Journal of Universal Language 27-1. March 2026, 1-25
DOI 10.22425/jul.2026.27.1.1
eISSN 2508-5344

When Linguistic Hierarchy Becomes Infrastructure: AI, Language, and the Reconfiguration of Linguistic Imperialism

Silo Chin
Sejong University, Korea

Abstract

Recent advances in AI translation and LLMs have generated expectations that computational systems may broaden access to global knowledge. Multilingual pretrained models process and generate text across dozens of languages, expanding linguistic coverage in contemporary NLP systems and potentially reducing

Silo Chin
Visiting Professor, Department of English Data Convergence, Division of International Studies, Sejong University, Korea
Email: silochin@sejong.ac.kr

Received January 16, 2026; Revised February 20, 2026; Accepted March 16, 2026

barriers that have historically favored dominant languages in education, publishing, and global communication. Despite these developments, expanded coverage does not entail structural equivalence among languages. Research in multilingual NLP documents persistent asymmetries in linguistic representation and performance. Studies show that datasets, benchmarks, and research attention remain concentrated in a small group of high-resource languages (Joshi et al.; Ruder, Vulić, and Søgaard), and empirical analyses suggest that multilingual LLMs often demonstrate stronger performance when prompted in dominant languages such as English (Gupta et al.; Rohera et al.). Survey research similarly observes that current systems remain largely English-centric due to the distribution of training data and evaluation benchmarks (Bird; Qin). These developments extend longstanding debates on linguistic inequality and linguistic justice within sociolinguistics and language policy research (Bourdieu; May; Phillipson; Piller). Rather than eliminating linguistic hierarchy, this paper argues that AI-mediated communication reconfigures it by relocating historical asymmetries within computational infrastructures that shape data production, model training, and evaluation. Through the concepts of algorithmic linguistic privilege and compensatory linguistic labor, the analysis interprets multilingual AI as a site where hierarchy emerges not through explicit policy but through infrastructural conditioning. The study is primarily conceptual and develops a theoretical framework for interpreting emerging empirical findings on linguistic asymmetries in AI-mediated communication.

Keywords: algorithmic linguistic privilege, compensatory linguistic labor, linguistic imperialism, multilingual AI, language technology, linguistic inequality, computational infrastructure

1. Introduction

Recent advances in artificial intelligence (AI) translation and large language models (LLMs) have generated expectations that computational systems may broaden access to global knowledge and

information. Multilingual pretrained models such as mT5 (Xue et al. 2021) and BLOOM (Scao et al. 2022) process and generate text across dozens of languages, expanding the linguistic coverage of contemporary natural language processing (NLP) systems and potentially reducing barriers that have historically favored dominant languages in education, publishing, and global communication.

This interpretation captures an important aspect of recent developments. Multilingual coverage has expanded, and translation quality has measurably improved. Yet expanded coverage does not entail structural equivalence among languages. Research in multilingual NLP documents persistent asymmetries in linguistic representation and performance. Joshi et al. (2020) demonstrate the uneven distribution of linguistic resources across NLP research, highlighting the concentration of datasets and research attention in a small group of high-resource languages. Surveys of cross-lingual modeling likewise note that multilingual representation learning is shaped by the availability and structure of cross-lingual corpora (Ruder, Vulić, and Søgaard 2019).

Recent empirical analyses suggest that multilingual LLMs often exhibit higher factual accuracy and lower hallucination rates when prompted in dominant languages such as English (e.g., Rohera et al. 2025). Other studies likewise report stronger task performance in English across multilingual evaluations, reflecting disparities in training data availability (Gupta et al. 2025). Survey research on multilingual language models similarly observes that current systems remain predominantly English-centric due to the concentration of training data and benchmarks in a small set of high-resource languages (Qin et al. 2025).

These asymmetries invite a broader theoretical question: Does AI translation eliminate linguistic hierarchy, or does it reorganize it under new conditions of mediation? This paper argues for the latter. Existing

discussions of language and AI largely focus on representational bias in model outputs. This paper instead examines how linguistic hierarchy becomes embedded within the infrastructures that organize data production, model training, and evaluation.

Rather than dissolving linguistic imperial relations, AI-mediated communication reconfigures them by relocating historical asymmetries within computational infrastructures that shape language processing and evaluation. Through the concepts of algorithmic linguistic privilege and compensatory linguistic labor, the analysis reframes multilingual AI as a new site of linguistic mediation in which hierarchy is not explicitly mandated but infrastructurally conditioned.

By shifting attention from language coverage to the differential positioning of languages within digitally structured environments, this study reconceptualizes linguistic imperialism in the AI era—not as a direct continuation of colonial policy, but as a rearticulation of linguistic authority through optimization processes, data density, and architectural design.

Debates on linguistic inequality and linguistic justice long predate the emergence of AI systems. Sociolinguistic and language policy scholarship has extensively examined how the dominance of particular languages can shape access to knowledge, participation, and social mobility (Bourdieu 1991; Chin 2023; May 2012; Park 2023; Phillipson 1992; Piller 2016). Recent work has begun to extend these concerns to AI-mediated communication environments, where language technologies may reproduce or intensify existing linguistic asymmetries (Blasi, Anastasopoulos, and Neubig 2022; Park 2024).

2. Theoretical Framework: From Linguistic Imperialism to Infrastructural Power

2.1. Linguistic Imperialism and the Institutional Location of Power

The concept of linguistic imperialism has traditionally been used to explain how language hierarchies are reproduced through institutional and geopolitical structures. Phillipson (1992) argues that the global dominance of English was neither accidental nor purely communicative in nature; rather, it was historically sustained through colonial administration, educational policy, international organizations, and the global publishing industry. Language expansion, in this account, is closely tied to asymmetrical power relations that extend beyond linguistic form into political economy.

Within this framework, English became embedded in educational curricula, academic accreditation systems, and international diplomacy. Its dominance was visible, codified, and often explicitly legitimized. Institutional arrangements conferred authority upon English, while other languages were positioned as peripheral or local. Linguistic hierarchy was therefore reproduced through formal governance structures.

Bourdieu's notion of linguistic capital deepens this analysis by emphasizing how languages acquire symbolic value within specific social fields (Bourdieu 1991). Linguistic competence is not merely communicative capacity; it is a resource convertible into institutional legitimacy and social mobility. Speakers of dominant languages gain access to opportunities that are less available to speakers of subordinated languages. Importantly, such capital is not inherent to the language itself but emerges from its position within structured power relations.

In classical formulations, then, linguistic imperialism is institutional: It operates through schools, policies, accreditation systems, and cultural industries. Authority is stabilized through visible structures.

Recent scholarship has revisited linguistic imperialism in the context of globalization and digital media, noting that English dominance persists in varied forms even where explicit policy enforcement has receded. These analyses extend Phillipson's framework beyond formal governance structures and suggest that linguistic hierarchy may persist even when explicit policy enforcement recedes.

While such work successfully demonstrates that linguistic power circulates through global media systems and digital platforms, it primarily examines the visibility and distribution of English across communicative networks. It leaves comparatively underexplored the internal architectures through which language is computationally processed, ranked, and generated.

The contemporary digital environment complicates the classical institutional model not only because communication now occurs online, but because linguistic mediation increasingly takes place within machine learning systems themselves. AI-mediated communication introduces a qualitatively different site at which hierarchy may be reproduced: computational infrastructure.

2.2. From Institutional Mandate to Infrastructural Default

The rise of large language models and machine translation systems marks a further shift in the location of linguistic mediation. Rather than relying primarily on educational institutions, publishing industries, or even media platforms, language interaction increasingly occurs within algorithmic systems trained on large-scale digital corpora. These systems are not neutral conduits; they are structured

environments shaped by dataset composition, annotation practices, model architecture, and optimization objectives.

Research on multilingual NLP has documented the uneven distribution of linguistic resources across languages. Joshi et al. (2020) demonstrate that the majority of NLP research attention and dataset availability is concentrated in a small subset of high-resource languages. This imbalance reflects broader inequalities in digital language presence, including disparities in web content, linguistic resources, and online textual production. Similarly, Ruder, Vulić, and Søgaard (2019) note that cross-lingual transfer effectiveness is closely linked to the availability and structure of multilingual training data.

Large multilingual models attempt to mitigate these imbalances by training on large-scale multilingual corpora. However, their training environments remain shaped by substantial differences in data availability across languages.

Documentation of large web-scale datasets further reveals how language representation is unevenly distributed across crawled corpora (Dodge et al. 2021). These asymmetries are not simply technical artifacts but reflect broader structural inequalities in the production and distribution of digital text.

This shift suggests that linguistic dominance may increasingly operate through infrastructural default rather than institutional mandate. English need not be formally declared the preferred language of AI systems in order to function as such. When training data, evaluation benchmarks, and alignment datasets are disproportionately concentrated in English, the resulting systems may systematically exhibit stronger performance in that language. Authority emerges not from explicit policy but from architectural embedding.

The relocation of power from institutions to infrastructure does not eliminate earlier forms of linguistic imperialism. Instead, it rearticulates them. What was once stabilized through curricula and accreditation

may now be stabilized through optimization gradients and data density. The site of hierarchy shifts from classrooms and policy documents to model architectures and datasets.

2.3. Algorithmic Linguistic Privilege and Infrastructural Amplification

To conceptualize this shift, this paper introduces the notion of algorithmic linguistic privilege. Drawing on Bourdieu's concept of linguistic capital (Bourdieu 1991), algorithmic linguistic privilege refers to the patterned advantage certain languages enjoy within AI systems due to their infrastructural embedding.

Where linguistic capital describes symbolic authority within social fields, algorithmic linguistic privilege describes infrastructural reinforcement within computational systems. Languages that occupy dense positions in training corpora, annotation pipelines, and evaluation benchmarks are more likely to generate outputs perceived as coherent, contextually nuanced, or authoritative. This advantage does not stem from intrinsic linguistic properties; it arises from differential representational saturation.

The concept of infrastructural amplification clarifies this dynamic. Historically accumulated linguistic capital intersects with computational infrastructure, producing a feedback loop between symbolic authority and model performance. English, already globally prestigious, is also densely represented in digital corpora. As Bommasani et al. (2021) observe, foundation models inherit the properties of their training data. When data concentration aligns with historical prestige, symbolic capital becomes computationally amplified.

Importantly, this process need not involve intentional bias. As Bender et al. (2021) caution in their critique of large-scale language models, training on vast but unevenly curated corpora can reproduce structural inequalities even in the absence of explicit design

preference. The neutrality of scale does not guarantee equity of representation.

Algorithmic linguistic privilege thus reframes linguistic dominance in infrastructural terms. It highlights how computational systems can stabilize performance hierarchies through inherited asymmetries in textual production.

2.4. Structural Asymmetry as Historical Sedimentation

The persistence of linguistic hierarchy in AI systems can therefore be understood as a form of structural asymmetry. Structural asymmetry refers to historically accumulated inequalities in data production, curation, evaluation, and alignment that become embedded within computational architectures.

These asymmetries are neither random nor purely technical. They reflect the geopolitical concentration of publishing industries, the dominance of English-language academic indexing, and the centrality of Anglophone media in the development of the modern internet. The digital sphere did not emerge in a linguistic vacuum; it inherited global power relations that continue to shape patterns of textual production and circulation.

Couldry and Mejias (2019) describe data colonialism as the appropriation and structuring of social life through large-scale data extraction. While their analysis does not focus specifically on language technologies, it demonstrates how digital infrastructures can reproduce and intensify existing global asymmetries. When certain languages dominate digital textual production, they are more likely to occupy central positions within data extraction processes and large-scale model training. Linguistic hierarchy thus becomes embedded within the data pipelines that sustain contemporary AI systems.

The uneven distribution of linguistic resources in AI systems is

often framed as a purely technical consequence of data availability. However, patterns of digital textual production are themselves historically shaped by geopolitical, economic, and linguistic power relations. Data asymmetries, therefore, cannot be understood independently from the broader historical conditions that structure the global production of knowledge and communication.

Understanding AI-mediated communication through structural asymmetry shifts analytical attention away from intrinsic linguistic properties and toward processes of historical sedimentation. The increasing multilingual capacity of contemporary AI systems does not automatically eliminate these asymmetries. While more languages may now be computationally supported, they remain unevenly positioned within the infrastructures that govern data density, model training, and evaluation regimes. Inclusion in principle does not guarantee equivalence in infrastructural depth.

In this sense, linguistic imperialism in the AI era may not appear as overt institutional enforcement. Instead, it manifests as differential embedding within computational systems. The logic of hierarchy persists, but its mechanisms become architectural rather than declarative, operating through data pipelines, model architectures, and optimization processes.

Linguistic hierarchy has therefore not disappeared in the age of AI; it has migrated from institutions into infrastructure.

3. Compensatory Linguistic Labor and the Infrastructural Conditioning of Choice

3.1. Re-Prompting and the Everyday Experience of Asymmetry

If algorithmic linguistic privilege operates through computational mediation, its effects become visible in everyday linguistic practice. Multilingual AI systems accept inputs in numerous languages and are often celebrated for this inclusivity (Scao et al. 2022; Xue et al. 2021). Yet users may encounter subtle performance gradients across linguistic pathways.

When certain languages appear to yield more detailed, stable, or contextually responsive outputs, language users may adjust their practices. Bilingual users may reformulate queries in English; others may translate prompts into English before submission and translate responses back into their primary language.

Empirical research complicates the assumption that such adaptation necessarily improves outcomes. A comparative benchmark study of Swahili-language AI systems shows that translating prompts into English does not consistently enhance performance and may introduce additional errors (Jaffer and Sayer 2025). Perceptions of English as an optimization pathway, therefore, do not always align with measurable system behavior.

Such adjustments may appear individually rational but collectively reproduce linguistic hierarchy. Linguistic accommodation itself is unremarkable in multilingual settings. However, when one language repeatedly functions—or is widely perceived to function—as a more reliable interface to computational authority, adaptation acquires structural significance. English becomes not merely a communicative option but a pathway through which linguistic value appears more

readily recognized.

The issue, then, is not formal freedom of choice but the patterned incentives shaping linguistic practice. Even absent explicit coercion, differential processing conditions may condition how users position their own languages in AI-mediated interaction.

3.2. From Tactical Adaptation to Structural Expectation

To clarify this dynamic, it is useful to distinguish between tactical adaptation and structural expectation. Tactical adaptation refers to occasional strategic adjustments; structural expectation emerges when such adjustments become normalized as necessary for effective participation.

Earlier forms of linguistic imperialism made dominance explicit. Academic publication, for example, privileged English as the language of international visibility, embedding translation within formal accreditation systems. In AI-mediated communication, by contrast, no policy mandates English input. Yet repeated encounters with performance differentials—or circulating assumptions about them—can generate a tacit expectation that higher-quality output requires operating through English.

Here, the distinction between coercion and conditioning becomes central. Classical linguistic imperialism operated through institutional structures that privileged English in education, publishing, and international communication (Phillipson 1992). In AI-mediated contexts, hierarchy operates through perceived optimization pressure. Users are not compelled to abandon their primary language, but they may infer advantages in temporarily doing so. Hierarchy is encountered as a gradient rather than a rule.

Even when empirical findings complicate claims of English superiority, symbolic associations between English and computational

robustness may persist. Structural expectation can thus stabilize independently of consistent performance differences. In this way, tactical adaptation can gradually crystallize into a structural expectation about how AI systems should be approached linguistically.

3.3. Compensatory Linguistic Labor: Effort, Epistemic Risk, and Symbolic Consequences

This paper conceptualizes these adaptive practices as compensatory linguistic labor: the additional cognitive, temporal, and translational effort undertaken by users whose languages occupy less privileged positions within computational mediation.

Compensatory linguistic labor may include:

i. Pre-submission translation of prompts into a higher-resource language.
ii. Iterative rephrasing to align with discourse patterns more densely represented in training data.
iii. Contextual simplification of culturally specific references.
iv. Post-output translation and reinterpretation within one's primary linguistic environment.

Each step entails interpretive responsibility. Unlike routine multilingual negotiation between speakers, this labor responds to infrastructural asymmetry. The aim is not interpersonal understanding but alignment with unevenly distributed representational density.

The term *labor* is not metaphorical. Digital systems frequently redistribute productive effort onto users while presenting such effort as voluntary participation (Scholz 2013; Terranova 2000). In multilingual AI interaction, language users absorb the burden of negotiating structural asymmetries embedded in model design.

Compensatory linguistic labor does not eliminate inequality; it redistributes its cost. Rather than correcting representational imbalance at the infrastructural level, the system permits users to approximate parity through additional effort.

Beyond effort, compensatory linguistic labor also entails epistemic risk. Translation involves interpretation and abstraction; reformulating knowledge claims within a linguistically privileged pathway may alter nuance or reshape culturally embedded concepts. Fricker's (2007) account of epistemic injustice clarifies this risk. When structural conditions undermine an individual's capacity to participate as a knower, injustice occurs. In AI-mediated interaction, the need to reformulate claims in a different linguistic register can shift interpretive burden onto the user. Distortion in translation may then be attributed to phrasing rather than to asymmetrical mediation.

Although AI systems appear linguistically inclusive, performance gradients can individualize responsibility. Structural asymmetry is reframed as a matter of user optimization. The system remains unchanged; linguistic adjustment becomes the user's task.

Compensatory linguistic labor also has symbolic consequences. When English repeatedly appears—or is widely believed—to yield more authoritative responses, associations between English and epistemic reliability may be reinforced through routine interaction. Bourdieu's (1991) framework is instructive here. Linguistic capital is sustained not only through formal recognition but through everyday reinforcement. In computational mediation, differential embedding may stabilize perceptions that certain languages are inherently more precise or analytically robust, even when such perceptions reflect infrastructural positioning rather than intrinsic superiority.

Over time, hierarchy can become naturalized. Linguistic dominance is experienced as pragmatic efficiency rather than historical contingency. English is not declared superior; it is encountered—or

assumed to be—more effective. This shift from explicit assertion to experiential inference marks a central feature of the infrastructural reconfiguration of linguistic imperialism. Compensatory linguistic labor, therefore, represents the everyday mechanism through which infrastructural asymmetries are translated into lived linguistic practice in AI-mediated communication.

These dynamics suggest that linguistic hierarchy in AI-mediated communication is not maintained primarily through explicit institutional enforcement but through the everyday organization of interaction within computational environments. Algorithmic linguistic privilege does not simply shape system performance; it also reshapes linguistic practice by redistributing the effort required to achieve epistemic recognition. Compensatory linguistic labor, therefore, becomes a routine mechanism through which infrastructural asymmetries are negotiated in practice. What appears as individual linguistic choice is thus conditioned by the structural organization of computational mediation. In this sense, the persistence of linguistic hierarchy in AI systems is not imposed through overt linguistic policy but reproduced through the subtle alignment of user behavior with unevenly distributed linguistic infrastructures.

4. Discussion: The Infrastructural Reconfiguration of Linguistic Imperialism

4.1. From Institutional Governance to Platform Governance

The preceding analysis suggests that linguistic hierarchy in the AI era operates through mechanisms distinct from those described in classical accounts of linguistic imperialism. In Phillipson's (1992)

formulation, language dominance was sustained through visible institutional governance—education systems, colonial administration, and international organizations. English gained authority through educational policy, institutional prestige, and geopolitical consolidation.

In AI-mediated contexts, linguistic authority is no longer stabilized primarily through ministries or diplomatic mandates. Instead, it is mediated through platforms, data pipelines, and model architectures. The site of linguistic hierarchy shifts from institutional governance to computational mediation.

This shift does not displace earlier dominance; it overlays it. English remains globally prominent in academia and commerce, but its authority is further reinforced through training data density, benchmark design, and optimization processes. Hierarchy is not confined to curricula or policy documents; it is embedded in systems that process and generate language at scale.

Such relocation alters the visibility of power. Institutional mandates are explicit and contestable. Infrastructural defaults are often encountered as technical features. Linguistic privilege may thus appear as an emergent property of scale rather than as the continuation of historical asymmetry.

4.2. Optimization and the Production of Linguistic Authority

A defining feature of this configuration is optimization. Foundation models are trained to maximize predictive performance across vast corpora (Bommasani et al. 2021). Because evaluation benchmarks often reflect the distribution of available annotated data, uneven corpora can reinforce representational density in already dominant languages.

Optimization does not require explicit preference. When English-language data are more abundant and diverse, performance gradients

may systematically favor English without a formal declaration of priority. Linguistic authority is thereby reproduced through statistical reinforcement rather than institutional decree.

Comparative analyses of multilingual LLMs suggest that model responses to moral evaluation tasks can reflect culturally specific value orientations associated with dominant training contexts (Aksoy 2024). Infrastructural asymmetry may therefore shape not only coherence but normative framing.

This marks a transformation in linguistic governance. Whereas classical linguistic imperialism relied on institutional exclusion (Phillipson 1992), AI-mediated hierarchy operates through incentive structures embedded in optimization. Users are not prohibited from using non-dominant languages, but they may perceive advantages in aligning with the infrastructurally privileged pathway. Optimization shapes linguistic practice through differential reward rather than prohibition.

As Couldry and Mejias (2019) argue in their analysis of data colonialism, digital infrastructures structure possibilities through design. When such design conditions the recognition of linguistic value and even normative authority, governance becomes embedded in mediation itself.

4.3. Redistribution of Responsibility and the Reconfiguration of Linguistic Imperialism

The movement from coercion to optimization entails a redistribution of responsibility. When performance disparities are navigated through compensatory linguistic labor, the burden of achieving parity shifts from infrastructural design to individual language users.

This redistribution parallels analyses of digital labor, in which user activity generates value while appearing voluntary and participatory

(Scholz 2013; Terranova 2000). In multilingual AI interaction, cognitive and translational work may be reframed as strategic competence rather than as a response to asymmetrical mediation.

Such framing individualizes inequality. If improved outcomes appear available through English input, the rational response becomes linguistic adjustment. Structural asymmetry is rendered manageable through personal effort, while infrastructural conditions remain intact.

When normative clarity or argumentative coherence is also associated with dominant linguistic pathways, this individualization deepens. Users may internalize not only performance expectations but epistemic hierarchies, aligning their practices with perceived centers of rationality. Because infrastructural conditioning is experienced as technical design rather than policy, it becomes more difficult to contest.

This transformation should not be reduced to simple continuity. The AI era does not merely replicate colonial language policy in computational form. Multilingual capacity represents a genuine expansion of participation (Scao et al. 2022; Xue et al. 2021). Languages previously excluded from digital systems now engage with them.

The term *reconfiguration*, therefore, signals both continuity and change. Historical concentrations of linguistic capital intersect with computational mediation in new ways. Institutional linguistic imperialism operated through explicit prestige hierarchies; infrastructural reconfiguration operates through routinized performance gradients and subtle normative calibrations.

AI systems neither abolish multilingual inclusion nor impose uniform conformity. Rather, they reorganize the terrain on which linguistic inequality unfolds. The central question is not whether AI is inherently imperial, but how inherited asymmetries interact with computational mediation to produce differentiated linguistic outcomes.

4.4. Implications for Linguistic Justice

Viewing linguistic imperialism as infrastructural reconfiguration shifts the normative focus. Linguistic justice cannot be assessed solely through metrics such as the number of supported languages or translation accuracy. It requires examining how languages are positioned within training datasets, evaluation benchmarks, alignment regimes, and normative calibration processes.

If structural asymmetry is embedded in data pipelines (Dodge et al. 2021; Joshi et al. 2020), redressing inequality demands infrastructural intervention: balanced dataset curation, multilingual evaluation standards, diversified annotation practices, and culturally plural alignment strategies.

More fundamentally, the discussion raises a design question. If linguistic hierarchy is mediated through computational environments, can linguistic justice be pursued through deliberate architectural choices? Historical experiments with auxiliary and constructed languages demonstrate that linguistic systems can be structured toward equity. In the AI context, this does not imply replacing natural languages, but interrogating the defaults through which linguistic authority is computationally stabilized.

The task is not to reject technological development but to render visible the conditions under which languages are differentially recognized. Only by making infrastructural asymmetry legible can linguistic justice become a matter of design rather than aspiration.

5. Conclusion: Linguistic Hierarchy in the Age of Infrastructure

The rapid development of multilingual large language models has generated widespread optimism that computational mediation may reduce longstanding linguistic inequalities. Because contemporary AI systems can process and generate text across dozens or even hundreds of languages, it is often assumed that technological mediation will gradually dissolve the linguistic hierarchies that historically structured global communication.

This article has argued that such expectations require qualification. While multilingual capacity has expanded significantly, linguistic hierarchy has not disappeared. Instead, it has been reconfigured. Drawing on Phillipson's (1992) account of linguistic imperialism and Bourdieu's (1991) theory of linguistic capital, the analysis has shown how historical asymmetries in textual production intersect with computational infrastructures. Languages that occupy dense positions within training corpora, benchmarks, and alignment datasets acquire algorithmic linguistic privilege, benefiting from infrastructural amplification within contemporary AI systems.

This infrastructural positioning does not operate through overt coercion. Rather, it emerges through optimization processes and data concentration. When certain linguistic pathways consistently yield more stable or detailed outputs, language users may adapt their interactional practices accordingly. The concept of compensatory linguistic labor captures this dynamic: Users whose languages occupy less privileged positions within computational infrastructures may undertake additional cognitive, translational, and interpretive work in order to obtain reliable responses.

Through these mechanisms, the burden of navigating linguistic asymmetry can shift from system design to individual users. Multilingual inclusion, therefore, coexists with infrastructural differentiation. AI systems may accept inputs in multiple languages while simultaneously rewarding certain linguistic pathways through performance gradients embedded in training data and model architecture.

Understanding this transformation requires reconceptualizing linguistic imperialism for the AI era. Classical accounts emphasized institutional governance—schools, policy regimes, and publishing systems—as the primary mechanisms through which language hierarchy was reproduced. In contemporary AI systems, by contrast, linguistic authority is increasingly mediated through computational infrastructures. Hierarchy persists not as formal exclusion but as differential embedding within data pipelines, model architectures, and optimization processes.

Linguistic hierarchy, in other words, has not disappeared in the age of AI; it has migrated from institutions into infrastructure.

Recognizing this shift has important implications for debates about linguistic justice. Evaluating multilingual AI cannot be limited to counting supported languages or measuring translation fluency. It requires examining how languages are positioned within training datasets, evaluation benchmarks, and alignment regimes. If infrastructural asymmetries are historically sedimented within digital textual production, then addressing linguistic inequality demands intervention at the level of computational design.

The challenge is therefore not simply to expand multilingual coverage but to interrogate the infrastructural conditions under which linguistic authority is produced. Making these conditions visible is a necessary first step toward designing AI systems that do not merely reproduce inherited asymmetries but actively confront them. Only by

recognizing the infrastructural character of contemporary linguistic hierarchy can debates about multilingual AI move beyond technological optimism toward a more critical understanding of language, power, and computation.

References

Aksoy, Meltem. 2024. “Whose Morality Do They Speak? Cultural Bias in Multilingual Large Language Models.” arXiv preprint arXiv:2412.18863.

Bender, Emily, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In *FAccT ’21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623. Association for Computing Machinery.

Bird, Steven. 2020. “Decolonising Speech and Language Technology.” In *Proceedings of the 28th International Conference on Computational Linguistics*, edited by Donia Scott et al., 3504–3519. International Committee on Computational Linguistics.

Blasi, Damián, Antonios Anastasopoulos, and Graham Neubig. 2022. “Systematic Inequalities in Language Technology Performance across the World’s Languages.” In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, edited by Smaranda Muresan et al., 5486–5505. Association for Computational Linguistics.

Bommasani, Rishi, Drew Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael Bernstein, et al. 2021. “On the Opportunities and Risks of Foundation Models.” arXiv preprint arXiv:2310.11829v4.

Bourdieu, Pierre. 1991. *Language and Symbolic Power*. Harvard University Press.

Chin, Silo. 2023. "Linguistic Diversity and Justice: The Role of Artificial Languages in Multilingual Societies." *Journal of Universal Language* 24(2): 71–89. https://doi.org/10.22425/jul.2023.24.2.71

Couldry, Nick and Ulises Mejias. 2019. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford University Press.

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1286–1305. Association for Computational Linguistics.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Gupta, Vansh, Sankalan Pal Chowdhury, Vilém Zouhar, Donya Rooein, and Mrinmaya Sachan. 2025. "Are Large Language Models for Education Reliable for All Languages?" arXiv preprint arXiv:2504.17720.

Jaffer, Sophie and Simeon Sayer. 2025. "Artificially Fluent: Swahili AI Performance Benchmarks between English-Trained and Natively-Trained Datasets." arXiv preprint arXiv:2509.04516.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. "The State and Fate of Linguistic Diversity in NLP." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky et al., 6282–6293. Association for Computational Linguistics.

May, Stephen. 2012. *Language and Minority Rights: Ethnicity,*

Nationalism and the Politics of Language. 2nd ed. Routledge.

Park, Sunyoung. 2023. "Multilingualism, Social Inequality, and the Need for a Universal Language." *Journal of Universal Language* 24(1): 77–93. https://doi.org/10.22425/jul.2023.24.1.77

Park, Sunyoung. 2024. "AI Chatbots and Linguistic Injustice." *Journal of Universal Language* 25(1): 99–119. https://doi.org/10.22425/jul.2024.25.1.99

Phillipson, Robert. 1992. *Linguistic Imperialism*. Oxford University Press.

Piller, Ingrid. 2016. *Linguistic Diversity and Social Justice*. Oxford University Press.

Qin, Libo, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip Yu. 2025. "A Survey of Multilingual Large Language Models." *Patterns* 6(1): 101118. https://doi.org/10.1016/j.patter.2024.101118

Rohera, Pritika, Chaitrali Ginimav, Gayatri Sawant, and Raviraj Joshi. 2025. "Better to Ask in English? Evaluating Factual Accuracy of Multilingual LLMs in English and Low-Resource Languages." arXiv preprint arXiv:2504.20022.

Ruder, Sebastian, Ivan Vulić, and Anders Søgaard. 2019. "A Survey of Cross-Lingual Word Embedding Models." *Journal of Artificial Intelligence Research* 65: 569-631. https://doi.org/10.1613/jair.1.11640

Scao, Teven Le, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, et al. 2022. "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model." arXiv preprint arXiv:2211.05100.

Scholz, Trebor, ed. 2013. *Digital Labor: The Internet as Playground and Factory*. Routledge.

Terranova, Tiziana. 2000. "Free Labor: Producing Culture for the Digital Economy." *Social Text* 18(2): 33–58.

https://doi.org/10.1215/01642472-18-2_63-33

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, et al. 2021. “mT5: A Massively Multilingual Pre-Trained Text-to-Text Transformer.” In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by Kristina Toutanova et al., 483–498. Association for Computational Linguistics.