpus Source**	**Corpus Word Count**	**Writing System(s)**
*Malayalam	Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics, n.d.	10,000	Malayalam Script
Manipuri (Meitei)	Statistical Machine Translation (StatMT), 2020	451,329	Bengali Script, Meitei Script
Marathi	Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics, n.d.	10,000	Devanagari Script
Nepali	Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics, n.d.	10,000	Devanagari Script
Oriya	Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics, n.d.	10,000	Oriya Script

Language	**Corpus Source**	**Corpus Word Count**	**Writing System(s)**
*Punjabi	Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics, n.d.	10,000	Gurmukhi Script
Tamil	Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics, n.d.	10,000	Tamil Script
Telugu	Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics, n.d.	10,000	Telugu Script
Santali	Google, n.d.	6,600	Devanagari Script, Bengali Script, Oriya Script, Roman Script, Ol Chiki Script
Sindhi	Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics, n.d.	10,000	Naskh Script, Devanagari Script

Language	Corpus Source	Corpus Word Count	Writing System(s)
Urdu	Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics, n.d.	10,000	Nasta'liq Script

**Note 1.* Although these languages are formally written in additional scripts outside of India, this study focuses primarily on the Indian context, and therefore, those alternative scripts were not included.

Note 2. Script information was obtained from Omniglot (n.d.).

The percentage of financially literate individuals per language group was calculated using publicly available data from the 2011 Indian Census, accessed through Langlex (n.d.), combined with state-level financial literacy scores from the 2019 Financial Literacy and Inclusion Survey (National Council for Financial Education, 2023). Since no publicly available data exist on financial literacy outcomes per language group, this estimation represents the most viable approach for conducting the analysis.

3.2 Design

The analysis employed simple and multiple linear regression models to examine the relationship between the percentage of financial literacy terms (independent variable) and the percentage of financially literate speakers (dependent variable) across language groups. Interaction term categories included language status, script type, geographic region, and population distribution. A literature review was then conducted to contextualize the regression findings.

3.3 Measures

Two measures were used to calculate the percentage of financial literacy terms and the percentage of financially literate individuals per language group. For the former, Perplexity AI (2023) was employed to identify terms

related to financial literacy in each dataset and compute their proportion relative to the total word count. For the latter, a weighted average proportion of financially literate individuals was calculated for each sampled language group. A simple linear regression was first conducted on these two variables to test for correlation. Subsequently, multiple linear regression models incorporating language vitality, script type, geographic region, and population distribution were further used to examine the relationship between the availability of financial literacy materials in local languages and levels of financial literacy in India.

3.4 Procedure

Financial Literacy Terms

The data used to calculate the percentage of financial literacy terms were obtained through a systematic search of corpora for each sampled language. Perplexity AI (2023) was then prompted with standardized instructions to generate the counts of financial literacy terms across corpora. This prompt was applied uniformly across languages to ensure comparability in the identification and quantification of financial literacy–related vocabulary. The standardized prompt is provided below:

> I need you to go through the attached file and look for words related to the following concepts ONLY in the corresponding language of the file (for example, if the file is in Hindi, look for Hindi financial literacy terms):

Table 2: Classification of Financial Literacy Concepts

Category	**Terms**
Time Value of Money	Present value, future value, discount rate, net present value, NPV, annuity, compounding, interest rate, time horizon, discounting
Interest	Interest, simple interest, compound interest, APR, annual percentage rate, fixed rate, variable rate, amortization, accrued interest

Category	Terms
Risk & Return	Risk, return, volatility, diversification, portfolio, yield, beta, alpha, capital gains, loss, reward, standard deviation, risk tolerance
Inflation	Inflation, CPI, Consumer Price Index, purchasing power, cost of living, price index, real income, nominal value, deflation, hyperinflation
Investing	Invest, investing, investment, stock, bond, mutual fund, ETF, index fund, dividends, capital gains, broker, asset allocation, diversification, portfolio, equity, securities, risk tolerance, IPO, return on investment, ROI, financial markets
Credit	Credit, borrower, lender, credit score, credit report, credit history, loan, lending, line of credit, APR, mortgage, student loan, payday loan, interest, credit card, balance transfer, debt-to-income ratio, cosigner
Saving	Saving, savings account, emergency fund, interest, compound interest, piggy bank, bank account, certificate of deposit, CD, money market account, automatic transfer, high-yield savings
Budgeting	Budget, budgeting, expense, income, fixed expenses, variable expenses, discretionary spending, financial goals, cash flow, savings goal, envelope system, zero-based budget
Debt Management	Debt, repayment, minimum payment, debt snowball, debt avalanche, credit counseling, bankruptcy, loan consolidation, delinquent, collections, interest rate, principal, default, forgiveness

Category	Terms
Financial Planning	Financial plan, retirement, pension, 401(k), IRA, Roth IRA, social security, annuity, insurance, life insurance, health insurance, disability insurance, premium, deductible, beneficiary, estate planning, will, trust, inheritance, financial advisor, planner, goal setting

This is not a fixed list.

Do a broad search and include a list of all terms relevant to financial literacy in the corresponding language.

Keep a count of the words related to these concepts and the total number of words available in the file.

When counting the total number of words, keep track of all the words in the file and do not count the numbers or special characters.

Most of the files should have around 10,000 words (I'll tell you if it doesn't), but still keep a count of the total number of words just in case. Please also account for ALL the words in the file.

Calculate the fraction and the percentage of the words related to these financial literacy concepts in the file. Do not count the frequency of the words. Only consider the word counts.

Note. The prompt was derived from information provided by the Reserve Bank of India (2017) and created using ChatGPT (OpenAI, 2025).

After obtaining the counts of financial literacy terms in each corpus, the values were converted into percentages relative to the total corpus size using the following formula:

$$\text{Percentage of Financial Literacy Terms} = \frac{\text{Number of Financial Literacy Terms}}{\text{Total Number of Words in the Corpus}} \times 100$$

Financially Literate Individuals

The data used to calculate the percentage of financially literate individuals per language group were obtained through two systematic searches: one for the financial literacy percentages reported for each state and the other for the nationwide language distribution data. The former was obtained from official statistics released by the National Council for Financial Education

(2023), while the latter drew on data from the 2011 Census (Langlex, n.d.). These two measures were combined to compute a weighted average using the following formula:

$$\text{W.A.} = \sum_{i=1}^{34} \frac{\text{Financial Literacy \% in State } i \times \text{Language Speakers \% in State } i}{34} \times 100$$

Here, the denominator represents the number of states in the sample. The weighted percentages were summed across states and divided by this total to obtain the weighted average.

3.5 Statistical Analysis

The study employed regression analysis to assess the relationship between the availability of financial literacy terms and the proportion of financially literate individuals across language groups. The models included a simple linear regression between financial literacy terms and financially literate individuals, as well as interaction term regression models considering factors such as language vitality, script type, geographic region, and population distribution. Models were accepted if the R-squared coefficient exceeded 0.10 and the explanatory variables were statistically significant. These metrics are consistent with social science research standards that aim to assess the impact of specific predictors on the dependent variable (Ozili, 2023).

3.6 Results

The analysis showed a significant correlation between the availability of financial literacy terms in local language materials and financial literacy outcomes among speakers across language groups. Furthermore, the interaction terms regression analyses showed significance for language vitality and script type, but not for geographic region or population distribution. Figure 1 presents the results of the simple regression analysis. The dependent variable, Speakers, represents the percentage of financially literate individuals in each language group, while the independent variable, Terms, denotes the proportion of financial literacy terms in the corpus.

Source	SS	df	MS	Number of obs	=	21
				F(1, 19)	=	7.34
Model	.000832913	1	.000832913	Prob > F	=	0.0139
Residual	.002155223	19	.000113433	R-squared	=	0.2787
				Adj R-squared	=	0.2408
Total	.002988136	20	.000149407	Root MSE	=	.01065

Speakers	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
Terms	12.02751	4.43859	2.71	0.014	2.73743	21.31758
_cons	-.0007884	.0045863	-0.17	0.865	-.0103877	.008811

Figure 1: Simple linear regression analysis between the percentage of financial literacy terms and the percentage of financially literate individuals across language groups. Data analyzed using Stata 19 (StataCorp, 2025).

In Figure 1, the absolute value of the t-statistic for *Terms* exceeds the critical threshold of 1.96, with a corresponding p-value lower than 0.05, indicating that the independent variable is statistically significant. The R-squared value is also greater than 0.10, confirming that the model meets the criteria for acceptance. Moreover, the positive coefficient for *Terms* demonstrates a positive linear relationship between the independent and dependent variables. Taken together, these results indicate a significant correlation between the availability of financial literacy terms in local language materials and financial literacy outcomes among individuals in India.

The correlation between the independent and dependent variables is largely driven by language vitality and script type. Figure 2 displays the regression results incorporating language vitality as an interaction term. In this model, the *languagestatus* variable denotes the endangerment status of the sampled language.

Source	SS	df	MS	Number of obs	=	21
				F(5, 15)	=	5.62
Model	.00194796	5	.000389592	Prob > F	=	0.0041
Residual	.001040176	15	.000069345	R-squared	=	0.6519
				Adj R-squared	=	0.5359
Total	.002988136	20	.000149407	Root MSE	=	.00833

financiallyliterateindividuals	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
financialliteracyterms	53.70361	13.37741	4.01	0.001	25.19034	82.21688
languagestatus						
2	.0606343	.0203542	2.98	0.009	.0172504	.1040183
3	.0591602	.0202285	2.92	0.010	.0160442	.1022761
languagestatus#c.financialliteracyterms						
2	-50.19174	14.51827	-3.46	0.004	-81.13669	-19.24679
3	-57.16648	17.0988	-3.34	0.004	-93.61172	-20.72125
_cons	-.0549644	.0195122	-2.82	0.013	-.0965537	-.0133751

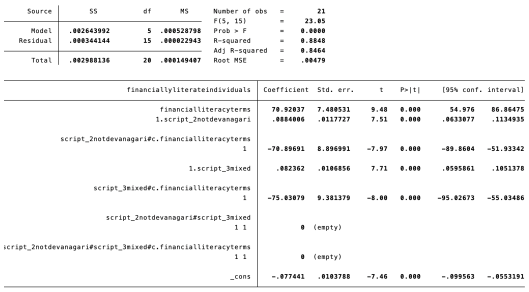
Figure 2: Interaction terms regression analysis of language vitality. Data analyzed using Stata 19 (StataCorp, 2025).

Before conducting the analysis, the sampled languages were grouped into three categories according to their endangerment status. Based on the data

from Ethnologue: Languages of the World (Eberhard et al., 2025), languages classified at level 1 on the Expanded Graded Intergenerational Disruption Scale (EGIDS) were coded as 1, those at level 2 as 2, and those at levels 3 through 5 as 3. This categorization was followed because languages beyond level 3 generally lacked sufficient data for meaningful comparison. In the regression, Groups 2 and 3 were compared against Group 1, which served as the control.

In Figure 2, the absolute values of the t-statistics for *financialliteracyterms*, *2*, *3*, and their interaction terms, all exceed the critical value of 1.96, with corresponding p-values less than 0.05, which indicates that the variables are statistically significant. The R-squared value is also higher than 0.10, showing that the model is acceptable. In the Stata output, the interaction terms are denoted by the symbol '#' (e.g., *languagestatus#financialliteracyterms*), which represents the combined effect of the two variables. The negative coefficient for category 2 of the interaction term *languagestatus#financialliteracyterms* shows that the effect of financial literacy terms on financial literacy outcomes is reduced for languages with an EGIDS level of 2 compared to that of languages with an EGIDS level of 1. The coefficient for category 3 is similarly negative and even smaller than that of category 2, suggesting that the effect of financial literacy terms on financial literacy outcomes is weakest for languages with an EGIDS level of 3–5.

Figure 3 presents the regression results incorporating script type as an interaction term. For this analysis, the sampled languages were classified into three categories: Devanagari, non-Devanagari, and mixed (Omniglot, n.d.). This grouping was designed to test whether the dominant writing system in India, Devanagari, affected the relationship between the availability of financial literacy terms and financial literacy outcomes. The variables were coded as *script_ 1devanagari* (not shown), *script_ 2notdevanagari*, and *script_ 3mixed*. Languages written exclusively in Devanagari were assigned to the first category, those written exclusively in a non-Devanagari script to the second, and those utilizing multiple formal writing systems to the third. *Script_ 1devanagari* was treated as the control variable, against which *script_ 2notdevanagari* and *script_ 3mixed* were compared.



Source	SS	df	MS	Number of obs	=	21
				F(5, 15)	=	23.05
Model	.002643992	5	.000528798	Prob > F	=	0.0000
Residual	.000344144	15	.000022943	R-squared	=	0.8848
				Adj R-squared	=	0.8464
Total	.002988136	20	.000149407	Root MSE	=	.00479

financiallyliterateindividuals	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
financialliteracyterms	70.92037	7.480531	9.48	0.000	54.976	86.86475
1.script_2notdevanagari	.0884086	.0117727	7.51	0.000	.0633077	.1134935
script_2notdevanagari#c.financialliteracyterms						
1	-70.89691	8.896991	-7.97	0.000	-89.8604	-51.93342
1.script_3mixed	.082362	.0106856	7.71	0.000	.0595861	.1051378
script_3mixed#c.financialliteracyterms						
1	-75.03079	9.381379	-8.00	0.000	-95.02673	-55.03486
script_2notdevanagari#script_3mixed						
1 1	0	(empty)				
script_2notdevanagari#script_3mixed#c.financialliteracyterms						
1 1	0	(empty)				
_cons	-.077441	.0103788	-7.46	0.000	-.099563	-.0553191

Figure 3: Interaction terms regression analysis of script type. Data analyzed using Stata 19 (StataCorp, 2025).

In Figure 3, the absolute values of the t-statistics for *financialliteracyterms*, *script_ 2notdevanagari*, *script_ 3mixed*, and their interaction terms all exceed the critical value of 1.96, with corresponding p-values less than 0.05, which indicates that the variables are statistically significant. The R-squared value is also higher than 0.10, showing that the model is acceptable. The interactions between *script_ 2notdevanagari#script_ 3mixed* and *script_ 2notdevanagari#script_ 3mixed#c.financialliteracyterms* yield an empty result due to perfect multicollinearity and were therefore excluded from the analysis. The negative coefficient for the interaction term *script_ 2notdevanagari#financialliteracyterms* indicates that, relative to languages written in Devanagari, the effect of financial literacy terms on financially literate individuals is reduced for languages not written in Devanagari scripts. The coefficient of the interaction term *script_ 3mixed#financialliteracyterms* is also negative and less than that of the interaction term *script_ 2notdevanagari#financialliteracyterms*. Therefore, the effect of financial literacy terms on financially literate outcomes is weaker for languages written in more than one formal script compared to those consistently written in a single script.

Factors such as geographic region and population distribution were hypothesized to influence the relationship between financial literacy terms and financial literacy outcomes; however, because the variable categories were not statistically significant, the models were rejected and excluded from further analysis.[1]

[1]The data were classified using resources from the Central Institute of Indian Languages (n.d.), National Institute of Urban Affairs (2022), and Maps of India (n.d.).

4 Discussion

The analysis shows a positive correlation between the availability of financial literacy terms and financial literacy outcomes, with language vitality and script type emerging as key explanatory factors. Languages with the lowest percentages of financial literacy terms were typically classified in category 3 for language vitality and in the mixed category for script type. These languages are typically spoken by tribal groups in India, such as Boro, Dogri, Konkani, Manipuri (Meitei), Santali, and Sindhi. The box plot in Figure 4 illustrates this disparity for language vitality.

In Figure 4, the median of category 3 is significantly lower than the medians of categories 1 and 2. This indicates that, on average, languages in category 3 contain fewer financial literacy terms than those in the higher categories.

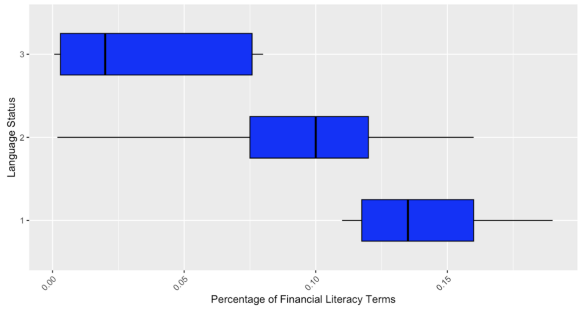


Figure 4: Comparison of financial literacy terms across language vitality categories. Data visualized using RStudio (R Core Team, 2025).

Figure 5 illustrates the disparity between the percentage of financial literacy terms across script categories. Similar to Figure 4, this figure shows that the median of languages in the mixed category is significantly lower than the medians of the Devanagari and non-Devanagari categories. This indicates that, on average, languages with no consistent formal writing system typically contain fewer financial literacy terms than those with strictly standardized ones.

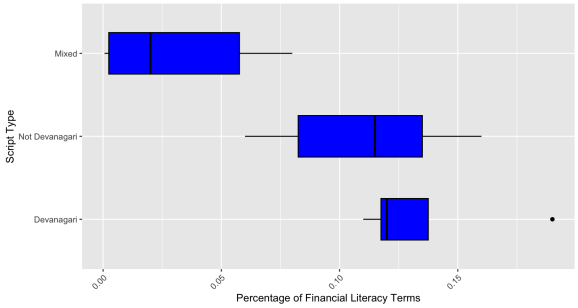


Figure 5: Comparison of financial literacy terms across script categories. Data visualized using RStudio (R Core Team, 2025).

The endangerment statuses and lack of standardized writing systems in tribal languages significantly constrain the presence of financial literacy terms in their written materials. Speakers of these languages were forced into these circumstances after being excluded from the public sphere by dominant linguistic groups, a process known as linguistic marginalization (Mohanty, 2010). This exclusion deprives them of economic opportunities and perpetuates cycles of intergenerational poverty. Furthermore, linguistic marginalization continues to hinder the development of these languages.

To understand how linguistic marginalization has contributed to the endangerment and stagnation of Indian tribal languages, it is necessary to consider their patterns of evolution. Languages, by nature, are inherently dynamic: word meanings evolve, dialects converge, and new forms emerge. However, if a language is not given the space to exist and progress, it eventually dies with the last generation of speakers (Mufwene, 2007). This occurs when it is denied access to the public sphere through the destruction of written records, exclusion from economic opportunities, and absence from formal education. These processes constitute linguistic marginalization and, in extreme cases, may culminate in linguistic genocide (Ghose et al., 2024).

Most of the sampled tribal languages in this study experienced historical linguistic marginalization. A specific instance is highlighted in the case of Konkani, the official language of India's western and coastal state, Goa (Omniglot, n.d.). During the Portuguese colonial period, Konkani was systematically devalued by the administration (Dubinsky & Starr, 2022). According to Sardessai (1983), Konkani started experiencing the effects of linguistic marginalization around the end of the seventeenth century when Viceroy Francisco de Tavora ordered that "the natives of the country should abandon the use of their language, and speak only Portuguese within 3 years."

By the middle of the 19th century, Portuguese had become the sole medium of instruction, the official language of the region and press, and the preferred language of the upper class. This pushed Konkani to the margins of society, where it began to decline.

The marginalization of the sampled tribal languages created significant barriers to the development and standardization of their writing systems. Many of these languages were once written in older scripts, while others existed primarily as oral traditions. However, prolonged efforts to erase them prevented the establishment of modern writing systems. Languages that were historically written in older scripts lost the continuity needed to sustain those traditions, while oral languages were marginalized before they had the opportunity to develop writing systems.

Revisiting the example of Konkani, which was traditionally written in the Goykanadi script, centuries of Portuguese rule led to the erasure of its historical writing traditions (Sardessai, 1983). During the colonial period, however, contact with other languages influenced its writing practices, eventually giving rise to multiple script forms. Today, Konkani is written in five different scripts: Devanagari, Kannada, Malayalam, Nasta'liq, and Roman (Omniglot, n.d.). The type of script employed is largely influenced by regional and religious factors (DeFrancis, 1984).

In contrast with Konkani, Boro is spoken by the Bodo people of Assam, a northeastern region of India (Omniglot, n.d.). Historically, it was predominantly an oral language. Following the colonial period, however, the Bodo people lost much of their political power and sovereignty, which pushed them to the fringes of society (Dwivedi, 2024). In response, the Bodoland movement emerged and ultimately led to the establishment of an autonomous region for the Bodo people. After gaining recognition, leaders sought to establish a standard writing system for Boro, but were initially unable to reach consensus (Garton et al., n.d.). In 1975, Devanagari was officially adopted as the standard script. Nevertheless, debates regarding the most appropriate script to represent the Bodo people continue to this day.

The challenges in adopting a script for Boro illustrate the complexities of establishing a standard writing system. In India, a script is likely to gain wide acceptance for a language only if it reflects and aligns with the sociocultural norms of its speakers (Garton et al., n.d.). Along with a standard writing system, orthographic conventions, which define the prescriptive norms of a script, must be determined. This prolonged process may contribute to a language's endangerment. Conversely, critical theories emphasize the potential dangers of imposing a writing system and orthography on a language community prematurely, as it may foster literacy elitism and rein-

force rigid hierarchical distinctions between "literate" and "illiterate" groups. Taken together, these considerations emphasize that timely, well-informed decision-making is essential to the process of language standardization.

Historical marginalization and delays in the standardization of tribal languages have contributed to a dearth of educational materials for these groups. As a result, many indigenous youths turn to neighboring dominant languages as a means of accessing greater opportunities and breaking cycles of intergenerational poverty. Since these dominant languages are often acquired after the mother tongue, features of the first language typically blend into the second. This may create hierarchies of "proper" versus "improper" use of the language and perpetuate standards that serve as a basis for discrimination.

An instance of a prescriptive language hierarchy can be evidenced by the Santal (an Anglicized version of the ethnic name *Sãotal*) tribe in West Bengal (Bagchi & Kumar, 2017). The Santals typically inhabit regions in Jharkhand, Bihar, western West Bengal, and the northeastern districts of Odisha and Assam. During the British colonial era, restrictions on forest resources and depictions of the tribe as "savage" and "barbaric" contributed to their societal marginalization and widespread poverty. To achieve a higher economic status, many Santals migrated to nearby industrial areas. This migration brought Santali into contact with Bengali, the official language of West Bengal, which many Santals adopted to ease communication in the workplace. Because of this contact, features of Santali influenced the Bengali spoken within the community. The influence is reflected in the choice of second-person pronouns. In Bengali, there are three kinds of second-person pronouns: */tui/*, */tumi/*, and */aapni/*. Each of these is used depending on the level of respect shown to the listener, with */tui/* being the most informal, and */aapni/* the most formal. However, the absence of such hierarchical distinctions in Santali leads Santals to use */tui/* across all contexts. Bengalis often perceive this as impolite, which establishes a linguistic standard that is used to discriminate against Santals.

The language contact between Santali and Bengali highlights a concept deeply embedded in complex societies: hierarchical structures. According to Turchin and Gavrilets (2009), hierarchies developed during the transition from "small-scale, 'simple' societies to large-scale, hierarchically complex ones." In India, these hierarchies have persisted through caste, culture, and gender. Efforts through modern progressive policies have sought to abolish such systemic hierarchies; nevertheless, the underlying structures remain intact. Today, these hierarchies are more commonly expressed through financial status rather than traditional social characteristics (Goodman & Kaplan, 2018). This system, often described as meritocracy, is more socially accepted

because it presumably ties success to effort. However, because the legacies of past hierarchies remain deeply embedded in social structures, those at the lowest rungs of society are often unable to achieve upward mobility, even with considerable effort. Language serves as a prime example of social weaponry that reinforces these barriers and prevents the upward mobility of the lower social classes.

Inequitable access to resources discourages many tribal members, leading them to resign their status to *fait accompli* and remain in their current positions (Mohanty, 2008). While some may succeed in accessing greater opportunities, this often comes at the expense of their linguistic identity. Recent efforts by the Indian government to establish a homogeneous national identity have contributed to the decline in linguistic diversity. This can be evidenced by data from the People's Linguistic Survey of India (PLSI), which reported that since the 1961 Census, more than 800 languages had already become extinct by 2016 (Ghose et al., 2024). Most of the languages that died out were indigenous languages, suggesting a clear pattern of erasure. Expert linguists are even viewing the situation as linguistic genocide, where the government is systematically eradicating tribal languages under the guise of promoting cultural unity. Tribal communities should not have to choose between their livelihoods and identities. Although acknowledging that not all languages can be used in professional settings, financial education in local languages is essential to promote equitable access to resources. Only then can India advance toward a genuine meritocratic society.

5 Conclusion

The correlational analysis of financial literacy terms and outcomes provides insight into how linguistic marginalization contributes to intergenerational poverty among India's Scheduled Tribe population. Language vitality and script standardization emerged as the most significant contributing factors, which evidence instances of historical linguistic marginalization. In the past, linguistic marginalization was carried out through explicit social hierarchies; however, these hierarchies are still socially ingrained in the minds of the Indian people, which creates barriers that prevent equitable access to opportunities. The newer system is more socially acceptable as it claims to tie success and effort together. Yet, in reality, this is simply a method of resuming age-old practices with a new demarcation. As a result, individuals belonging to linguistically vulnerable groups continue to struggle with poverty. For India to truly become a meritocratic nation, it is essential to

provide tribal communities with access to financial education in their local languages. This helps preserve linguistic diversity and offers opportunities for upward economic mobility.

6 Limitations

The limitations of this study include potential errors associated with AI, the absence of recent census data, and estimation inaccuracies. After testing multiple AI platforms, including ChatGPT, Gemini, and Google Cloud, Perplexity Pro produced the most comprehensive and accurate results. Numerous trials were also for verification. Regardless, an average uncertainty margin of $\pm 0.05\%$ remains in the estimated percentage of financial literacy terms per language group. With respect to the 2011 Census data used to calculate the percentage of financially literate individuals, its datedness may not fully reflect current linguistic demographics across states. This limitation was unavoidable due to the absence of more recent census data. Furthermore, the estimation of the percentage of financially literate individuals per language group may contain errors, as it does not fully account for the possibility that higher or lower levels of financial literacy may be concentrated within specific linguistic groups. However, averaging across groups helps reduce these effects.

7 Implications

The findings from this quantitative analysis provide critical evidence of the detrimental effects of linguistic marginalization as a mechanism of economic oppression in India. The study underscores the importance of delivering financial education to tribal communities in their local languages. Moreover, the ethical implications of linguistic erasure in the pursuit of homogeneity and economic growth warrant further examination. Future research should identify the most vulnerable tribal populations and evaluate policy interventions that ensure equitable access to opportunities. Overall, this study offers a comprehensive foundation for future economic, political, and linguistic inquiry.

References

Bagchi, T., & Kumar, R. (2017). Marginalisation, exclusion and identity of santals.

Central Institute of Indian Languages. (n.d.). Hindi demography [Accessed August 20, 2025]. http://lisindia.ciil.org/Hindi/Hindi_demo.html

DeFrancis, J. (1984). Digraphia. *Word*, *35*(1), 59–66.

Department of Linguistics, K.M. Institute of Hindi and Linguistics. (2024). Bodo [github repository]. https://github.com/kmi-linguistics/bodo

Dubinsky, S., & Starr, H. (2022). Weaponizing language: Linguistic vectors of ethnic oppression. *Global Studies Quarterly*, *2*(2), ksab051.

Dwivedi, A. (2024). The marginalization of the bodos: A struggle for ethnic identity [PDF retrieved from Journal website]. *International Journal of History*, *6*(2), 168–175. https://www.historyjournal.net/article/307/6-2-29-357.pdf

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2025). Ethnologue: Languages of the world (28th ed.) [Accessed August 20, 2025]. https://www.ethnologue.com/

Farooqi, I. (2017). Speak but don't write: Understanding linguistic exclusion in a metropolitan city.

Garton, R., Dale, M., Roy, L. S., & Basumatary, P. (n.d.). Endangered languages in the digital public sphere: A case study of the writing systems of boro and manipuri.

Ghose, A., Bharadwaj, S., & Ali, S. A. (2024). Linguistic genocide of minority and mother tongue languages: Unravelling international implications on indian laws through a critical discourse. *Journal of Asian and African Studies*, 00219096241243058.

Goodman, R., & Kaplan, S. (2018). The mantra of meritocracy [Accessed August 20, 2025]. *Stanford Social Innovation Review*. https://doi.org/10.48558/66KC-EG44

Google. (n.d.). Corpuscrawler [github repository] [Accessed August 20, 2025]. https://github.com/google/corpuscrawler?tab=readme-ov-file

Government of India, Ministry of Education. (n.d.). Indian languages [Accessed August 20, 2025].

Langlex. (n.d.). Home [Accessed August 20, 2025]. https://langlex.com/

Leipzig University, Saxon Academy of Sciences and Humanities, & Institute for Applied Informatics. (n.d.). Leipzig corpora collection (wortschatz leipzig / deutscher wortschatz) [Accessed August 20, 2025].

Maps of India. (n.d.). Zonal map of india [Accessed August 20, 2025]. https://www.mapsofindia.com/maps/india/zonal-map.html

Mohanty, A. K. (2008). Perpetuating inequality: Language disadvantage and capability deprivation of tribal mother tongue speakers in india. *Language and poverty*, 102–124.

Mohanty, A. K. (2010). Languages, inequality and marginalization: Implications of the double divide in indian multilingualism.

Montaut, A. (2005). Colonial language classification, post-colonial language movements and the grassroot multilingualism ethos in india. *Living Together Separately. Cultural India in History and Politics*, 75–116.

Mufwene, S. S. (2007). How languages die. *Combat pour les langues du monde-Fighting for the world's languages: hommage a Claude Hagege*, 377–388.

National Council for Financial Education. (2023). Executive summary: Ncfe financial literacy and inclusion survey 2019 [Accessed August 20, 2025]. https://ncfe.org.in/wp-content/uploads/2023/12/ExecSumm_.pdf

National Institute of Urban Affairs. (2022). Handbook of urban statistics 2022 [Accessed August 20, 2025]. https://niua.in/intranet/sites/default/files/2802.pdf

Omniglot. (n.d.). The online encyclopedia of writing systems and languages [Accessed August 19, 2025]. https://www.omniglot.com/

OpenAI. (2025). Chatgpt [large language model]. https://chat.openai.com/

Open-Speech-EkStep. (n.d.). Vakyansh-models [github repository] [Accessed August 20, 2025]. https://github.com/Open-Speech-EkStep/vakyansh-models?tab=readme-ov-file

Ozili, P. K. (2023). The acceptable r-square in empirical modelling for social science research. In *Social research methodology and publishing results: A guide to non-native english speakers* (pp. 134–143). IGI global.

Perplexity AI. (2023). Perplexity [large language model]. https://www.perplexity.ai

R Core Team. (2025). R: A language and environment for statistical computing [software]. https://www.r-project.org/

Rahman, A., Ali, M., & Kahn, S. (2018). The british art of colonialism in india: Subjugation and division. *Peace and Conflict Studies*, *25*(1), 5.

Ranjan, A. (2021). Language as an identity: Hindi–non-hindi debates in india. *Society and Culture in South Asia*, *7*(2), 314–337.

Rao, S. S. (2008). India's language debates and education of linguistic minorities. *Economic and Political Weekly*, 63–69.

Reserve Bank of India. (2017). Financial literacy in india – data & policy [pdf]. https://www.rbi.org.in/

Sardessai, M. L. (1983). The portuguese influence on konkani. *Journal of South Asian Literature*, *18*(1), 155–158.

StataCorp. (2025). Stata statistical software: Release 19 [software].

Statistical Machine Translation (StatMT). (2020). Pmindia v1 parallel corpus [data set] [Accessed August 20, 2025]. https://data.statmt.org/pmindia/v1/parallel/

Sujatha, K. (2002). Education among scheduled tribes.

Talbot, I., & Singh, G. (2009). *The partition of india*. Cambridge University Press Cambridge.

Turchin, P., & Gavrilets, S. (2009). Evolution of complex hierarchical societies. *Social Evolution & History*, *8*(2), 167–198.