New Academia: An International Journal of English Language, Literature and Literary Theory

ONLINE ISSN 2347-2073 **Vol. 15, Special Issue 1, January 2026**

(Peer Reviewed and Refereed Journal)

DATA, POWER, AND THE FUTURE OF LANGUAGE SURVIVAL THROUGH ARTIFICIAL INTELLIGENCE AS LINGUISTIC INFRASTRUCTURE

Bilure Suhasini Ramchandra
Research Scholar,
School of Languages & Literature,
Punyashlok Ahilyadevi HolkarSolapur University, Solapur.
biluresuhasini@gmail.com

Abstract

Artificial Intelligence (AI) today transcends its initial definition as a set of technical instruments to become a pervasive socio-technological infrastructure that mediates communication, knowledge production, governance, economy, and representation on a global scale. Language is the mediating factor and plays the role of the primary input and output medium of AI systems. In this paper, the author proposes to develop an extensive and critical argument that AI needs to be understood and studied as linguistic infrastructure because the latter is the foundation that determines the linguistic visibility and survival within the digital world. In this regard, the paper relies on trans-disciplinary scholarly work within the domains of sociolinguistics, critical data studies, political economy, science and technology studies, and AI ethics and examines the role of data regimes and the corresponding structures and systems within the platform. The author claims within the context of this paper that the contemporary AI systems and structures exhibit a hierarchical bias towards "a very short list of high-resource languages with strong global presence and the rest to the significantly marginalized and endangered language systems." Additionally, this paper proposes a methodological approach to the study with a qualitative research approach wherein the author conducts critical discourse analysis and theoretical work employing the structures and frameworks of infrastructural theory. In doing so, the author relies on the use of analytical tables and structures to support the theoretical argument made within the paper. In fact, the entire argument depends on the development of analytical tables and structures to reach the conclusion. Ultimately, the study culminates on its conclusion to propose a long-term vision and policy-orientation to reimagine AI as an inclusive linguistic infrastructure that supports linguistic diversity and epistemic justice within the algorithmic world.

Keywords: Artificial Intelligence; Linguistic Infrastructure; Language Survival; Data Power; Algorithmic Governance; Linguistic Inequality; Digital Divide; AI Ethics

Introduction:

The first two decades of the twenty-first century will be remembered for the proliferation of Artificial Intelligence in all walks of social life. AI-based systems now shape all language use in social media through machine translation, speech recognition technology, automated transcription, predictive text, content review, recommendation systems, and conversational AI. There was a time when such AI systems were on the fringes of communication. But today, they are integral to the use, value, and transmission of language.

Language has traditionally been and continues to be an important medium to and through which knowledge, identity, governance, and culture are encoded and maintained in societies. In the modern digital age, the viability and survival of languages are fast becoming dependent on their representation in technology and technological systems, especially where languages are compatible with artificial intelligence systems to have access to the spheres of power like education, governance, healthcare, and the digital economy.

AI also represents a new and qualitatively different method of selecting language(s) than previous language loss due to colonialism, nation-state language policy, industrialization, and globalization (Although all of these methods continue to exist). AI's selection of language(s) occurs through algorithmically-driven methods operating within a larger context than ever before; AI normalises linguistic forms at such scale(s) as well as provides indirect influence over choices of language(s) via the interface used (i.e., an interface that looks "neutral" or "inevitable"). Therefore, the hierarchies of language(s) are increasingly reproduced/recreated through the technological medium of AI and not through explicit political coercion.

While there has been a significant increase in academic interest surrounding the intersection of

AI and language, most of this interest is still focused on AI's use as a tool for documenting/revitalizing endangered language(s), therefore having a technological instrumentalist focus or perspective, rather than an understanding of how AI systems produce/exist from and consolidate existing power structures within the global political economy through the use of AI. I argue for a complete paradigmatic shift whereby AI will be viewed as a language infrastructure, with this perspective focusing on the political economy of AI, the inherent power relations that exist within AI through the data and algorithms used to operate AI, and the political structures/governance mechanisms that will ultimately determine which language(s) will be sustained by AI.

Theoretical and Conceptual Framework: AI as Linguistic Infrastructure:

The study of infrastructure has highlighted that the foundational systems for organizing social life are structured by their operation below the realm of everyday visibility, but in the sense that; for example, by virtue of their existence, the roads, electricity grid or distribution of electricity, water & sewer, and communication, which are referred to as Communication, how they allow us to access, move and be provided with as a member of our community. Therefore, by analogy, AI is to Linguistics, as Infrastructure is to Society.

To consider AI as a form of linguistic infrastructure requires three distinct but interrelated levels of analysis. First, a Data Infrastructure refers to the sociotechnical process of producing, gathering, digitizing, standardizing, storing, owning, and governing linguistic data. Political decisions as to what languages will be digitized, whose speech will be recorded, and what constitutes valid linguistic data are political in nature and reflect the inequality of access and opportunity due to the power dynamics of the Social Structure.

Second, Algorithmic Infrastructure includes the computational models, learning architectures, and optimization and evaluation metrics used to convert linguistic data into predictions and outputs. Algorithms encode the assumptions made about language, efficient functioning, and normative structures, often privileging the Standardised Varieties of language, as well as the Dominant Languages. These assumptions will determine which languages will be considered computationally tractable and will therefore be generated, and which will be considered too complex or economically unsustainable; therefore, Algorithmic Infrastructure represents the total collapse of Linguistic Infrastructure.

Corporate and institutional frameworks under which Artificial Intelligence (AI) systems are established, developed and operated, comprise what is referred to as the "platform infrastructure." The entities involved (search engines, social media platforms, voice assistant applications and digital service providers) hold a "gatekeeper" role, where they ultimately decide what languages have the greatest opportunity for support, promotion, monetisation or are systematically "marginalised" by blocking access to these resources. Thus, it becomes apparent that platform policies will play an important role in determining the level of linguistic visibility and sustainability afforded to various languages through time.

On the surface, several levels of consideration underlie linguistic inclusion in AI. In particular, infrastructural access is the main determinant of whether particular languages will be included in the development and ongoing operation of an AI system, rather than just the individual vitality of the speakers. As a result, it is consistent with the broader theoretical perspective on infrastructure and power, which asserts that access to the basic elements that comprise the infrastructure of a society are fundamental to the creation of social inequalities.

Given the rise of AI technology, the cultural future of civilizations will be heavily influenced by the institutional framework that provides access to linguistic resources.
Page 3- Review of Previous Literature

Review of Literature:

There is no contention among Peer-Reviewed Research that language loss occurs as the result of Sociopolitical Factors leading to Language Loss rather than through Natural Processes. Crystal argues (2000) that Languages will die or lose their vitality when they lose access to the Sources of Power, Prestige, and Modernity. Within Sociolinguistic Research, later studies demonstrated that Languages which are able to Support and Sustain themselves through: Technological Adaptability; Economic Relevance; and Institutional Support will be more sustainable than Languages that do not.

Technological and Linguistically Related Studies:

The relationship between Technology and Language has been the focus of Continuing Debate in the Scientific Community for decades; however, with the inception of the internet, many Early Scholars believed that the internet would provide Minority Language Speakers with an opportunity to have their Languages Supported through the use of Digital Platforms. However, Subsequent Studies demonstrated that there was a Distinct Pattern of Linguistic Concentration via the Internet, and that the Dominant Languages (Especially those spoken by Majorities) are Disproportionately Represented Online (Warschauer, 2003).

Further complicating the matter are Current Studies Related to Ethical Data Use and AI; for example, Boyd and Crawford (2012) found that Data are influenced by a community's Power Structure, influencing what Data will be Collected, Stored and analyzed. Noble (2018) demonstrated how Algorithmic Systems reinforce the Racial and Linguistic Hierarchies of a given community whilst also creating an illusion of Objectivity. Recent Studies Related to Large Language Models were conducted to demonstrate that Language Model Training Data is Overwhelmingly Represented by and Large Language Models are Most Viewed by Spanish and a Small Number of Other Highly Judged/Global Languages (Bender et al, 2021).

The previous paragraphs have revealed two main insights to this study; first, that current literature has tended to portray AI as a separate technological object whereas this research is suggesting that AI should be viewed as part of an Infrastructure for Language and Culture. Secondly, current literature typically considers the infrastructure of Language and Culture as 'critical', yet, the findings from this research highlight the need to consider how different forms of infrastructures have been shaped by different modes of technology development. The goal of this research is to bring together aspects of Sociolinguistics,

Critical AI Studies and a Theory of Infrastructure into a unified framework, focusing specifically on Language Survival as an Infrastructure and Political Issue.

Objectives and Research Question:

Objectives:

1. To understand Artificial Intelligence as an Infrastructure for Language.
2. To explore the relationship between Data Regime/Algorithmic Power and Language Inequality.
3. To examine the relationship between Platform Governance and Language Survival.
4. To create a framework with multiple levels of Linguistic Sustainability for Artificial Intelligence.

Research Questions:

1. What are the methods used by these AI data practices to privilege certain languages over others?
2. How do the methods of Algorithmic Governance provide legitimacy to linguistic hierarchies?
3. What Ethical, Policy and Infrastructure models must be established to ensure that Languages survive in the Era of AI?

Research Methodologies:

This research utilizes a mixed method, policy analytical method that utilizes critical qualitative analysis, comparative policy analysis and structured secondary data analysis. Using this method allows the researcher to triangulate the data and provide more analytical robustness to the research and meet the empirical/conceptual expectations of Scopus indexed interdisciplinary journals.

Qualitative critical analytical approaches are first used to conceptualize AI as a language infrastructure. This analysis uses critical discourse analysis, the political economy of technology and theory and infrastructural theory to investigate peer-reviewed articles, narratives of AI and governance frameworks in order to understand how power structures, data regimes, and algorithmic norms can influence language inclusion and exclusion.

The research also contains a comparative policy analytical method to examine the numerous major international policy documents on the subject such as UNESCO's Recommendation on the Ethics of AI, OECD AI policy reports, draft regulatory proposals such as the proposed EU AI Act etc., and compares these with national initiatives, with a particular emphasis on India's multilingual digital governance system. This comparison

provides insight into how linguistic diversity is either implemented or ignored at different levels of policy.

Next, an organized secondary data analysis is performed. A thorough investigation of Large Scale Linguistic Datasets, AI Training Corpora, Documented Low-Resource Language Projects, and Published Performance Evaluations of NLP Systems are conducted as a means of establishing empirical context for the theoretical assertions made in this paper. Although this research does not produce a primary corpus, this layer provides additional methodological triangulation, thereby adding to the analytical credibility of this paper.

This combination of methods places the research as a Mixed Methods Policy Analysis Paper and establishes a firm foundation in theory and empirical relevance; thus, it is eligible for consideration for publication in high-impact, peer-reviewed journals indexed in UGC-CARE&Scopus.

Data are inherently driven by the research across all languages within the data ecosystem, with a clear difference of availability to in-state or near-state (high) resource languages versus other linguistic(s) represented by all other (low) resource linguistic data. Data availability allows high-resource languages to be present historically (due to being better supported by historical power, state, industry, institutions and/or economy) Digital Availability of standardised orthography allowing Institutional data to be represented in digital corpus' preventing these linguistic from being visible within AI Systems. As a result, low-resource language data are disproportionately represented within AI.

The result of this consequence is Algorithm Bias that creates disparity between performances of AI systems for high and low resource languages. Data has become a vehicle for Linguistic Capital consistent with Bourdieu's (1991) theory of Symbolic Power disadvantages of low resource languages unable to compete for access to High Resource Language gains.

Table 1 illustrates the difference in linguistic representation in AI Ecosystems by Dimension:

Dimension	High Resource Languages	Low Resource Languages
Data Availability	Extensive, standardized corpora	Spates, fragmented data
Institutional Support	Strong	Minimal or absent
AI Applications	Translation, speech, NLP, catboats	Limited or nonexistent

Algorithmic bias and unequal performance of systems are two direct results of these discrepancies. Just as Bourdieu (1991) defined language as a 'social capital' that can be used

for 'symbolic power', therefore, data serves as a form of linguistic capital. As a result, languages with an abundance of data enjoy greater symbolic and material advantage than languages with a scarcity of data, leaving the latter languages unable to enter the system.

Algorithmic Power and Platform Governance:

Technology plays a huge role in shaping how we speak through ranking, predicting languages, recommending languages and moderating our language. As a result of all the interactions we have with AI systems on a daily basis, individuals are encouraged to speak in languages that support AI. Technology companies are placing greater emphasis on large-scale businesses, efficient practices, and making money than they are on supporting linguistic diversity within their business models.

As Noble (2018) states, in addition to the issue of supporting the network of dominant language users, technology also makes it very difficult for the audience to see the power dynamics and, because they cannot examine the power dynamics, the audience is more likely to accept the exclusion of that language as "natural" or part of the landscape of languages. This creates a feedback mechanism in which dominant languages are continually being given additional value and visibility, while marginalized languages go into extended periods of decline. Furthermore, without appropriate regulatory oversight, technology companies will continue to develop AI technologies and promote homogeneity of languages.

Consequences for Language Survival:

The impact of AI driven language technology on threatening the existence and continuity of lesser known, endangered languages is both huge and complicated. Because AI technology includes domestic, educational and commercial interactions, languages without AI technology will become extinct and disappear as modern-day generations continue to use, communicate with and engage with daily life through AI technology.

From the view of social-linguistics, exclusion from AI further reduces the size and scope of a language's domains, which is one of the primary indicators of the endangerment of a language (Crystal, 2000). If a language has no capacity for findings, artificial capabilities on the internet and generating information, its value in society decreases further, and thus more youth will begin to shift out of their cultural and linguistic backgrounds. Thus, if new languages are able to coexist with AI technology, they will gain a new status, new areas in their domains of use and strength in transferring those domains across generations.

The Importance of India as a Case Study in the Relationship between Language, Artificial Intelligence, and Digital Governance:

India is an important case study for exploring the relationship between artificial intelligence (AI) and language survival, as one of the most linguistically-diverse countries in the world, with hundreds of languages spanning a large spectrum of political, economic, technological, and other forms of inequality in terms of hierarchy. As the nature of AI-related digital technologies continues to evolve, India has also begun to comprehend that language and multilingualism are critical to achieving digital inclusion, as is reflected in the adoption of its National Education Policy (NEP) 2020, establishment of the Bashing (National Language Translation Mission) initiative and ongoing efforts under the Digital India Program.

Through the introduction of NEP 2020, India has made an official commitment to a mother-tongue, multilingual educational approach for children, which recognizes that the ability for individuals to use their languages must be supported through institutional and technological resources. The Bhashini initiative represents India's first foray into developing AI-based translation and speech technology to be used in Indian languages, and as such, places AI language technology as part of the framework of digital infrastructure within India. However, this analysis also shows that, while such initiatives are focused on constitutionally-recognized and commercially-viable languages, they, in effect, may create a new layer of hierarchy between different languages in India.

This case exemplifies a larger systemic issue; without ongoing funding of grassroots generated data, community-centric participation, and open-access AI systems, government-sponsored multilingual artificial intelligence initiatives will continue to prop up existing dominant- language ecosystems. Therefore, India demonstrates both the extent of opportunity represented by policy-based efforts to support language inclusion within AI platforms, as well as the limitations imposed by these efforts. Because of this, India serves as both an opportunity and limitation for global debates regarding the preservation of endangered languages.

Therefore, language preservation in the age of AI is more than just a cultural issue: it is an issue of infrastructural justice.

Multi-Level Framework for Linguistically Sustainable AI:

In this article, I propose a multi-level framework for addressing infrastructural linguistic inequality. The graded framework consists of an intersection of documentation, technological inclusion, platform integration, and long-term governance. The framework can be scaled to apply to both national and global contexts.

Table 2: Framework for AI-based Linguistic Sustainability:

Level	Focus	Strategic Intervention	Stakeholders
I	Longitudinal Documentation	Development of Digitized, Corpus Development, Standardization of Orthography	Linguist, Communities, Universities
II	Inclusion	Development of Natural Language Processing Tools, Automatic Speech Recognition, Machine Translation, Automatic Speech Recognition and Translation (NLP, ASR, & MT) for Less Resourceful Languages.	AI researchers/Startups
III	Integration of Platform	Deployment & Service Integration	Technology Platforms, Government
IV	Sustainability Practice	Ethical Governance, Funding, and Policy Coordination	POLICYMAKERS / UNESCO / CIVIL SOCIETY

This multi-level graded framework posits that linguistic sustainability cannot occur through isolated interventions, but will require coordination across all four levels (as opposed to only one level). For example, while documentation of marginalized languages may provide an avenue for preservation of data, the failure to integrate documented language data onto a digital platform may result in continued marginalization of the archived language. Similarly, if technology inclusion occurs without meaningful governance or accountability, extractive data practices will likely continue. Ultimately, Sustainable Outcomes can only be achieved through Coordination across All Four Levels of the Framework.

The tenets of the Digital Age have caused scholars to consider Language as a form of Linguistic Infrastructure, replacing previously held notions of how Languages will continue to exist into the Digital Age (i.e., Linguistic Inequality). Rather than being an incidental result

of technological innovation, Linguistic Inequality is created through structurally produced Data Regimes (i.e., the production of data; the collection and analysis of data, and the collection of vocabulary items from Digital and other sources) and through Algorithmic Normalization, and Platform Governance (Gatekeeping). In this light, when Data is categorised as capital; Algorithms are the agents that Normalise linguistic data; and Platforms are the Gatekeepers of linguistic legitimacy, therefore, the structure of the Digital Age creates Linguistic and Sociocultural Dominance within a community.

This model reflects Pierre Bourdieu's (1991) view that, in today's world, it is possible to continue reproducing power through Institutional Structures rather than through overt forms of force or coercion. Furthermore, it is through the development and use of Technical Standards and Optimization Criteria that Dominance is reproduced; and, Interface Design, specifically, can facilitate the reproduction of Dominance through the Digital Age. The discussion by Safiya Noble (2018) regarding the discriminatory consequences of Algorithms has also highlighted how the systems that seem neutral can also promote patterns of Inequality (i.e., racism).

Nevertheless, the debate points out the constructive role of AI, when based on ethics, in the transformation of the linguistic landscape. UNESCO guidelines on AI and linguistic diversity underpin the value of integrating linguistic rights in the development process. AI architecture focused on the support of multilingual data creation, community engagement, and open language resources contributes to the enhancement rather than the degradation of linguistic vitality.

Conclusion:

This paper has shown that a decisive role in the future of linguistic diversity is played by Artificial Intelligence. The future viability of languages in the AI era not only hinges on the resilience of communities but also on their construction in data and algorithmic systems as infrastructural features. Through its understanding of the concept of AI as linguistic infrastructure, it has been shown that the future of linguistic diversity is not predetermined but designed in certain ways.

"A future with enough linguistic diversity must be intentionally structured with infrastructure changes and data governance policies. Otherwise, AI could fuel the loss of language with its reinforcement of the status quo. Yet with an [.] AI infrastructure vision of inclusivity, there is the possibility of linguistic and epistemic justice and sustainability." The future of language survival is intertwined with the future of AI governance.

References:

Bender, E. M., Gebru, T., McMillan-Major, A., &Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 610–623.

Bird, S. (2020). Decolonising speech and language technology. Proceedings of the 28th International Conference on Computational Linguistics, 3504–3519.

Bourdieu, P. (1991). Language and symbolic power. Harvard University Press.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. Information, Communication & Society, 15(5), 662–679.

Couldry, N., &Mejias, U. A. (2019). The costs of connection: How data is colonizing human life and appropriating it for capitalism. Stanford University Press.

Crystal, D. (2000). Language death. Cambridge University Press.

Floridi, L. (2019). Translating principles into practices of digital ethics. Philosophy & Technology, 32(3), 411–431.

Joshi, P. et al. (2020). The state and fate of linguistic diversity in NLP. Proceedings of ACL, 6282–6293.

Kornai, A. (2013). Digital language death. PLoS ONE, 8(10), e77056.

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. New York University Press.

OECD. (2022). AI and the future of work and language. OECD Publishing.

UNESCO. (2021). Recommendation on the ethics of artificial intelligence. UNESCO Publishing.

UNESCO. (2023). World report on languages. UNESCO Publishing.

vanDijk, J. (2020). The digital divide. Polity Press.

Warschauer, M. (2003). Technology and social inclusion: Rethinking the digital divide. MIT Press.

Zuboff, S. (2019). The age of surveillance capitalism. PublicAffairs.

Government of India. (2020). National Education Policy 2020. Ministry of Education.

Government of India. (2022). Bhashini: National Language Translation Mission. Ministry of Electronics and Information Technology