International Journal of
Computational Linguistics Research

DLINE JOURNALS

Print ISSN: 0976-416X
Online ISSN: 0976-4178

IJCLR 2026: 17 (3)
https://doi.org/10.6025/ijclr/2026/17/3/133-150

Demographic Inequality and Linguistic Diversity in Indian States and Union Territories

K.Kiruthika
KSR College of Technology
Tiruchengode. Tamil Nadu
India
kkiruthika3108@gmail.com

ABSTRACT

India's multilingual landscape and uneven demographic distribution present unique challenges for governance, yet few studies integrate these dimensions at the sub-national level. This study proposes an integrated analytical framework to examine demographic inequality and linguistic diversity across India's 36 States and Union Territories. The acquired dataset encompasses administrative identifiers, socioeconomic metrics, and widely spoken languages. Demographic concentration is evaluated using the Gini coefficient, Lorenz curve, cumulative distribution, and Pareto analysis. Concurrently, linguistic heterogeneity is quantified through lexical diversity indices, including the Shannon and Simpson Diversity Indices, Type Token Ratio, Herdan's C, and Pielou's Evenness. Additionally, network-based visualizations, such as state language bipartite networks and co-occurrence maps, elucidate structural patterns of multilingual coexistence. Results reveal pronounced demographic inequality (Gini = 0.6149), confirming that a small subset of highly populated states accounts for the vast majority of the national population. In stark contrast, linguistic analysis identifies 43 distinct languages exhibiting high lexical richness and equitable distribution, demonstrating minimal dominance by any single language. These findings highlight a critical divergence between concentrated demographics and broadly distributed linguistic ecosystems. Ultimately, this reproducible computational framework provides policymakers with vital insights for evidence-based governance, equitable resource allocation, educational planning, and language preservation, demonstrating how computational analytics can support the study of complex socio-demographic systems.

Keywords: Demographic Inequality, Linguistic Diversity, Gini Coefficient, Lexical Diversity Indices, Multilingualism, Network Analysis, Computational Linguistics, Regional Development

Received: 20 January 2026, Revised: 1 May 2026, Accepted: 11 June 2026

1. Introduction

India is one of the most linguistically diverse countries in the world, characterized by hundreds of languages and dialects spoken across its States and Union Territories. This multilingual landscape reflects the country's rich cultural heritage while presenting significant challenges for governance, education, public administration, digital inclusion, and regional development. At the same time, India's demographic distribution is highly uneven, with a relatively small number of States accounting for a substantial proportion of the national population. These demographic disparities influence socioeconomic development, resource allocation, and the preservation and distribution of linguistic communities.

Understanding the relationship between demographic inequality and linguistic diversity is essential for developing effective public policies and promoting inclusive development. While previous studies have extensively examined global language endangerment, multilingualism, and demographic change, relatively few have investigated how population concentration and linguistic diversity interact at the sub-national level within India. Consequently, there remains a need for a comprehensive quantitative framework that integrates demographic indicators with measures of linguistic diversity to better characterize regional variations across Indian States and Union Territories.

To address this gap, the present study proposes an integrated analytical framework that combines measures of demographic inequality with computational linguistic diversity analysis. Population inequality is evaluated using the Gini coefficient, Lorenz curve, cumulative population distribution, population ranking, and Pareto analysis, while linguistic diversity is assessed using established lexical diversity indices, including the Type Token Ratio, Herdan's C, Guiraud's Index, Shannon Diversity Index, Simpson Diversity Index, Pielou's Evenness, and Dominance Index. In addition, network-based visualizations, including language co-occurrence networks, state language bipartite networks, word clouds, and diversity heatmaps, are employed to reveal patterns of multilingual coexistence and regional language distribution.

The findings contribute to a deeper understanding of India's demographic and linguistic heterogeneity and provide a reproducible analytical framework for studying multilingual societies. The proposed approach offers valuable insights for policymakers, researchers, and planners concerned with language preservation, regional development, educational planning, and evidence based governance while demonstrating how computational analytics can support the study of complex socio demographic systems.

2. Literature Review

2.1 Global Linguistic Diversity and Language Endangerment

Linguistic diversity constitutes one of humanity's most valuable cultural resources, reflecting centuries of social, historical, and cultural evolution. However, like the global decline in biodiversity, linguistic diversity faces an unprecedented threat. Of the approximately 7,000 documented languages worldwide, nearly half are currently classified as endangered, indicating an alarming decline in linguistic heritage [1-5] [(Sutherland, W.; Eberhard, D. M.; Moseley, C.; endangeredlanguages.com; Campbell, L.).

Although language change and language shift are natural components of cultural evolution, the rapid decline in language diversity has been significantly accelerated by colonisation, globalisation, urbanisation, and socioeconomic transformation. Historical factors, including colonial expansion, political environments, educational policies, and governmental support for bilingual education, substantially influence patterns of language endangerment across different regions [6] (Romaine, S.). Consequently, language vitality cannot be understood solely through linguistic characteristics but must also be examined within broader social, political, and economic contexts.

To better understand these patterns, Bowern, [7] analysed a comprehensive dataset comprising 6,511 languages, representing more than 90% of the world's spoken languages. Their study evaluated 51 predictor

variables associated with language maintenance and vitality, providing one of the most extensive quantitative investigations into factors influencing language survival. Similarly, Mufwene, [8] emphasized that under standing global threats to language diversity requires developing a macroecology of language endangerment, enabling researchers to investigate language loss through large scale ecological and sociolinguistic perspectives.

Although the mechanisms driving linguistic decline differ from those affecting biodiversity, both fields face similar analytical challenges in identifying global patterns of endangerment [9-13] (Grenoble, L. A.; Hua, X.; Cardillo, M.; Hua, X.; Amano, T.; Loh, J.). Extending this perspective, Bromham, [14] L. employed global analyses to investigate the complex interactions among environmental, demographic, and socioeconomic factors influencing language vitality and to predict the future trajectory of the world's languages over the coming century.

2.2 Socioeconomic and Environmental Drivers of Language Decline

Research has consistently demonstrated that language endangerment is influenced by multiple interconnected demographic and environmental factors rather than isolated linguistic characteristics. Languages are more likely to become endangered when they are located within regions where neighbouring languages are also under threat, suggesting that regional socioeconomic processes substantially influence language vitality.

Among these factors, transportation infrastructure has emerged as an important predictor. Increased road density facilitates human mobility, migration, commercial interaction, and integration into centralized governance, thereby accelerating language shift toward dominant regional or national languages. Interestingly, this relationship is not simply attributable to language contact, as languages with greater overlap with neighbouring linguistic communities often exhibit lower levels of endangerment. Similarly, other geographical variables such as landscape roughness, altitudinal variation, waterways, urbanization, GDP, and life expectancy do not consistently explain language decline.

Instead, road connectivity appears to represent increasing socioeconomic integration between previously isolated communities and urban centres. Limited transportation infrastructure has been identified as a protective factor for Indigenous language maintenance because it restricts the spread of dominant lingua francas, such as Tok Pisin in Papua New Guinea [15] (Aikhenvald). Conversely, improved transportation facilitates migration from rural communities to urban employment opportunities while simultaneously increasing external influence within traditionally isolated populations, thereby accelerating language shift toward dominant commercial and administrative languages [16-19] [(Muysken, P.; Gal, S.; Holmquist, J.; Dobrin, L.).

Furthermore, Skirgård's [20] analyses indicate that language loss is not geographically uniform. Instead, reductions in linguistic diversity are expected to vary considerably across the world's major linguistic regions, highlighting substantial regional disparities in language vulnerability.

2.3 Multilingualism, Linguistic Landscapes, and Language Visibility

Beyond language preservation, contemporary research increasingly examines how multilingualism is represented within public spaces and institutional environments. Linguistic landscapes (LLs), consisting of publicly displayed written languages, have become valuable resources for understanding language visibility, power relations, and ethnolinguistic vitality.

Brinkmann [21] investigated the pedagogical role of linguistic landscapes in multilingual regions by examining educational settings in Germany and the Netherlands. The study demonstrated that linguistic landscapes encourage critical reflection on linguistic power structures while supporting plurilingual pedagogies within mainstream secondary education (Brinkmann).

Similarly, Dafouz [22] examined how multilingualism and linguistic diversity are conceptualized across different institutional contexts. The findings revealed considerable variation in the visibility of linguistic diversity, with utilitarian and economic objectives generally receiving greater emphasis than identity-related

or cultural considerations (Dafouz). These studies collectively suggest that multilingualism extends beyond communication, encompassing issues of identity, educational policy, cultural representation, and social inclusion.

2.4 Linguistic Diversity During Public Health Emergencies

Recent global events have further highlighted the societal importance of linguistic inclusion. During the COVID-19 pandemic, multilingual communication became essential for ensuring equitable access to public health information. However, several studies documented significant linguistic inequalities in crisis communication.

The concepts of *disaster linguicism* [23] (Uekusa, 2019) and *linguademic* [24] (Phyak, 2020) describe how linguistic minorities experienced unequal access to emergency information during the pandemic. Chen [25] (2020) reported that Taiwan provided very limited public health communication in Indigenous languages, while migrant languages were similarly neglected. Comparable findings were observed in Qatar, where government communication was generally effective but largely excluded migrant languages such as Indonesian and Amharic [26] (Ahmad and Hillman, 2020). These studies illustrate that linguistic diversity remains a critical consideration for public policy, emergency communication, and social equity.

2.5 Artificial Intelligence and Linguistic Diversity

The rapid advancement of generative artificial intelligence has introduced new challenges for preserving linguistic diversity. Guo [27] investigated the consequences of recursively training language models using synthetic text generated by preceding models, a practice that is becoming increasingly common in large language model development. Rather than focusing exclusively on conventional performance metrics, the study evaluated lexical, syntactic, and semantic diversity using novel diversity metrics across multiple English natural language generation tasks.[28] The findings suggest that recursive synthetic data training may gradually reduce linguistic diversity, highlighting an emerging concern regarding the long term sustainability and representativeness of AI-generated language.

2.6 Research Gap

The existing literature provides substantial evidence regarding the global decline of linguistic diversity, determinants of language endangerment, multilingual education, linguistic landscapes, public health communication, and the emerging influence of artificial intelligence on language preservation. Nevertheless, relatively few studies integrate demographic inequality and linguistic diversity within a unified analytical framework, particularly at the sub national level in highly multilingual countries such as India. Existing research has primarily focused either on global language endangerment or on localized multilingual contexts, leaving a limited understanding of how demographic disparities intersect with linguistic diversity across Indian states and Union Territories.

Accordingly, the present study addresses this gap by systematically examining demographic inequality alongside linguistic diversity across Indian states and Union Territories, thereby providing a comprehensive regional perspective on the distribution, variation, and potential determinants of linguistic diversity.

3. Architecture

3.1 Research Design and Analytical Architecture

The proposed analytical framework integrates demographic inequality analysis with computational linguistic diversity analysis to examine the relationship between population distribution and multilingual diversity across the 36 Indian States and Union Territories. The framework consists of six sequential layers, beginning with data acquisition and ending with statistical interpretation and visualization.

The first layer involves acquiring demographic and linguistic information from official Indian government datasets and validated secondary sources. The collected data include administrative identifiers, population statistics, literacy rates, geographical areas, GSDP, population densities, and widely spoken languages for each State and Union Territory. Data cleaning and preprocessing are then performed to remove inconsistencies, standardize state names and ISO codes, normalize numerical variables, and tokenize multilingual language entries.

The second layer focuses on demographic inequality assessment. Population values are analyzed using inequality measures including the Gini coefficient and Lorenz curve, followed by ranking analysis, cumulative population distribution, and Pareto analysis to evaluate the concentration of population across administrative units.

The third layer extracts linguistic information from the "Widely Spoken Languages" attribute. Languages are tokenized into individual lexical units, normalized, and transformed into structured representations suitable for quantitative linguistic analysis.

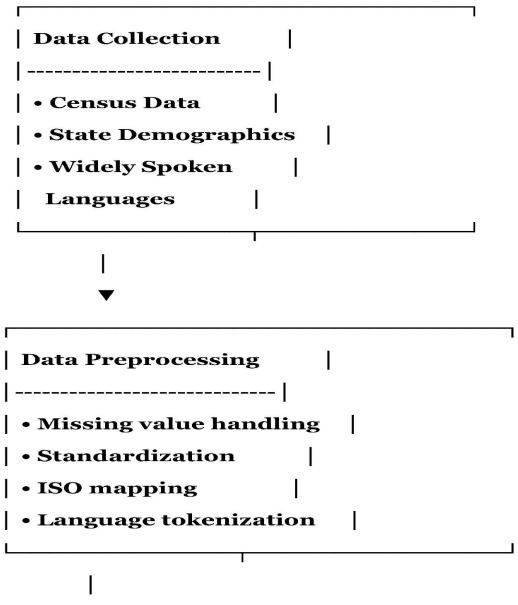
The fourth layer computes lexical diversity indicators, including:

- Type–Token Ratio (TTR)
- Herdan's C
- Guiraud's Index
- Shannon Diversity Index
- Simpson Diversity Index
- Pielou's Evenness
- Dominance Index

These metrics collectively quantify linguistic richness, diversity, uniformity, and dominance.

The fifth layer constructs network based representations of multilingual relationships, including language co-occurrence networks, state language bipartite networks, language frequency distributions, and language diversity heatmaps. These visualizations reveal patterns of multilingual coexistence across Indian regions.

Finally, all statistical outputs, diversity indices, and network measures are synthesized to explain demographic concentration and linguistic heterogeneity, enabling evidence based interpretation of regional language diversity.



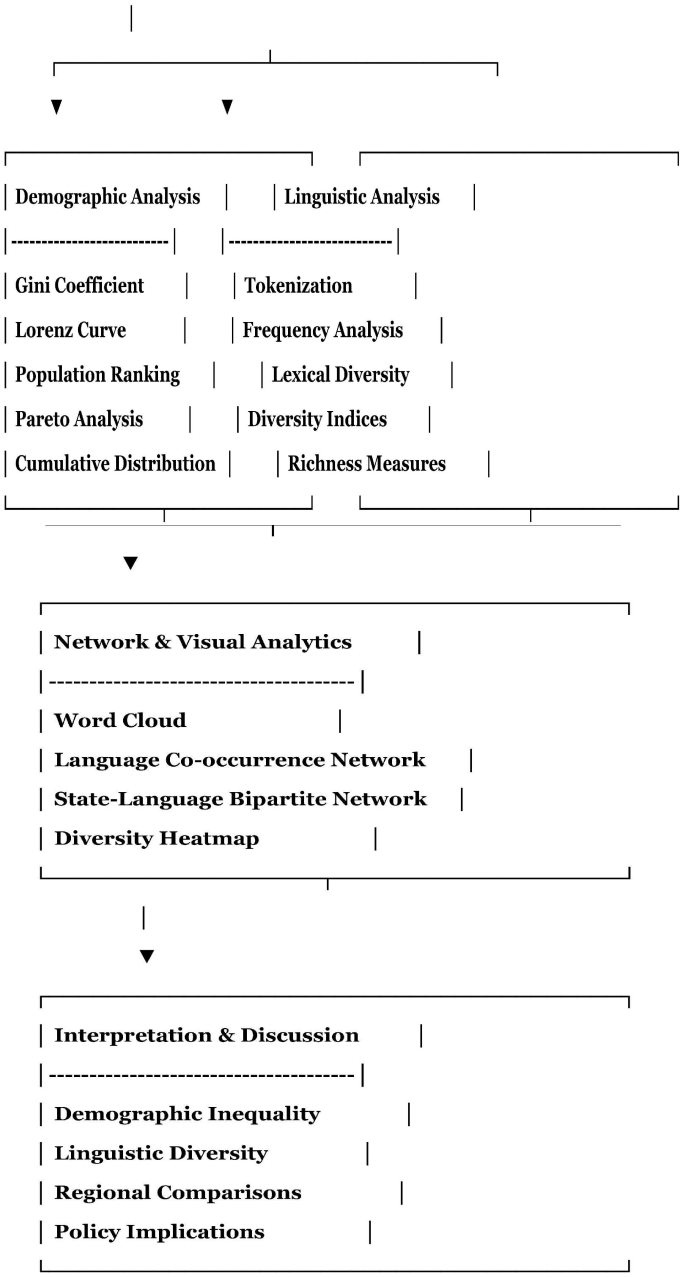


Figure 1. Workflow

3.2 Dataset Description

The dataset is structured as a single, relational flat-file (typically in comma separated values format) optimized for algorithmic processing. Each row represents a distinct geopolitical entity (State or Union Territory), mapped across several attribute dimensions. The core variables can be categorized into three primary domains:

3.2.1 Administrative and Geographic Identifiers

• Geopolitical Name: The official designation of the State or Union Territory, serving as the primary key for string matching.

• Alpha-2 / ISO Code: Standardized two-letter codes (e.g., IN-MH for Maharashtra, IN-DL for Delhi) utilized to ensure interoperability with standard Geographic Information System (GIS) shapefiles and GeoJSON boundary objects.

• Zonal Classification: A categorical variable dividing the country into macro regions (North, South, East, West, Northeast, Central) to facilitate spatial stratification.

3.2.2 Demographic and Morphological Metrics

• Total Population (p): Absolute population counts, often disaggregated by gender (P_m, p_f) and residential status (Urban vs. Rural).

• Geographical Area (A): Expressed in square kilometers (Km^2), establishing the physical scale of the entity.

• Population Density (D): A derived metric representing spatial pressure, calculated as:

$$D = \frac{P}{A}$$

• Sex Ratio (SR): The demographic metric expressing the balance between sexes, defined as the number of females per 1,000 males:

$$SR = (\frac{P_f}{P_m}) \text{ x } 1000$$

3.2.3 Socioeconomic and Development Indices

Literacy Rate (%): The proportion of the population aged seven and above capable of reading and writing with understanding.

• Gross State Domestic Product (GSDP): The monetary value of all finished goods and services produced within the state borders over a specific financial year, serving as a proxy for regional economic capacity.

3.3 Data Applications and Research Utility

3.3.1 Geospatial and Choropleth Mapping

The integration of standardized ISO codes and state names allows researchers to seamlessly join this dataset with spatial vector data. It provides an immediate pipeline for rendering choropleth maps that visually isolate clusters of high or low development or severe demographic imbalances.

3.3.2 Exploratory Econometric Modeling

The dataset enables cross sectional econometric analysis. Researchers can run regression models to assess the strength of relationships between structural inputs (e.g., spatial density or regional location) and human capital outputs (e.g., literacy rates or GSDP per capita).

3.3.3 Normalization and Auxiliary Merging

In larger macro studies such as those evaluating pan Indian public health datasets, crime statistics, or climate vulnerability this dataset acts as a vital auxiliary lookup table. Raw absolute figures from external sources can be normalized by merging them with this dataset's population and area metrics, transforming raw metrics into comparable per capita or per unit area rates.

4. Methodology

4.1 Data Acquisition

Demographic and linguistic information for all 36 Indian States and Union Territories was collected from publicly available government datasets and validated statistical sources. The dataset contains administrative identifiers, population statistics, geographical area, literacy rate, population density, Gross State Domestic Product (GSDP), and widely spoken languages.

4.2 Data Preprocessing

The acquired dataset underwent several preprocessing operations to ensure consistency and analytical reliability.

These operations included:

- duplicate removal
- missing value verification
- ISO code validation
- normalization of numerical variables
- standardization of state names
- language tokenization
- language spelling normalization
- feature extraction

4.2.1 Population Inequality Analysis

4.3 Gini Coefficient

The Gini coefficient was used to quantify inequality in population distribution across the 36 Indian States and Union Territories. This measure evaluates the extent to which the observed population distribution deviates from a perfectly equal distribution. The Gini coefficient ranges from 0 (perfect equality) to 1 (perfect inequality), with larger values indicating greater demographic concentration.

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2\mu}$$

where:

- x_i = population of state, i
- n= number of states/UTs
- μ= mean population

Gini	Interpretation
0.00	Perfect equality
0.10–0.30	Low inequality
0.30–0.50	Moderate inequality
0.50–0.70	High inequality
>0.70	Very high inequality

Table 1. Gini Coefficient Results

The coefficient was computed using the population values of all States and Union Territories after arranging the observations in ascending order. The calculation considers pairwise differences between population values relative to the overall population mean and provides a single summary measure of demographic inequality.

4.4 Lorenz Curve

The Lorenz curve was constructed to provide a graphical representation of population inequality. Initially, all States and Union Territories were ranked in ascending order of population. Subsequently, cumulative percentages for administrative units and population were calculated.

The Lorenz curve was generated by plotting cumulative percentages of States/UTs on the horizontal axis against cumulative percentages of population on the vertical axis. A 45° reference line representing perfect equality was included to facilitate comparison. The greater the deviation of the Lorenz curve from this equality line, the greater the inequality in the population distribution.

4.5 Population Rank Analysis

Population rank analysis was performed by arranging all States and Union Territories in descending order of population. This analysis provides a straightforward visualization of demographic dominance and enables comparison between highly populated and sparsely populated administrative regions.

4.5.1 Cumulative Population Distribution

To evaluate demographic concentration, cumulative population percentages were computed after sorting States and Union Territories by population in descending order. This analysis illustrates how rapidly the nati -onal population accumulates among the highest ranked States.

4.5.2 Pareto Analysis

Pareto analysis was conducted to determine whether a relatively small proportion of States accounts for the majority of the total population. Population values were arranged in descending order, and cumulative percentages were computed. A horizontal reference line representing the 80% cumulative threshold was incorporated to evaluate the applicability of the Pareto principle within the demographic distribution.

4.5.3 Lexical Diversity Analysis

Lexical diversity analysis was conducted to quantify the richness, heterogeneity, and distribution of widely spoken languages across the 36 Indian States and Union Territories. The *Widely Spoken Languages* attribute was first preprocessed by separating multiple language entries within each state into individual language tokens. Duplicate language names within the entire corpus were removed only for estimating vocabulary size (types), while repeated occurrences across different States/UTs were retained for frequency based analyses.

The resulting corpus was represented using two fundamental lexical units: tokens, representing the total number of language occurrences, and types, representing the number of distinct languages identified in the dataset. Based on these representations, several established lexical diversity measures were computed to evaluate linguistic richness and distributional characteristics.

The Type Token Ratio (TTR) was calculated as the ratio of unique language types to total language tokens, providing a preliminary estimate of lexical richness. Since TTR is sensitive to corpus size, Herdan's C and Guiraud's Index were also computed to obtain more robust estimates of lexical diversity less influenced by the total number of observations.

To assess both richness and evenness of language occurrences, two widely adopted ecological diversity measures were employed. The Shannon Diversity Index quantified the uncertainty associated with predicting a randomly selected language occurrence, thereby reflecting both the number of languages and their relative frequencies. The Simpson Diversity Index measured the probability that two randomly selected language occurrences belonged to different languages, emphasizing dominance and diversity within the corpus.

Furthermore, Pielou's Evenness Index was calculated to evaluate how uniformly language occurrences were

distributed among the identified languages. Finally, the Dominance Index was estimated to determine whether a small number of languages dominated the linguistic landscape or whether language occurrences were relatively balanced across the dataset.

Together, these complementary measures provide a comprehensive quantitative assessment of lexical richness, diversity, and evenness, enabling a robust characterization of the linguistic landscape represented by the Indian States and Union Territories.

5. Results and Discussion

5.1 Population Inequality

The computed Gini coefficient was 0.6149, indicating a high level of population inequality among Indian States and Union Territories. This value suggests that population is concentrated within a relatively small number of highly populated States, whereas many Union Territories contain comparatively small populations. According to commonly accepted interpretation thresholds, a Gini coefficient exceeding 0.50 represents substantial inequality, confirming pronounced demographic concentration within India.

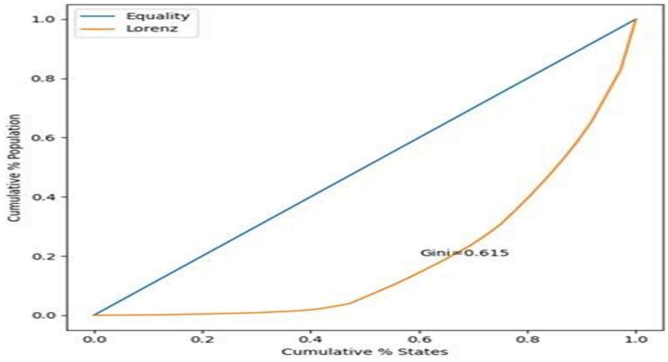


Figure 2. Lorenz Curve

Figure 2 presents the Lorenz curve together with the 45° line of perfect equality. The curve deviates considerably from the equality line, indicating that the population is unevenly distributed among the States and Union Territories. The sizeable area between the Lorenz curve and the equality line corresponds to the computed Gini coefficient, providing graphical evidence of significant demographic inequality. These findings suggest that a limited number of States account for a disproportionately large share of the national population, which has implications for regional development, infrastructure planning, and public resource allocation.

5.3 Population Ranking

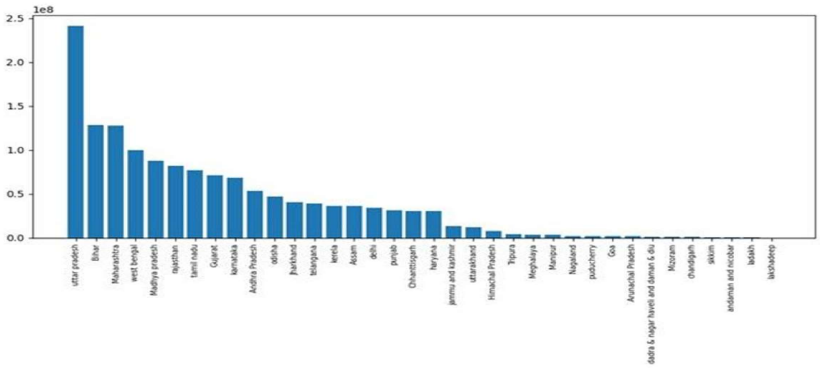


Figure 3. Ranking of states by popluation

Figure 3 ranks all States and Union Territories according to population. The ranking shows a steep decline from the most populous States to smaller administrative regions, highlighting substantial differences in population size. The higher ranked States account for a dominant share of the national population, whereas numerous Union Territories represent only a small fraction of the total population.

5.4 Cumulative Population Distribution

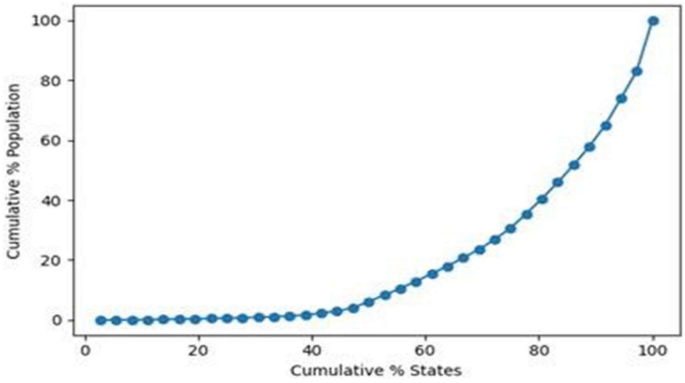


Figure 4. Cumulative percentage of States/UTs

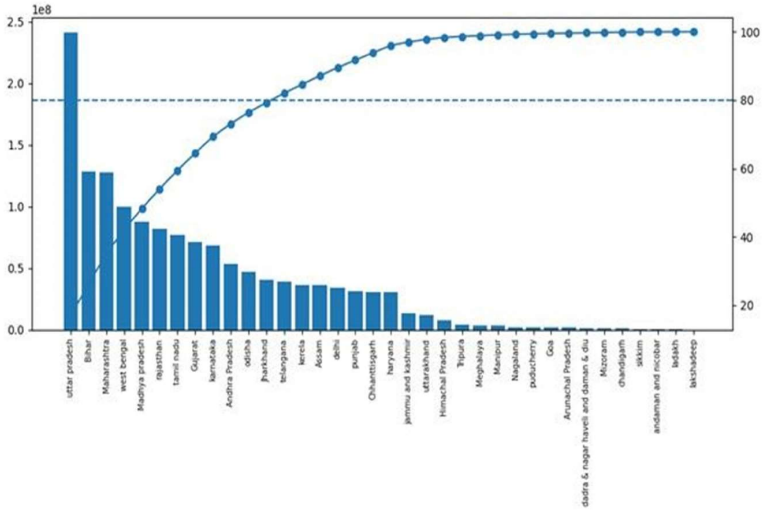


Figure 5. Cumulative percentage of population

Figures 4 and 5 illustrate the cumulative distribution of population across Indian States and Union Territories. The cumulative population curve rises sharply during the initial ranks, indicating that a relatively small proportion of States contributes a majority of the national population. The observed pattern confirms strong demographic concentration and supports the inequality identified by the Gini coefficient.

5.5 Pareto Analysis

The bars represent State populations in descending order, while the cumulative percentage curve shows the cumulative growth of the national population. The intersection with the 80% reference line indicates that a relatively small subset of States accounts for approximately four-fifths of India's population. This finding confirms the presence of a Pareto type distribution, where demographic concentration is heavily skewed toward a limited number of highly populated States.

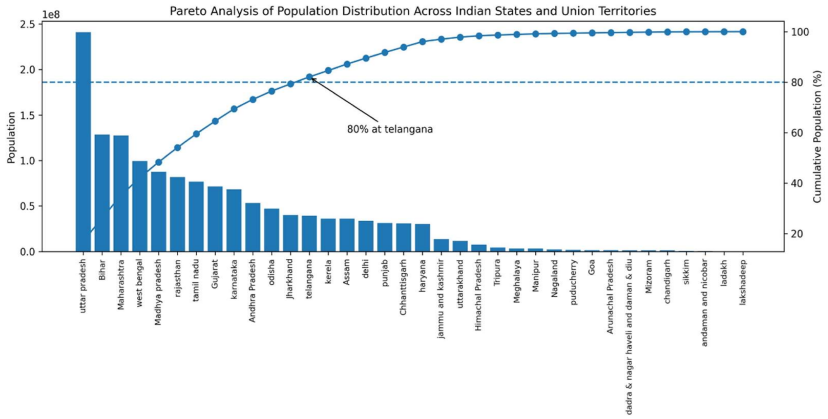


Figure 6. Pareto chart of population distribution

5.6 Discussion

The combined evidence from the Gini coefficient, Lorenz curve, population ranking, cumulative population distribution, and Pareto analysis consistently indicates pronounced demographic inequality among Indian States and Union Territories. While a few highly populated States dominate the national demographic landscape, many smaller States and Union Territories contribute comparatively little to the overall population. Such demographic concentration has important implications for equitable resource allocation, infrastructure development, governance, healthcare delivery, education planning, and regional policy formulation. These findings also provide a quantitative foundation for subsequent analyses of linguistic diversity, language distribution, and socio demographic relationships within the dataset.

5.7 Lexical Diversity Statistics

Table 2 presents the lexical diversity statistics computed from the dataset.

Metric	**Value**
Total States/Union Territories	36
Total Language Occurrences (Tokens)	71
Unique Languages (Types)	43
Average Languages per State/UT	1.972
Minimum Languages per State/UT	1
Maximum Languages per State/ UT	4
Type–Token Ratio (TTR)	0.6056
Herdan's C	0.8824
Guiraud's Index	5.1032
Shannon Diversity Index	3.4422

Simpson's Diversity Index	0.9490
Pielou's Evenness	0.9152
Dominance Index	0.0510

Table 2. Lexical Diversity Statistics

5.8 Results and Interpretation

The lexical diversity analysis reveals a highly heterogeneous linguistic composition across the Indian States and Union Territories. A total of 71 language occurrences corresponding to 43 distinct languages were identified, demonstrating substantial linguistic richness within the dataset. On average, each State/UT is associated with approximately 1.97 widely spoken languages, with the number of reported languages ranging from one to four, indicating considerable regional variation in multilingual representation.

The computed Type Token Ratio (0.6056) exceeds the commonly accepted threshold for high lexical diversity, suggesting that more than 60% of all language occurrences correspond to unique language types. This finding indicates that the linguistic landscape is characterized by considerable variety rather than repeated occurrences of only a few dominant languages.

Because TTR may be influenced by corpus size, additional richness measures were examined. Herdan's C (0.8824) and Guiraud's Index (5.1032) consistently indicate a rich and diverse vocabulary, confirming that the observed lexical richness is not merely a consequence of the relatively small dataset. These complementary measures reinforce the conclusion that the linguistic inventory represented across Indian States and Union Territories is extensive and varied.

The diversity indices further highlight the balanced distribution of language occurrences. The Shannon Diversity Index (3.4422) reflects substantial linguistic heterogeneity by accounting for both the number of distinct languages and their relative frequencies. Similarly, the Simpson Diversity Index (0.9490) approaches its theoretical maximum, indicating a very high probability that two randomly selected language occurrences belong to different languages. These results demonstrate that the dataset exhibits not only high richness but also strong diversity without excessive concentration among a limited set of languages.

The Pielou's Evenness Index (0.9152) further indicates that language occurrences are distributed relatively uniformly across the identified language inventory. Such a high evenness value suggests that linguistic representation is balanced rather than dominated by only a few frequently occurring languages. This observation is corroborated by the Dominance Index (0.0510), whose low value indicates minimal concentration and confirms that no single language overwhelmingly dominates the corpus.

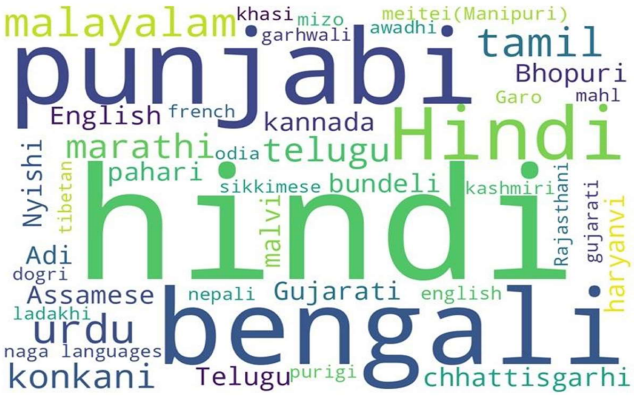


Figure 7. Word Cloud of Language Occurrences

Overall, the lexical diversity measures consistently demonstrate that the dataset captures a broad and well-balanced representation of India's multilingual landscape. The combination of high lexical richness, strong diversity, high evenness, and low dominance provides empirical evidence of substantial linguistic heterogeneity across the 36 Indian States and Union Territories. These findings establish a strong foundation for subsequent analyses, including language frequency analysis, co-occurrence networks, community detection, clustering, correspondence analysis, and regional language similarity modelling.

The size of each language label is proportional to its frequency across the Indian States and Union Territories, providing a visual summary of the most widely represented languages.

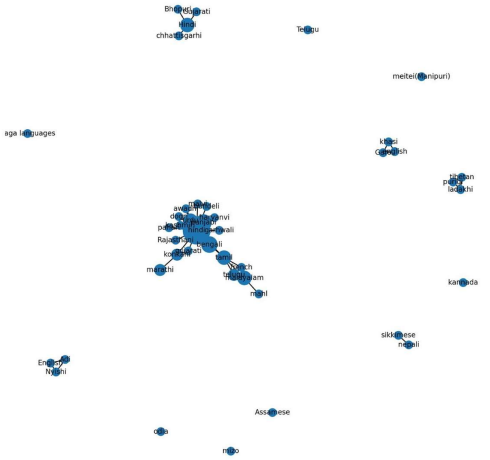


Figure 8. Language Co-occurrence Network

Nodes represent languages, while edges connect languages that co-occur within the same State or Union Territory. Larger and more connected nodes indicate languages that frequently appear alongside others, highlighting patterns of multilingual coexistence.

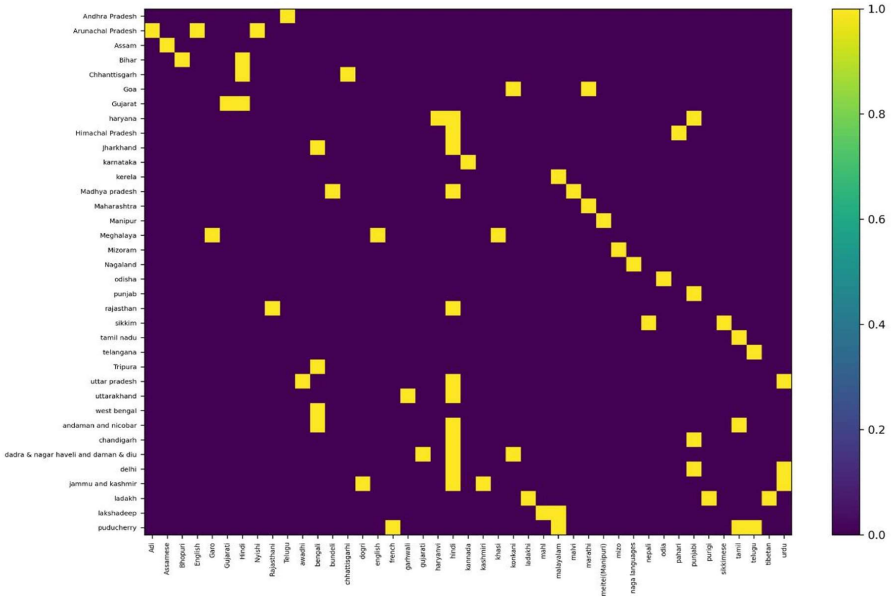


Figure 9. Language Diversity Heatmap

The heatmap illustrates the presence (1) or absence (0) of each language across all States and Union Territories. Rows correspond to States/UTs and columns represent unique languages, enabling comparison of regional linguistic diversity.

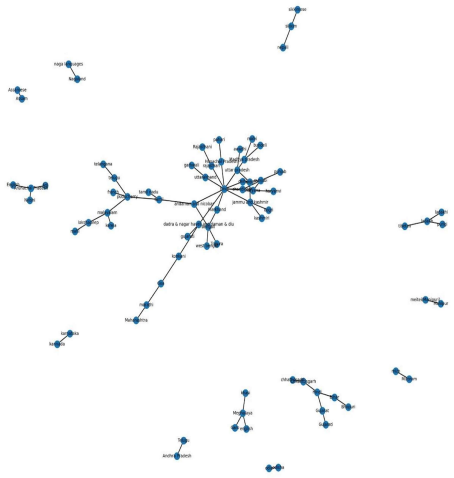


Figure 10. State Language Bipartite Network

The bipartite graph connects States/Union Territories with their widely spoken languages. One node set represents administrative regions and the other represents languages, revealing the distribution and sharing of languages across different regions.

6. Discussion

The present study provides a comprehensive quantitative assessment of demographic inequality and linguistic diversity across the 36 Indian States and Union Territories by integrating demographic statistics with computational lexical diversity analysis. The findings reveal that India's demographic landscape is highly concentrated, while its linguistic landscape remains remarkably diverse and well distributed. The computed Gini coefficient (0.6149), supported by the Lorenz curve, cumulative population distribution, and Pareto analysis, demonstrates substantial population inequality, indicating that a relatively small number of States account for a disproportionately large share of the national population. Such demographic concentration has important implications for governance, infrastructure development, healthcare delivery, educational planning, and equitable allocation of public resources.

In contrast to the unequal demographic distribution, the lexical diversity analysis reveals a rich and balanced multilingual landscape. The identification of 43 unique languages from 71 language occurrences, together with a high Type Token Ratio (0.6056), Herdan's C (0.8824), and Guiraud's Index (5.1032), demonstrates considerable linguistic richness across Indian States and Union Territories. Furthermore, the high Shannon Diversity Index (3.4422), Simpson Diversity Index (0.9490), and Pielou's Evenness (0.9152), coupled with the low Dominance Index (0.0510), indicate that India's linguistic diversity is characterized not only by richness but also by an equitable distribution of languages without excessive dominance by a small number of languages.

The complementary visual analyses further reinforce these observations. The word cloud highlights the most frequently represented languages, while the language co occurrence network illustrates multilingual interactions among languages spoken within the same States and Union Territories. Similarly, the language diversity heatmap and state language bipartite network demonstrate the widespread distribution and sharing of languages across regions, emphasizing India's multilingual interconnectedness rather than isolated linguistic clusters. These visual representations provide structural insights that extend beyond conventional descriptive statistics by revealing patterns of multilingual coexistence and regional language associations.

Overall, the findings indicate that demographic concentration and linguistic diversity represent two distinct but interconnected dimensions of India's regional heterogeneity. While population is concentrated within relatively few administrative regions, linguistic diversity remains broadly distributed throughout the country. This distinction highlights the importance of considering demographic and linguistic characteristics simultaneously when designing public policies related to education, language preservation, regional development, digital inclusion, and multilingual governance. The proposed analytical framework further demonstrates that combining demographic inequality measures with computational linguistic diversity metrics provides a robust and scalable methodology for investigating multilingual societies at regional and national scales.

7. Conclusion

This study presented an integrated analytical framework for examining demographic inequality and linguistic diversity across the Indian States and Union Territories. By combining measures of demographic inequality, lexical diversity indices, and network based visual analytics, the study provides a comprehensive quantitative understanding of India's population distribution and multilingual landscape. The results demonstrate pronounced demographic inequality, with population concentrated in a relatively small number of States, while simultaneously revealing a highly diverse, balanced, and interconnected linguistic ecosystem.

The proposed framework illustrates the value of integrating statistical inequality measures with computational linguistic analysis to evaluate regional diversity from multiple perspectives. Beyond describing population and language distributions, the study provides reproducible analytical procedures that can support evidence-based policymaking in regional planning, education, language preservation, digital governance, and inclusive public service delivery. The combination of inequality analysis, ecological diversity metrics, and network visualization also offers a generalizable methodology applicable to other multilingual countries.

Despite these contributions, the present study is based on cross sectional data and focuses primarily on widely spoken languages. Future research could incorporate temporal census data to examine longitudinal demographic and linguistic changes, include additional socioeconomic indicators such as migration, income, education, and urbanization, and apply advanced machine learning, spatial econometric models, graph neural networks, and geospatial analyses to investigate the complex relationships among demographic dynamics, economic development, and language evolution. Such extensions would provide a deeper understanding of multilingual societies and support more effective policies for preserving linguistic diversity while promoting balanced regional development.

References

[1] Sutherland, W. J. (2003). Parallel extinction risk and global distribution of languages and species. *Nature*, *423*(6937), 276–279. https://doi.org/10.1038/nature01607.

[2] Eberhard, D. M., Simons, G. F., Fennig, C. D. (Eds.). (2019). *Ethnologue: Languages of the world* (22nd ed.). SIL International. https://www.ethnologue.com.

[3] Moseley, C. (Ed.). (2010). Atlas of the world's languages in danger (3rd ed.). UNESCO Publishing. http://www.unesco.org/culture/en/endangeredlanguages/atlas.

[4] University of Hawai»i at Mânoa. (2020). Catalogue of Endangered Languages. http://www.end angeredlanguages.com.

[5] Campbell, L., Okura, E. (2018). New knowledge producing a new catalog of the world's endangered languages. In L. Campbell A. Belew (Eds.), *Cataloguing the world's endangered languages* (1st ed., p. 79–84). Routledge.

[6] Romaine, S. (2017). Biodiversity and linguistic diversity. In A. F. Fill H. Penz (Eds.), *The Routledge handbook of ecolinguistics* (Chap. 3). Routledge.

[7] Bowern, C. (2017). Language vitality: Theorizing language loss, shift, and reclamation (Response to Mufwene). *Language*, *93*(4), e243–e253. https://doi.org/10.1353/lan.2017.0069.

[8] Mufwene, S. S. (2017). Language vitality: The weak theoretical underpinnings of what can be an exciting research area. *Language*, *93*(4), e202–e223. https://doi.org/10.1353/lan.2017.0066.

[9] Grenoble, L. A., Whaley, L. J. (1998). Toward a typology of language endangerment. In L. A. Grenoble & L. J. Whaley (Eds.), *Endangered languages: Language loss and community response* (p. 22–54). Cambridge University Press.

[10] Hua, X., Greenhill, S. J., Cardillo, M., Schneemann, H., & Bromham, L. (2019). The ecological drivers of variation in global language diversity. *Nature Communications*, *10*, Article 2047. https://doi.org/10.1038/s41467-019-09842-2.

[11] Cardillo, M., Bromham, L., Greenhill, S. J. (2015). Links between language diversity and species richness can be confounded by spatial autocorrelation. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1801), Article 20142986. https://doi.org/10.1098/rspb.2014.2986.

[12] Amano, T., Sandel, B., Eager, H., Bulteau, E., Svenning, J. C., Dalsgaard, B., Sutherland, W. J. (2014). Global distribution and drivers of language extinction risk. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1793), Article 20141574. https://doi.org/10.1098/rspb.2014.1574.

[13] Loh, J., Harmon, D. (2014). *Biocultural diversity: Threatened species, endangered languages*. WWF.

[14] Bromham, L., Dinnage, R., Skirgård, H., Luke, A., Hodgetts, J., Greenhill, S. J., Cardillo, M. (2022). Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology Evolution*, *6*(2), 163–173. https://doi.org/10.1038/s41559-021-01604.

[15] Aikhenvald, A. Y. (2004). Language endangerment in the Pacific Rim: An overview. In O. Sakiyama F. Endo (Eds.), *Lectures on endangered languages: 5. Endangered languages of the Pacific Rim* (p. 97–142). ELPR.

[16] Muysken, P. (1981). Halfway between Quechua and Spanish: The case for relexification. In A. Highfield & A. Valdman (Eds.), *Historicity and variation in Creole studies* (p. 52–78). Karoma.

[17] Gal, S. (1979). *Language shift: Social determinants of linguistic change in bilingual Austria*. Academic *Press*.

[18] Holmquist, J. (1985). Social correlates of a linguistic variable: A study in a Spanish village. *Language in Society*, *14*(2), 191–203. https://doi.org/10.1017/S004740450001112.

[19] Dobrin, L. M. (2014). The socioeconomics of linguistic heritage. In P. K. Austin J. Sallabank (Eds.), *Endangered languages: Beliefs and ideologies in language documentation and revitalization* (Chap. 7). British Academy. https://doi.org/10.5871/bacad/9780197265765.003.0008.

[20] Skirgård, H., Haynie, H. J., Blasi, D. E., Hammarström, H., Collins, J., Latarche, J. J., Gray, R. D. (2023). Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, *9*(16), eadg6175. https://doi.org/10.1126/sciadv.adg6175.

[21] Brinkmann, L. M., Duarte, J., Melo-Pfeifer, S. (2022). Promoting plurilingualism through linguistic landscapes: A multi-method and multisite study in Germany and the Netherlands. *TESL Canada Journal*, *38*(2), 88–112. https://doi.org/10.18806/tesl.v38i2.1354.

[22] Dafouz, E. (2024). Exploring the conceptualisation of linguistic diversity and multilingualism in the construction of (Transnational) European Universities: The case of UNA Europa. *Journal of Multilingual and Multicultural Development*, *45*(4), 759–774. https://doi.org/10.1080/01434632.2021.1920964.

[23] Uekusa, S. (2019). Disaster linguicism: Linguistics minorities in disasters. *Language in Society*, *48*(3), 353–375. https://doi.org/10.1017/s0047404519000150.

[24] Phyak, P. (2020, December 10). *Citizen multilingualism: Communication during the COVID-19 pandemic in Nepal*. Australia-Himalaya Research Network. https://aushimalaya.net/2020/12/10/citizenmultilingualism-communication-during-the-covid-19-pandemic-in-nepal/.

[25] Chen, C.-M. (2020). Public health messages about COVID-19 prevention in multilingual Taiwan. *Multilingua*, *39*(5), 597–606. https://doi.org/10.1515/multi-2020-0092.

[26] Ahmed, A. (2019, October 15). UAE now hosts more than 13,000 Korean residents as it becomes their favourite destination. *The Gulf News*. https://gulfnews.com/uae/uae-now-hosts-more-than-13000-korean-residents-as-it-has-become-their-favourite-destination-1.66850583.

[27] Guo, Y., Shang, G., Vazirgiannis, M., Clavel, C. (2024). The curious decline of linguistic diversity: Training language models on synthetic text. *In* Findings of the Association for Computational Linguistics: NAACL *2024* (p. 3589–3604). *Association for Computational Linguistics*. https://doi.org/10.18653/v1/2024.findi ngs-naacl.228

[28] van Driem, G. (2007). Endangered languages of South Asia. In M. Brenzinger (Ed.), *Language diversity endangered* (Chap. 14). Mouton de Gruyter.