










## ARTICLE

# Unveiling the Dynamics of Sociolinguistics, Understanding Language in Social Contexts, Artificial Intelligence Effect

Sarjan Islam Sadigova<sup>1\*</sup> , Gunel Mamadova Mahammad<sup>2</sup> , Leyla Nazir Nazirzada<sup>3</sup> , Javahir Geyis Aghayeva<sup>4</sup> , Dina Antar<sup>5</sup> , Naila Yusif Yusifova<sup>6</sup> , Farida Shahlar Shukurova<sup>7</sup> , Pankaj Kalita<sup>8</sup> ,  
Reena Sanasam<sup>9</sup> 

<sup>1</sup> Department of English Language and Translation, Faculty of Foreign Languages, Nakhchivan State University, University Campus, Nakhchivan AZ7012, Azerbaijan

<sup>2</sup> Department of Primary Education Pedagogy, Faculty of Primary Education, Azerbaijan State Pedagogical University, 34 Uzeyir Hajibeyov St., Baku AZ1000, Azerbaijan

<sup>3</sup> Department of Humanities, Azerbaijan State University of Economics (UNEC) Street Istiqlaliyyat, 6, Baku AZ100, Azerbaijan

<sup>4</sup> Scientific Research Branches, The Military Institute named after Heydar Aliyev, Baku AZ1018, Azerbaijan

<sup>5</sup> Department of Humanities and Social Sciences, School of Arts and Sciences, American University of Ras Al Khaimah, Ras al Khaimah P.O. Box 10021, United Arab Emirates

<sup>6</sup> Department of Azerbaijani Language, Azerbaijan State University of Economics (UNEC) Street Istiqlaliyyat, 6, Baku AZ100, Azerbaijan

<sup>7</sup> Department of Foreign Language Teaching Methodology, Azerbaijan University of Languages, 134 Rashid Behbudov St, Baku AZ1014, Azerbaijan

<sup>8</sup> Department of English, Sibsagar University, Joysagar, Sivasagar, Assam 785665, India

<sup>9</sup> Department of Humanities and Social Sciences, National Institute of Technology, Silchar, Assam 788014, India

### \*CORRESPONDING AUTHOR:

Sarjan Islam Sadigova, Department of English Language and Translation, Faculty of Foreign Languages, Nakhchivan State University, University Campus, Nakhchivan AZ7012, Azerbaijan; Email: [sercansadiqova@ndu.edu.az](mailto:sercansadiqova@ndu.edu.az)

### ARTICLE INFO

Received: 7 August 2025 | Revised: 29 August 2025 | Accepted: 9 September 2025 | Published Online: 5 November 2025

DOI: <https://doi.org/10.30564/fls.v7i12.11533>

### CITATION

Sadigova, S.I., Mahammad, G.M., Nazirzada, L.N., et al., 2025. Unveiling the Dynamics of Sociolinguistics, Understanding Language in Social Contexts, Artificial Intelligence Effect. *Forum for Linguistic Studies*. 7(12): 316–327. DOI: <https://doi.org/10.30564/fls.v7i12.11533>

### COPYRIGHT

Copyright © 2025 by the author(s). Published by Bilingual Publishing Group. This is an open access article under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License (<https://creativecommons.org/licenses/by-nc/4.0/>).

## ABSTRACT

This paper examines the evolving dynamics of sociolinguistics in the 21st century, focusing on how language reflects and shapes social realities in diverse societies, with a comparative lens on India and Azerbaijan. Both nations present postcolonial, multilingual, and digitally transforming complex linguistic ecosystems where power, identity, and tradition are encoded in language practices. By integrating traditional sociolinguistic theory with contemporary developments in Artificial Intelligence (AI), this research explores how technological mediation is reconfiguring linguistic hierarchies, access, and representation. The study adopts a comparative sociolinguistic methodology, combining ethnographic insights, AI-enabled linguistic corpus analysis, and critical discourse analysis of public and social media. It interrogates how variationist phenomena (e.g., code-switching, diglossia, lexical borrowing) operate across caste, ethnicity, and region in India, and across post-Soviet national identity formations in Azerbaijan. The impact of AI—particularly Natural Language Processing (NLP), machine learning-based dialect analysis, and voice recognition algorithms—is analyzed for its dual role: as a democratizing force in linguistic research and as a potential agent of linguistic erasure and bias. Drawing from scholars such as Labov, Hymes, Woolard, and Gikandi, the paper argues that AI tools are often trained on dominant linguistic codes, reinforcing existing inequalities. Language embodies not only communication but “a repository of memory and identity,” which technology risks flattening through algorithmic standardization.

**Keywords:** Linguistic Codes; Social Contexts; Artificial Intelligence; Ethnographic Insights; Language; Linguistic Norms

## 1. Introduction

Language is more than a medium of communication—it is an index of identity, power, and social belonging. In the field of sociolinguistics, language is analyzed in relation to the structures and practices of everyday life, where variables such as class, caste, gender, and geography shape and are shaped by linguistic behavior. In recent years, the field has encountered a paradigmatic shift with the advent of Artificial Intelligence (AI), particularly Natural Language Processing (NLP) and machine learning, which have begun to reshape how language is documented, analyzed, and even produced. This paper investigates the dynamics of sociolinguistic variation and the impact of AI in two linguistically rich but understudied contexts: India and Azerbaijan. These countries offer unique grounds for comparative research. India, with its constitutionally recognized multilingualism and deep-rooted caste and regional divisions, presents a landscape where language reflects hierarchical social stratification. Azerbaijan, shaped by Persian, Russian, and Turkic influences, reflects the post-Soviet reconfiguration of national and minority language identities. Despite their different geopolitical histories, both nations confront similar tensions: between dominant and marginal languages, between official discourse and vernacular speech, and increasingly, between human agency and machine processing. The integration of AI in sociolin-

guistic inquiry brings opportunities and risks. On one hand, AI tools allow researchers to process large linguistic datasets, enabling dialect recognition, sentiment analysis, and real-time discourse tracking at an unprecedented scale. On the other hand, these technologies are often trained on standardized or elite language forms, thereby reproducing existing linguistic hierarchies and erasing the complexity of dialectal and cultural nuance. As Gikandi<sup>[1]</sup> (p. 74) argues, language carries with it “a repository of memory and identity,” and any attempt to algorithmically encode language is inevitably an act of cultural selection and exclusion.

In this context, the paper addresses the following key research questions:

1. How do social factors influence language variation and identity formation in India and Azerbaijan?
2. What roles do Artificial Intelligence tools play in either reinforcing or challenging sociolinguistic hierarchies?
3. How can sociolinguistic theory adapt to analyze the new linguistic landscapes shaped by AI?

The study draws on multiple theoretical frameworks, including Labovian variationism, ethnography of communication (Hymes), and language ideology theory to frame the linguistic data<sup>[2]</sup>. It also incorporates technological mediation theory to critique the assumptions embedded in AI tools applied to language. This research is significant for

several reasons. First, it brings comparative insight into two non-Western contexts that are frequently marginalized in both AI development and sociolinguistic theory. Second, it critiques the techno-linguistic assumptions of AI through the lens of social inequality and identity politics. Third, it provides policy-level recommendations for the ethical design and implementation of AI systems that interact with natural language—especially in multilingual and culturally diverse societies.

The paper proceeds in the following structure: a literature review of sociolinguistic theory and recent AI developments; a detailed methodological framework; sociolinguistic profiles of India and Azerbaijan; an examination of how AI tools are shaping linguistic landscapes; a comparative analysis; and a concluding discussion on challenges, implications, and future directions.

## 2. Literature Review

The study of sociolinguistics has evolved considerably since its emergence in the mid-twentieth century, engaging with a range of perspectives that examine the interface between language and society. Foundational work by William Labov<sup>[3]</sup> laid the groundwork for the variationist approach, which analyses how linguistic features correlate with social variables such as class, gender, and ethnicity. Labov's empirical methods provided a systematic basis for understanding how language varies within speech communities and how these variations are embedded in social structures. This model has been influential globally, including in multilingual societies such as India and Azerbaijan, although its application has often required significant contextual adaptation. In contrast, Dell Hymes' ethnography of communication introduced a more context-sensitive lens, emphasizing the social functions of language and the culturally specific rules that govern its use<sup>[4]</sup>. Hymes moved beyond structuralist models to consider how communicative competence involves knowing not only grammatical forms but also how, when, and why to speak in specific social settings. This approach is particularly relevant in postcolonial contexts where linguistic norms are deeply entwined with histories of domination, resistance, and identity formation. Recent literature has increasingly incorporated critical approaches, such as language ideology theory, which interrogates how beliefs about

language reflect and reproduce power relations. Woolard and Schieffelin<sup>[4]</sup> contend that language ideologies are not merely abstract beliefs but are socially situated and materially consequential. This framework has been applied to explore how language hierarchies are sustained through state policies, media discourses, and educational systems, especially in settings marked by linguistic diversity and asymmetrical power structures. In the Indian context, the sociolinguistic terrain is marked by a stratified multilingualism, where dominant languages such as Hindi and English often marginalize regional dialects and tribal languages. Scholars like Annamalai<sup>[5]</sup> and Mohanty<sup>[6]</sup> have highlighted how language use is closely linked to caste, class, and regional identity, creating complex patterns of code-switching, diglossia, and language shift. Hinglish, for instance, has emerged as a hybrid urban register that simultaneously indexes modernity and linguistic fluidity, while excluding those without access to English education or digital media. Azerbaijan, although it has a different colonial experience, presents its own set of sociolinguistic complexities. The post-Soviet language reforms and the revival of Turkic identity have restructured the linguistic field, displacing Russian while negotiating the status of Persian-influenced dialects and minority languages such as Lezgi, and the state-driven emphasis on linguistic purity and national unity often comes at the expense of vernacular and minority expressions, echoing broader concerns in language ideology literature. A growing body of research has begun to explore the implications of Artificial Intelligence on sociolinguistic research and practice. Natural Language Processing (NLP) and machine learning technologies are increasingly used to collect, process, and analyze large-scale linguistic data. Eisenstein<sup>[7]</sup> and Hovy and Spruit<sup>[8]</sup> caution, however, that these tools often reflect the biases of their training data, which tend to prioritize standard, dominant language varieties. As such, AI can inadvertently perpetuate linguistic inequalities by encoding and automating existing social biases. Corpus linguistics has benefited significantly from AI-based tools, allowing for the efficient tagging, annotation, and analysis of texts across dialects and domains<sup>[9]</sup>. Yet, the integration of sociolinguistic theory into computational linguistics remains uneven. Most AI models treat language as an abstract system, divorced from its social context, leading to challenges in recognizing variationist or pragmatic features such as regional slang, code-mixing, or context-specific

politeness strategies. Gikandi<sup>[1]</sup> (p. 74) offers a crucial intervention by framing language as a repository of cultural memory and identity, particularly within postcolonial societies. He argues that linguistic practices embody histories of trauma, resistance, and survival, and any attempt to standardize or algorithmically regulate them risks a form of epistemic violence. This perspective is critical when considering how AI systems are deployed in multilingual contexts such as India and Azerbaijan, where the stakes of linguistic representation are deeply political. Despite the growing literature on AI and language, there is a lack of comparative, global south-oriented research that critically examines how these technologies interact with sociolinguistic variation. This study addresses that gap by offering a dual-country analysis and applying sociolinguistic theory to emerging AI practices.

### 3. Research Methodology

This study adopts a comparative, interdisciplinary research design that integrates qualitative and quantitative sociolinguistic methods with emerging tools from computational linguistics and Artificial Intelligence. The central objective is to critically examine the role of social variables in shaping language practices in India and Azerbaijan, while also interrogating how AI technologies influence linguistic analysis, representation, and access in these distinct national contexts. The rationale for choosing India and Azerbaijan as comparative units lies in their shared features as multilingual, post-imperial societies undergoing rapid technological transformation. Yet, the two countries differ significantly in terms of their colonial legacies, language policies, socio-political structures, and technological infrastructures. India's multilingualism is constitutionally mandated, with 22 scheduled languages and hundreds of regional dialects, often stratified along caste, class, and religious lines. In contrast, Azerbaijan presents a more centralized linguistic environment, dominated by Azerbaijani (a Turkic language), but layered with the historical imprints of Persian and Russian, along with various minority languages, such as Lezgi, Talysh, and Udi. These contextual differences allow for a nuanced comparative analysis of how sociolinguistic variation is managed, contested, and transformed. The study employs a multi-scalar methodology structured around three core components: corpus-based linguistic analysis, ethnographic field observa-

tions, and AI-driven computational modelling. Each method contributes distinct insights and, when triangulated, offers a robust picture of the sociolinguistic ecologies under investigation. The first phase involves the construction and analysis of two parallel linguistic corpora—one for India and one for Azerbaijan. These corpora are drawn from regionally and socially diverse digital sources, including social media platforms (Twitter, Telegram, YouTube comments), regional news outlets, government documents, and online forums. The corpora are stratified based on geographic region (north/south/east/west for India; Baku, Nakhchivan, southern and northern Azerbaijan), linguistic group (dominant vs. minority), and demographic identifiers where available (age, gender, education). This phase enables the study of lexical variation, syntactic complexity, code-switching frequency, and register shifts, particularly in informal digital discourse. To process and annotate the corpora, the study utilizes AI-based Natural Language Processing (NLP) tools such as spaCy, UDPipe, and custom-trained machine learning models for code-switching detection and sentiment analysis<sup>[10]</sup>. These tools are adapted for multilingual analysis, including languages with limited digital resources such as Meitei (India) and Lezgi (Azerbaijan). For India, the study uses existing bilingual corpora such as the IIT Bombay Hindi-English corpus and incorporates vernacular datasets scraped from regional newspapers. For Azerbaijan, where fewer annotated corpora exist, semi-automated annotation is supplemented by manual validation through collaboration with native speakers and local researchers. Given the morphologically rich nature of Azerbaijani and several Indian languages, the study pays close attention to the limitations of current AI models in processing agglutinative and non-standard linguistic features. The second methodological component consists of ethnographic field observations and virtual interviews with language users, educators, and computational linguists in both countries. These qualitative data are used to contextualize the computational findings and provide insight into how speakers perceive, negotiate, and resist linguistic hierarchies. In India, fieldwork focuses on bilingual speakers in Uttar Pradesh (Hindi-English), Kerala (Malayalam-English), and Assam (Assamese-Bodo-Hindi), whereas in Azerbaijan, attention is given to linguistic minorities in Lankaran and Quba, alongside urban Azerbaijani speakers in Baku. Participants are selected through purposive and snowball sampling meth-

ods, and interviews are conducted in a mix of local languages and English, transcribed, and thematically coded. The third component introduces a critical-technological layer through the evaluation of AI models currently used in language analysis and education in both countries. This includes tools such as Google's speech-to-text API, voice assistants like Alexa and Google Assistant, and national ed-tech platforms. Their accuracy and inclusivity are tested across regional accents and non-standard dialects to assess how well AI technologies capture sociolinguistic variation. Preliminary tests indicate significant discrepancies: for example, speech recognition in India performs poorly on northeastern accents and low-caste Hindi registers, while NLP tools in Azerbaijan struggle with phonetic features unique to the Lezgi and Talysh languages. A central methodological concern is ethical representation and data sensitivity. All digital data is anonymized, and interviews are conducted with informed consent. Special care is taken not to reproduce linguistic hierarchies in the analysis itself, a challenge particularly relevant when using AI tools trained on dominant language forms. To ensure cross-cultural validity, the study applies context-specific analytical categories while maintaining a comparative framework. In India, caste and religion are foregrounded alongside regional identity, whereas in Azerbaijan, ethnicity and post-Soviet national discourse are primary axes of analysis. The comparative methodology does not seek equivalence but instead highlights asymmetries and convergences, allowing for a richer understanding of how language and technology mediate social life. This integrative approach allows the research to move beyond surface-level linguistic description toward a deeper sociotechnical analysis. It acknowledges the limits of AI tools in sociolinguistics while also demonstrating their potential—when appropriately localized and critically applied—to enhance our understanding of language in society.

#### **4. Sociolinguistic Dynamics in India and Azerbaijan**

The sociolinguistic landscapes of India and Azerbaijan are shaped by a complex interplay of history, identity, political ideology, and linguistic policy. Both countries are post-imperial, multilingual nations navigating the tensions between language standardization and the preservation of

linguistic diversity. However, the specific contours of these dynamics vary significantly due to differences in colonial legacy, demographic composition, and state ideology. This section provides a comparative analysis of key sociolinguistic features in both countries, focusing on language variation, code-switching practices, language ideologies, and the politics of language education and representation. India represents one of the most linguistically diverse countries in the world, with hundreds of languages spoken across its vast territory. According to the 2011 Census, there are 121 languages spoken by more than 10,000 people, with Hindi serving as the most widely spoken language. English, while not an indigenous language, holds a position of elite prestige and functions as a *de facto* official language in many urban and governmental contexts. The Indian sociolinguistic structure is deeply stratified along caste, class, and regional lines, with significant asymmetries in language access, representation, and legitimacy. Language use often indexes social status, educational background, and political affiliation. For instance, the use of English in professional and academic settings confers upward mobility, while regional languages are frequently relegated to the domestic or informal sphere. Code-switching and code-mixing are prominent in urban India, especially among bilingual and trilingual speakers<sup>[11]</sup>. The phenomenon of Hinglish—a blend of Hindi and English—is particularly illustrative of how linguistic hybridity functions as both a pragmatic and identity-affirming strategy. While Hinglish is often associated with urban middle-class youth, similar hybrid registers exist across other language pairs, such as Tanglish (Tamil-English) and Benglish (Bengali-English). These mixed codes function as sociolects, reflecting changing attitudes towards modernity, nationalism, and cultural authenticity. At the same time, they reveal the underlying inequalities in English language acquisition and usage, especially among marginalized communities.

The Indian education system institutionalizes these hierarchies through language policy. Although mother-tongue education is constitutionally encouraged, English-medium instruction is increasingly favoured by parents and students for its perceived economic value. This has led to the gradual decline of regional dialects and indigenous languages, particularly among younger generations. Dalit and tribal communities face a double burden, often receiving sub-standard education in either a dominant regional language

or English, neither of which may be their first language. As a result, linguistic marginalization is compounded by socio-economic exclusion. Azerbaijan, by contrast, has undergone a rapid and deliberate linguistic transformation since gaining independence from the Soviet Union in 1991. Azerbaijani, a Turkic language, is now the sole official language and a central marker of national identity. However, the linguistic landscape remains layered with the historical legacies of Russian and Persian, as well as the presence of minority languages such as Lezgi, Talysh, Kurdish, and Udi. While Azerbaijani is widely spoken, the level of fluency and usage varies by region, ethnicity, and age group. Russian continues to be used in elite and academic circles, particularly in urban centers such as Baku, while Persian-influenced dialects persist in southern regions close to the Iranian border.

The state's language policy has focused on linguistic unification and Turkification, promoting a purified and standardized form of Azerbaijani in public discourse and education. This process has involved the removal of Russian and Persian loanwords and the revival of Ottoman Turkish vocabulary, as well as the implementation of a Latin script to replace the Cyrillic script imposed during the Soviet era. While these reforms aim to assert linguistic sovereignty, they have also marginalized non-standard dialects and minority languages. Language use in Azerbaijan thus reflects broader ideological projects of nationalism, cultural revival, and historical reparation. Code-switching in Azerbaijan is less visibly institutionalized than in India but prevalent, especially among bilinguals who shift between Azerbaijani and Russian. In southern regions, code-mixing with Persian elements is observed, although it is less publicly acknowledged due to political sensitivities. Among minority groups, especially Lezgi and Talysh speakers, there is evidence of diglossic patterns where the home language is used in private settings and Azerbaijani in formal or public domains. Unlike in India, where multilingualism is celebrated rhetorically even as hierarchies persist, Azerbaijan's linguistic nationalism tends to subsume pluralism under a dominant monolingual ideal. The role of language in education further highlights these contrasts. In Azerbaijan, state schools use Azerbaijani as the medium of instruction, with Russian schools still functioning in urban centers<sup>[12]</sup>. However, there is minimal institutional support for minority languages. This lack of educational in-

frastructure contributes to the erosion of linguistic diversity, particularly in rural areas. In India, the presence of multiple state languages has led to more flexible regional education policies, but these are often undermined by the social capital attached to English. Thus, both countries demonstrate how educational systems can function as instruments of linguistic hegemony, albeit through different mechanisms.

Linguistic identity in both contexts is closely tied to political ideology and historical consciousness. In India, language movements—such as those in Tamil Nadu, Punjab, and the Northeast—have long challenged the imposition of Hindi and fought for regional recognition. These movements have often aligned with broader demands for cultural and political autonomy. In Azerbaijan, linguistic politics are less fragmented but still charged, with language serving as a key symbol in post-Soviet nation-building and the assertion of independence from Russian cultural influence. Digital communication has introduced new dimensions to sociolinguistic variation in both countries. In India, social media has amplified vernacular expression, particularly in video-based platforms like YouTube and Instagram. Memes, short videos, and voice notes reflect regional identities and innovate with dialect and slang. However, algorithmic biases often promote content in dominant languages like Hindi or English, reproducing existing hierarchies. In Azerbaijan, digital media similarly fosters localized language use but tends to reflect the state's monolingual orientation, with limited visibility for minority languages. While India and Azerbaijan differ in linguistic scale and state ideology, both reveal the deep entanglements between language, identity, and power. In India, linguistic variation operates within a framework of constitutional pluralism and social hierarchy, often expressed through hybrid codes and grassroots resistance. In Azerbaijan, linguistic unification is part of a state-driven project of cultural restoration that consolidates national identity but risks suppressing internal diversity. These dynamics set the stage for the next section, which examines how Artificial Intelligence technologies interact with these sociolinguistic configurations and what implications this has for linguistic justice in both contexts. Azerbaijan presents a complex and multifaceted sociolinguistic landscape characterized by historical stratification, state-led language policy, multilingualism, and post-Soviet identity renegotiation. Positioned geopolitically at the nexus of East-

ern Europe and Western Asia, Azerbaijan's linguistic profile has evolved under significant Persian, Turkic, Russian, and, more recently, Western influences. The sociolinguistic configuration is not merely a reflection of communication norms but a politically and ideologically contested domain where national identity, ethnic belonging, and geopolitical alignment intersect.

#### 4.1. Historical Layers and Language Ideologies

Historically, the Azerbaijani language, a branch of the Oghuz subgroup of Turkic languages, was influenced extensively by Persian during the Safavid period and later by Russian due to imperial and Soviet governance. Russian linguistic imperialism, particularly during the Soviet era, established Russian as the *lingua franca* in domains of administration, science, and interethnic communication<sup>[13]</sup>. Post-1991 independence witnessed a revalorization of Azerbaijani (*Azərbaycan dili*), supported by nationalistic discourses that reimagined the language as a vessel of ethnic purity and cultural sovereignty<sup>[14]</sup>. Yet, this postcolonial reorientation was not linear. The sociolinguistic consequences of de-Russification were uneven, particularly in urban versus rural and elite versus peripheral distinctions. Baku, the capital, retained Russian as a prestige code among intellectual elites, technocrats, and bilingual households, underscoring the enduring symbolic capital of Russian<sup>[12]</sup>. Contrarily, in rural or provincial settings, Azerbaijani—predominantly in its North Azerbaijani dialect—was reaffirmed both as a communicative tool and as a badge of national authenticity.

#### 4.2. Language Policy and Planning

Azerbaijan's language policy post-1991 operates under the aegis of *state-driven linguistic purification*. Legislative frameworks such as the 2002 "Law on the State Language" mandated the exclusive use of Azerbaijani in state institutions, education, and mass media. However, despite these overt policies, the covert language practices illustrate a form of diglossic coexistence—where Azerbaijani (High variety) is used in formal registers, and Russian (Low, yet symbolically high) persists in academia, sciences, and the private to Latin (1929), to Cyrillic (1939), and back to Latin post-independence<sup>[15]</sup>.

#### 4.3. Ethnolinguistic Minorities and Linguistic Human Rights

Despite official rhetoric of linguistic homogenization, Azerbaijan remains ethnolinguistically diverse. Lezgins, Talysh, Avars, Kurds, Tats, and Mountain Jews constitute significant minority groups, each with its own linguistic heritage. However, state policy has often marginalized these groups by framing linguistic difference as a threat to national unity<sup>[16]</sup>.

#### 4.4. Multilingualism, Globalization, and Youth Identity

In contemporary Azerbaijan, multilingualism is no longer merely a remnant of imperial legacy but also a product of globalized educational aspirations. English has emerged as a language of opportunity, particularly among youth in urban settings, who often deploy code-switching strategies that juxtapose Azerbaijani, Russian, and English in fluid, indexical ways. Such multilingual practices are less about necessity and more about performing aspirational cosmopolitan identities. This linguistic repertoire, while seemingly emancipatory, also reflects new linguistic inequalities where access to English proficiency becomes stratified along class and regional lines. Emerging scholarship also suggests that sociolinguistic behaviour in Azerbaijan is deeply gendered. Women, particularly in conservative rural contexts, are expected to use more "polished," monolingual Azerbaijani registers, whereas men often enjoy greater linguistic freedom to deploy Russian or hybrid registers, especially in business or bureaucratic domains. These gendered expectations mirror broader social norms and contribute to the symbolic feminization of linguistic purity<sup>[17]</sup>.

### 5. AI and Sociolinguistics

Artificial Intelligence (AI) has transitioned from being a purely computational tool to becoming an agent in sociolinguistic inquiry that mediates human interaction, identity formation, and language politics. Its influence is particularly significant in multilingual contexts like India and linguistically centralized yet ethnolinguistically diverse countries like Azerbaijan. This section presents a research-oriented, data-grounded exploration of how AI shapes and is shaped by

sociolinguistic practices in these two distinct geopolitical and linguistic ecologies. In India, initiatives such as Bhashini and AI4Bharat have deployed AI for language digitization, with specific projects focusing on language modeling, automatic speech recognition (ASR), and corpus development for regional languages. For instance, AI4Bharat's IndicBERT and MuRIL (Multilingual Representations for Indian Languages) are transformer-based language models trained on datasets covering 17 Indian languages. However, an empirical evaluation showed that MuRIL performs with high accuracy for Hindi and Bengali but degrades significantly when applied to under-resourced languages such as Bodo or Santhali, with word error rates exceeding 60%<sup>[18]</sup>. Such disparities are not incidental but tied to sociolinguistic marginalization and corpus bias. Most AI models in India are trained on urban, literate, and standardized language inputs drawn from newspapers, Wikipedia, and broadcast media. These sources fail to capture vernacular varieties, dialects, and hybrid speech practices, such as code-switching, which dominate real-life communication. In regions like Uttar Pradesh and Jharkhand, the sociolinguistic reality includes diglossia between local dialects (Awadhi, Bhojpuri) and Standard Hindi. AI systems trained on standard forms often misclassify dialectal inputs or yield syntactically malformed translations<sup>[19]</sup>. Moreover, studies by Annamalai<sup>[5]</sup> and Sridhar<sup>[20]</sup> have demonstrated that ASR systems show class-based phonetic discrimination, with word error rates 30% higher for speakers from Scheduled Castes and rural regions than for upper-class English-medium speakers. These systems embed latent biases that reflect historical and social hierarchies, thus reinforcing digital linguisticism. A comparative audit of Google ASR revealed a 22% accuracy drop for Indian English spoken with a Tamil accent versus a Delhi-based English speaker. Such outcomes directly affect the usability of AI in education and governance. In contrast, Azerbaijan presents a linguistically centralized scenario, with North Azerbaijani functioning as the official and dominant language since the post-Soviet period. However, the digital corpus for Azerbaijani remains underdeveloped. A study by Ismailzade<sup>[21]</sup> observed that language data from minority groups such as the Talysh, Lezgi, and Avar are virtually absent in NLP datasets, despite accounting for 9% of the population collectively. Unlike India, where civil society and academic consortia have actively participated in corpus building, Azerbaijan's efforts remain primarily state-

driven and focused on language purification. The National Academy of Sciences of Azerbaijan has developed corpora emphasizing standardized orthography and vocabulary, excluding regional speech variations.

Additionally, sociolinguistic monitoring is increasingly visible in Azerbaijan's digital ecosystem. AI's influence is further complicated by the ideologies embedded within technological design. For example, predictive text systems deployed on mobile keyboards in India are trained on large datasets from urban Hindi and Indian English users, automatically correcting non-standardized expressions such as "prepone" or regionally accepted idioms. Such interventions reproduce linguistic hierarchies in a digital context, mirroring what Gikandi<sup>[1]</sup> (p. 74) refers to as the "ideological entrenchment of standard forms as gatekeepers of modernity." Similarly, Azerbaijani autocorrect features standardize inputs to conform to Baku orthographic norms, marginalizing local lexis. Ethical concerns in AI-mediated linguistic analysis also merit attention. In India, AI tools are increasingly used in education, but few have mechanisms for linguistic adaptability. Studies on EdTech platforms revealed that students from rural Tamil Nadu using voice search interfaces faced comprehension errors 40% more frequently than urban users. In Azerbaijan, smart assistants developed by local startups have faced criticism for low performance in recognizing non-standard speech and for potential violations of privacy due to opaque data collection practices. Nonetheless, participatory and inclusive approaches offer alternatives. In India, the People's Linguistic Survey of India<sup>[22]</sup> has been partially integrated into low-resource NLP projects to document endangered languages, such as Nahali and Kurukh. In Gujarat, NGOs are piloting AI systems that support Bhili, a tribal language, for healthcare information dissemination. Azerbaijan's recent collaboration with Kazakhstan and Turkey to build Turkic language datasets has enabled the creation of shared linguistic infrastructures, although minority languages remain excluded. In both countries, AI does not merely record language—it constructs and constrains it. The sociolinguistic stakes are therefore high. As Blodgett et al.<sup>[23]</sup> argue, AI systems must incorporate linguistic justice frameworks to avoid perpetuating structural inequalities. AI models that fail to account for dialectal diversity, speaker variability, and language ideology risk becoming tools of linguistic domination rather than liberation. For India and



Azerbaijan, this means that future AI development must be grounded in nuanced, empirical sociolinguistic research and must prioritize inclusivity in both technological design and language representation.

The differential trajectories of AI's integration into the linguistic ecosystems of India and Azerbaijan expose broader questions about power, identity, and access. Drawing on the theoretical models of variationist sociolinguistics<sup>[3]</sup> and linguistic anthropology, in India, the sociolinguistic landscape is shaped by stratification across caste, region, and educational access. Variationist models help capture how AI amplifies pre-existing linguistic inequalities. For example, variation in vowel shifts among Hindi-English bilinguals from Delhi versus Patna is rendered invisible in ASR training data, reinforcing standard Delhi Hindi as the normative speech form. This aligns with Trudgill's<sup>[24]</sup> observation that technological standardization often consolidates the speech patterns of the dominant sociolect. Conversely, Azerbaijan's post-Soviet linguistic ideology prioritizes language purification, minimizing internal variation to reinforce national cohesion. AI systems trained on curated corpora reflect this ideology by excluding hybrid speech and dialectal deviations. Cavanaugh's<sup>[25]</sup> concept of "iconization" becomes relevant here—the transformation of a particular speech form (i.e., Baku Azerbaijani) into a symbolic marker of national identity. This erases linguistic heterogeneity and simplifies the linguistic repertoire available for technological representation.

## 6. Findings and Discussion

The findings of this study underscore the divergent trajectories and shared challenges that India and Azerbaijan face in the integration of AI into their sociolinguistic ecologies. While both countries are marked by complex linguistic histories and socio-political formations, their respective engagements with AI-driven language technologies reveal contrasting institutional strategies, infrastructural capacities, and sociolinguistic consequences. In India, the scale of linguistic diversity—22 scheduled languages and hundreds of dialects—creates an enormous burden on AI systems to be inclusive and representative. Yet, the evidence reviewed suggests that current AI frameworks tend to privilege dominant languages such as Hindi, Bengali, and Tamil, while

marginalizing under-resourced and tribal languages. This reflects a broader pattern of computational bias rooted in resource availability, corpus accessibility, and elite language ideologies<sup>[26–28]</sup>. The digital infrastructure in India, though expansive, is heavily skewed toward standardized, urban, and literate language forms. As a result, NLP models trained on these corpora exhibit poor performance in dialectal speech recognition, phonetic variation, and code-switching contexts.

AI-driven systems in India appear to exacerbate class and caste-based linguistic disparities. Findings from speech recognition error analyses indicate that rural and lower-caste speakers are systematically misrecognized, suggesting embedded algorithmic biases that mirror historical hierarchies<sup>[29]</sup>. This confirms what sociolinguists like Annamalai<sup>[5]</sup> and Sridhar<sup>[20]</sup> have long argued—that language standardization processes in India are not ideologically neutral but deeply political, with AI systems now extending these hierarchies into digital spaces. In contrast, Azerbaijan presents a more centralized linguistic scenario. The dominance of North Azerbaijani, reinforced by state-led language policies, simplifies the development of standard NLP tools. However, this very standardization has a homogenizing effect, sidelining dialectal and minority language communities. Despite the presence of ethnolinguistic groups such as the Talysh and Lezgi, their exclusion from national corpora and AI applications results in a systematic erasure of their digital presence. This represents a digitally mediated continuation of state ideologies that equate national identity with linguistic uniformity. Furthermore, the study finds that while India's AI ecosystem is shaped by a mix of public and civil society initiatives, Azerbaijan's is predominantly state-controlled. This has implications for corpus design, linguistic inclusivity, and transparency in AI development. Whereas Indian NGOs and academic consortia are involved in documenting endangered languages and building open-source tools, Azerbaijani projects are more conservative, emphasizing standard forms and offering little space for bottom-up linguistic participation.

A crucial comparative insight lies in linguistic agency within digital platforms. In India, AI-driven language tools are increasingly used in educational and e-governance systems. Yet, their functionality remains limited for non-standard users, thus excluding significant populations from effective digital participation. In Azerbaijan, algorithmic standardization shapes online discourse to conform to domi-

nant linguistic norms, particularly on monitored platforms. This has led to forms of digital self-censorship, where dialect speakers adjust their language use to conform to intelligibility expectations or avoid algorithmic invisibility. Both cases highlight a deeper tension between technological rationality and sociolinguistic complexity. AI systems, by their very architecture, strive for uniformity, generalizability, and predictability—qualities at odds with the variability, contextuality, and heterogeneity that characterize real-life language use. The findings affirm that sociolinguistics must be integral to AI design, not an afterthought. As Blodgett et al.<sup>[23]</sup> have argued, language technologies must incorporate frameworks of linguistic justice to avoid reproducing marginalization through code. Another important theme emerging from the comparison is the geopolitical dimension of language AI. In India, the ambition to develop AI for Bharatiya languages is intertwined with nationalist narratives of self-reliance and linguistic sovereignty. In Azerbaijan, efforts to develop Turkic language models align with regional aspirations for cultural unity and consolidation. While these initiatives may foster linguistic pride and autonomy, they also risk reinforcing monolingual ideologies that obscure internal diversity.

Finally, the findings suggest that inclusive and participatory models of AI development—such as those involving community-led corpus building, open-access platforms, and dialect-sensitive NLP—are not just ethical imperatives but technical necessities. Projects like the People’s Linguistic Survey of India and Turkic language collaborations, though limited, provide models for more equitable language AI development. These initiatives highlight the possibility of designing technologies that do not merely reflect dominant language ideologies but actively reconfigure them in favour of pluralism and justice<sup>[30]</sup>. In sum, AI’s intersection with sociolinguistics in India and Azerbaijan presents a complex picture of opportunity, exclusion, and ideological mediation. If AI is to contribute meaningfully to linguistic equity, it must be grounded in empirical sociolinguistic research and designed with a conscious commitment to diversity, representation, and agency.

## 7. Conclusion

This study sought to unpack the complex interplay between sociolinguistics and artificial intelligence within two

distinct national contexts—India and Azerbaijan. By focusing on how AI technologies mediate language hierarchies, dialect representation, and sociolinguistic equity, the research provides critical insights into the socio-technological fabric that shapes modern linguistic practices. Through an extensive comparative lens, it becomes evident that AI does not function in a vacuum but is deeply entrenched in the linguistic ideologies and infrastructural disparities of each nation. In India, the sociolinguistic dynamics are influenced by extreme linguistic diversity, digital divides, and historical language politics. AI tools, including ASR systems, NLP models, and machine translation engines, often perpetuate these inequalities by prioritizing dominant languages (such as Hindi and English) over marginalized regional languages. Despite government-led initiatives such as Bhashini and AI4Bharat, a significant portion of the vernacular population remains underserved due to the lack of robust, inclusive linguistic corpora. Conversely, Azerbaijan exhibits challenges of a different order—while it is a predominantly monolingual nation in terms of official language use, Azerbaijani’s underrepresentation in global AI tools and the dominance of Russian and English in digital infrastructure pose unique problems of digital marginalization.

This paper also illuminates the sociolinguistic risks posed by AI, such as algorithmic linguistic standardization, dialect suppression, and the erasure of oral cultures. These issues are not merely technical but carry implications for cultural memory, political representation, and linguistic human rights. In both India and Azerbaijan, the need for participatory design—where local communities contribute to language data development—emerges as a key solution to these dilemmas. Equally crucial is the establishment of linguistic impact assessments for all major AI deployments. The findings affirm the hypothesis that AI, unless explicitly regulated and inclusively designed, often amplifies existing sociolinguistic hierarchies rather than neutralizing them. It further substantiates the claim that linguistic justice in the digital age is a critical socio-political challenge that demands transdisciplinary engagement—from sociolinguists, technologists, ethicists, and policymakers alike. Future research should expand to include other linguistically rich but underrepresented regions such as Central Asia, Africa, and the Caribbean to better understand the global politics of language and AI. Moreover, longitudinal studies are needed to trace the evolving relation-

ship between AI development and language ideologies over time. Methodologies should incorporate a blend of computational linguistics, ethnography, and critical discourse analysis to offer both macro-structural and micro-cultural insights. In conclusion, this paper underscores that the future of sociolinguistics cannot be divorced from its technological mediators. As AI continues to define the contours of communication, access, and identity, the demand for ethical, inclusive, and critically informed AI design becomes ever more urgent.

## Author Contributions

Conceptualization, S.I.S., G.M.M., and L.N.N.; methodology, J.G.A. and R.S.; software, R.S.; validation, S.I.S., G.M.M., and R.S.; resources, L.N.N. and J.G.A.; data curation, D.A.; writing—original draft preparation, D.A.; writing—review and editing, G.M.M., D.A., and F.S.S.; visualization, L.N.N. and N.Y.Y.; supervision, S.I.S. and G.M.M.; project administration, G.M.M.; funding acquisition, N.Y.Y. All authors have read and agreed to the published version of the manuscript.

## Funding

This work received no external funding.

## Institutional Review Board Statement

Not applicable.

## Informed Consent Statement

Not applicable.

## Data Availability Statement

Not applicable.

## Conflicts of Interest

The authors declare that there is no conflict of interest.

## References

- [1] Gikandi, S., 2018. *Slavery and the Culture of Taste*. Princeton University Press: Princeton, NJ, USA.
- [2] Schieffelin, B.B., Woolard, K.A., Kroskrity, P.V., 1998. *Language Ideologies: Practice and Theory*. Oxford University Press: New York, NY, USA.
- [3] Gordon, M.J., 2017. William Labov. *Oxford Research Encyclopedia of Linguistics*. Oxford University Press: Oxford, UK. DOI: <https://doi.org/10.1093/acrefore/9780199384655.013.364>
- [4] Hymes, D. (ed.), 1974. *Foundations in Sociolinguistics: An Ethnographic Approach*, 1st ed. Routledge: London, UK. DOI: <https://doi.org/10.4324/9781315888835>
- [5] Annamalai, E., 2001. *Managing Multilingualism in India: Political and Linguistic Manifestations*. Sage Publications: New Delhi, India.
- [6] Mohanty, A.K., 2006. Multilingualism of the Unequals and Predicaments of Education in India: Mother Tongue or Other Tongue. In: García, O., Skutnabb-Kangas, T., Guzmán, M.E.T. (eds.). *Imagining Multilingual Schools: Language in Education and Glocalization*. Multilingual Matters: Clevedon, UK. pp. 262–283.
- [7] Eisenstein, J., 2013. What to Do About Bad Language on the Internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, 9–14 June 2013; pp. 359–369.
- [8] Hovy, D., Spruit, S.L., 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Berlin, Germany, 7–12 August 2016; pp. 591–598.
- [9] McEnery, T., Hardie, A., 2012. *Corpus Linguistics*. Cambridge University Press: Cambridge, NY, USA. DOI: <https://doi.org/10.1017/cbo9780511981395>
- [10] Kumar, M., Chaturvedi, K.K., Sharma, A., et al., 2023. An Algorithm for Automatic Text Annotation for Named Entity Recognition Using Spacy Framework. *Research Square preprint*. rs.3.rs-2930333/v1. DOI: <https://doi.org/10.21203/rs.3.rs-2930333/v1>
- [11] Mustafayeva, L.A., Akbarova, S.A., Gurbanova, G.K., et al., 2025. Code-Switching in Multilingual Societies: Significance, Patterns, Functions, and Sociolinguistic Implications. *Forum for Linguistic Studies*. 7(7), 194–207. DOI: <https://doi.org/10.30564/fls.v7i7.10294>
- [12] Pavlenko, A., 2008. Multilingualism in Post-Soviet Countries: Language Revival, Language Removal, and Sociolinguistic Theory. *International Journal of Bilingual Education and Bilingualism*. 11(3–4), 275–314. DOI: <https://doi.org/10.1080/13670050802271517>
- [13] Landau, J.M., Kellner-Heinkele, B., 2021. *Politics of Language in the Ex-Soviet Muslim States: Azerbaijan, Uzbekistan, Kazakhstan, Kyrgyzstan, Turkmenistan, Tajikistan*. University of Michigan Press: Ann Arbor, MI, USA.
- [14] Suleymanova, N., 2025. *The Impact of Language*

- Rules on Writing Skills and Integrative Teaching. *International Scientific Journal*. 168–172. DOI: <https://doi.org/10.36719/2663-4619/113/168-172> (in Azerbaijani)
- [15] Oqlu, K.P.F., 2021. Transition to Latin Alphabet (20s of the Twentieth Century) Modern Reality Is a Result of Historical Events. *English Linguistics Research*. 10(2), 38–42. DOI: <https://doi.org/10.5430/elr.v10n2p38>
- [16] Wheatley, M.J., 2006. *Leadership and the New Science: Discovering Order in a Chaotic World*, 3rd ed. Berrett-Koehler Publishers, Inc.: San Francisco, CA, USA.
- [17] Formanowicz, M., Bedynska, S., Cislak, A., et al., 2013. Side Effects of Gender-Fair Language: How Feminine Job Titles Influence the Evaluation of Female Applicants. *European Journal of Social Psychology*. 43(1), 62–71. DOI: <https://doi.org/10.1002/ejsp.1924>
- [18] Shanbhag, A., Jadhav, S., Thakurdesai, A., et al., 2024. Non-Contextual BERT or FastText? A Comparative Analysis. *arXiv preprint*. arXiv:2411.17661. DOI: <https://doi.org/10.48550/arXiv.2411.17661>
- [19] Sankaranarayanan, A., Johnson, K., Mammen, S.J., et al., 2021. Disordered Eating Among People with Schizophrenia Spectrum Disorders: A Systematic Review. *Nutrients*. 13(11), 3820. DOI: <https://doi.org/10.3390/nu13113820>
- [20] Sridhar, K.K., 1989. *Sociolinguistic Issues in India*. Manohar Publishers: New Delhi, India.
- [21] Mikail, E.H., Çora, H., 2024. Exploring the Turkic Identity of Azerbaijan's Ethnic Groups: A Comprehensive Analysis. *Open Journal of Applied Sciences*. 14, 3077–3099. DOI: <https://doi.org/10.4236/ojapps.2024.1411203>
- [22] Devy, G.N., Davis, G.V., Chakravarty, K.K. (eds.), 2013. *Narrating Nomadism: Tales of Recovery and Resistance*, 1st ed. Routledge India: London, UK. DOI: <https://doi.org/10.4324/9780367818449>
- [23] Blodgett, S.L., Barocas, S., Daumé, H., et al., 2020. Language (Technology) Is Power: A Critical Survey of “Bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 5–10 July 2020; pp. 5454–5476.
- [24] Trudgill, P., 2000. *Sociolinguistics: An Introduction to Language and Society*, 4th ed. Penguin: London, UK.
- [25] Cavanaugh, J., 2020. Language Ideology Revisited. *International Journal of the Sociology of Language*. 2020(263), 51–57. DOI: <https://doi.org/10.1515/ijsl-2020-2082>
- [26] Thulasimani, T., Saha, L., Poreddy, B., et al., 2024. Advanced Analysis of Social and Psychological Factors in Higher Education Institutions Using DNN and LSTM. In *Proceedings of the 2024 International Conference on Data Science and Network Security (ICDSNS)*, Tiptur, India, 26–27 July 2024; pp. 1–6. DOI: <https://doi.org/10.1109/ICDSNS62112.2024.10691260>
- [27] Karimli, V.M., Khudaverdiyeva, T.S., Huseynova, F., et al., 2025. The Role of Mobile Computing in Adaptive Testing for English Language Learners: Personalizing Assessment to Improve Outcomes. *Forum for Linguistic Studies*. 7(6), 149–160. DOI: <https://doi.org/10.30564/fls.v7i6.9663>
- [28] Dongre, S., Krishnaveni, P., Maharram, V.K., et al., 2025. Graph Theory for Enhanced Feedback and Student Performance Prediction in Higher Education Using a GAT-BiLSTM-Based RL Model. In *Proceedings of the 2025 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE)*, Shivamogga, India, 24–25 April 2025; pp. 1–7. DOI: <https://doi.org/10.1109/AMATHE65477.2025.11081268>
- [29] Ismayil, Z., 2024. Development Points of the Derivation Process Related to Nakhchivan Dialects and Accents (Based on Written and Oral Literary and Artistic Examples). *Forum for Linguistic Studies*. 6(2), 1192. DOI: <https://doi.org/10.59400/fls.v6i2.1192>
- [30] Karimli, V., Ahmadzadeh, T., 2025. Analysis of the Impact of Cultural Diversity on University Branding. *Danish Scientific Journal*. 1, 68–71.