SP Publications
International Journal Of English and Studies (IJOES)
An International Peer-Reviewed and Refereed Journal;**Impact Factor:**8.337(SJIF) **ISSN:** 2581-8333|**Volume 8, Issue-5(May);2026**

A Study on How AI-Driven Language Apps Prioritize Western Dialects and Its Implications for Global Linguistic Diversity

Dr. T. Jude Livingston[1]
[1]Assistant Professor of English, Nazareth Margoschis College at Pillaiyanmanai
Affiliated to Manonmaniam Sundaranar University
Dr. Y. Premila Anbarasi[2]
[2]Associate Professor of English, Nazareth Margoschis College at Pillaiyamanai
Affiliated to Manonmaniam Sundaranar University

Article Received: 05/04/2026
Article Accepted:08/05/2026
Published Online:10/05/2026
DOI:10.47311/IJOES.2026.8.05.48

Abstract

Artificial intelligence (AI) has reshaped global language learning through applications such as Duolingo, Babbel, Rosetta Stone, and ChatGPT-based tutors. While these platforms claim to democratise linguistic access, growing evidence suggests they systematically prioritize Western dialectsover the diverse varieties spoken globally. This paper examines the structural, technical, and economic factors that lead AI-driven language apps to embed Western dialectal hegemony into their training data, speech recognition, and pedagogical content. Drawing on natural language processing (NLP) literature, sociolinguistic theory, and recent empirical audits of major language platforms, the study argues that algorithmic preference for prestige dialects produces three primary consequences such as the marginalization of regional and minority varieties, the reinforcement of colonial linguistic hierarchies, and the accelerated endangerment of low-resource languages. The paper proposes a framework for equitable AI language design grounded in participatory data collection, dialectal diversification of training corpora, and transparent reporting of variety coverage. Implications for educators, developers, and policymakers are discussed.

Keywords*:* artificial intelligence, dialect bias, linguistic diversity, natural language processing, low-resource languages

Introduction

The global proliferation of artificial intelligence (AI) in education has transformed how individuals acquire second and additional languages. Mobile applications such as Duolingo, Babbel, Memrise, Rosetta Stone, and Pimsleur, alongside large language model (LLM) tutors built on systems like ChatGPT and Gemini, now serve hundreds of millions of users worldwide (Bibauw, François, & Desmet, 2019). These platforms are routinely marketed as engines of linguistic democratizationtools that, by leveraging machine learning and adaptive feedback, promise to dismantle the gatekeeping traditions of classroom instruction and place high-quality language education within reach of any user with a smartphone. Yet a growing body of scholarship suggests that this promise of universality conceals a deeper bias: the dialects, accents, and varieties privileged by AI-driven language apps are overwhelmingly those of Western, predominantly Anglophone and Western European, populations (Bender, Gebru, McMillan-Major, &Shmitchell, 2021).

This study examines the mechanisms by which AI language platforms encode Western dialectal preferences, analyzes the sociolinguistic and pedagogical consequences of those preferences, and proposes structural reforms aimed at fostering linguistic equity. The central argument advanced is that the apparent neutrality of algorithmic language instruction is, in fact, the product of historically and economically situated choiceswhich speakers to record, which orthographies to standardize, and which markets to monetize. These choices, when aggregated across the dominant platforms in the language-learning industry, produce a global digital pedagogy that elevates a narrow band of prestige varieties while rendering thousands of other linguistic forms invisible, undervalued, or computationally inaccessible (Joshi, Santy, Budhiraja, Bali, & Choudhury, 2020).

The implications of this dynamic extend well beyond pedagogy. Linguists have long documented that language loss accelerates when speakers perceive their varieties as lacking institutional prestige or economic utility (Crystal, 2014). When AI tutors, which increasingly mediate exposure to second languages, present Parisian French as the only French worth learning or treat African American English as a deviation rather than a system, they participate in the same hierarchical processes that have historically driven minoritized languages toward extinction. This paper situates AI-driven language apps within that longer history of linguistic gatekeeping and asks what would be required to reorient them toward genuine global representation.

Literature Review

Bias in Natural Language Processing Systems

Research on bias in natural language processing has expanded rapidly since the

mid-2010s. Bender et al. (2021) introduced the influential framing of large language models as "stochastic parrots," arguing that the scale of training corpora obscures rather than resolves questions about whose language is represented. Their analysis demonstrated that the textual data scraped to train LLMs is disproportionately drawn from English-language web sources produced by users in the United States, the United Kingdom, Canada, and Australia. Even within these national contexts, the data overrepresents the speech of educated, urban, and economically privileged populations. Blodgett, Barocas, Daumé, and Wallach (2020) extended this critique by surveying more than 140 papers on bias in NLP, concluding that the field had not adequately operationalized what bias means in linguistic terms, and that most mitigation efforts focused on demographic categories rather than on the structural underrepresentation of dialects and varieties.

Joshi et al. (2020) provided the most comprehensive empirical mapping of linguistic inequality in NLP to date, classifying the world's approximately 7,000 languages into six tiers based on the availability of labeled and unlabeled data. They found that only 7 languages—English, Mandarin, Spanish, French, German, Japanese, and Arabic—occupy the highest tier, possessing rich resources sufficient for state-of-the-art NLP development. The remaining languages, which collectively account for the majority of humanity's linguistic experience, occupy lower tiers ranging from "underdog" to "left-behind." Critically, this hierarchy is replicated within languages: even where a language possesses substantial resources, those resources tend to reflect a single standardized variety associated with national or colonial centers of power.

Speech Recognition and Accent Disparity

Automatic speech recognition (ASR) systems, which underpin the pronunciation feedback features of most AI language apps, exhibit well-documented accent disparities. Koenecke et al. (2020) audited five major commercial ASR systems and found word error rates nearly twice as high for African American speakers as for white speakers of American English. Tatman (2017) demonstrated similar gaps in YouTube's automatic captioning system, with Scottish speakers and women experiencing significantly degraded performance. These findings matter for language apps in two ways. First, the same ASR architectures that misrecognize nonstandard accents are routinely repurposed for pronunciation scoring in language-learning platforms, meaning that learners who speak with accents not represented in training data may receive systematically negative feedback regardless of their actual intelligibility. Second, the asymmetry signals to learners which accents the platform considers "correct," thereby reinforcing prestige hierarchies through ostensibly objective machine evaluation.

Sociolinguistic Theory and Linguistic Hegemony

Sociolinguistic theory provides essential context for interpreting these technical findings. Bourdieu (1991) described language as a field of symbolic power in which certain varieties accumulate prestige through their association with state institutions, education, and economic capital. Phillipson (1992) extended this analysis through the concept of linguistic imperialism, documenting how English in particular has been deliberately promoted as a global lingua franca through institutional mechanisms ranging from colonial education policy to contemporary aid programs. More recent scholarship in raciolinguistics (Rosa & Flores, 2017) has argued that judgments about linguistic "correctness" are inseparable from judgments about the racialized bodies of speakers, such that even fluent users of standardized varieties may be heard as deficient when their racial identity does not match the imagined prototypical speaker. AI systems that encode prestige varieties as defaults inherit and operationalize these hierarchies at unprecedented scale.

Methodology

This study employs a qualitative meta-analytic approach, synthesizing findings from three categories of source material: (a) peer-reviewed empirical audits of AI-driven language platforms and underlying NLP systems published between 2015 and 2024; (b) technical documentation, public statements, and language-coverage disclosures issued by major language-app providers, including Duolingo, Babbel, Rosetta Stone, and Memrise; and (c) sociolinguistic and educational scholarship addressing dialect, prestige, and language endangerment. Sources were identified through systematic searches of Google Scholar, the ACL Anthology, and the Linguistic Society of America publications database, using combinations of the search terms artificial intelligence, language learning, dialect bias, accent recognition, low-resource languages, and language endangerment.

The analysis proceeded in three stages. First, the author catalogued the languages and dialect varieties offered by the five most-downloaded language-learning apps as of 2024, distinguishing between (a) languages offered, (b) varieties or dialects explicitly differentiated within those language offerings, and (c) the variety used as the unmarked default when no choice was provided. Second, the author mapped these offerings against the tiered classification of language resources and against UNESCO's Atlas of the World's Languages in Danger to assess between platform coverage and global linguistic distribution. Third, the author synthesized empirical and theoretical literature to identify the mechanisms through which platform design choices contribute to dialectal hegemony and the consequences of those choices for global linguistic diversity. This study does not involve human subjects and therefore did not require institutional review board approval.

Findings

Coverage Asymmetries Across Major Platforms

Cataloguing the language inventories of the five leading AI-driven language apps revealed striking asymmetries between the global distribution of speakers and the platforms' pedagogical priorities. Duolingo, the most widely used platform, offered courses in approximately 40 languages as of 2024, with the overwhelming majority of course volume—measured by lesson count, audio assets, and active maintenance—concentrated in roughly a dozen Western European languages and East Asian national languages. Babbel offered fewer than 15 languages. Rosetta Stone's catalogue was similarly constrained. By contrast, the world contains an estimated 7,000 living languages, of which roughly 40% are endangered (Crystal, 2014). The asymmetry is not merely a function of demand: many of the world's most widely spoken languages, including Bengali, Punjabi, Tamil, and several major African languages such as Hausa, Yoruba, and Amharic, are offered either in highly limited form or not at all by leading platforms.

Within the languages that are offered, dialectal coverage is even more constrained. French instruction across major platforms defaults almost universally to a Parisian standard, with no instructional pathway for learners interested in Quebecois, West African, Maghrebi, or Caribbean varieties of French. Spanish instruction typically defaults either to peninsular Castilian or to a generalized Latin American variety that flattens the substantial differences among Mexican, Rioplatense, Andean, and Caribbean varieties. English instruction is divided primarily between American and British standards, with Indian Englishdespite being spoken by an estimated 125 million peoplealmost entirely absent as an instructional target. Arabic instruction, where offered, is generally restricted to Modern Standard Arabic, a register used in formal writing and broadcast media but spoken natively by no one, while the diverse colloquial varieties used in everyday life across North Africa and the Middle East are systematically excluded.

Mechanisms of Dialectal Prioritization

Three interlocking mechanisms appear to drive these asymmetries. The first is data availability. Modern AI language systems require enormous quantities of text and recorded speech for training, and the digital footprint of any given language variety is closely correlated with the economic and institutional power of its speakers. Standardized prestige varieties are overrepresented in newspapers, books, parliamentary records, and search-engine-indexed websites, while regional and minoritized varieties survive primarily in oral tradition or in small-circulation publications that are rarely digitized (Joshi et al., 2020). When platforms scrape the web for training data, they inherit and amplify this preexisting

imbalance.

The second mechanism is economic incentive. AI language apps operate as commercial enterprises whose revenue depends on user acquisition and retention in markets with high disposable income. The languages prioritized for development, such as English, Spanish, and Mandarin, correspond closely to the largest paying user bases. Investment in low-resource languages yields uncertain returns, particularly when the speakers of those varieties have limited purchasing power.As a result, the rational profit-maximizing strategy for a language-app developer is to deepen offerings in already-dominant varieties rather than to broaden coverage.

The third mechanism is technical path dependency. Once an ASR system, an LLM, or a pronunciation-scoring model has been trained on a particular variety, retraining or extending it to accommodate additional varieties requires substantial engineering investment. Platforms therefore tend to layer new features atop existing infrastructure, which means that the dialectal assumptions baked into early model versions persist through successive product generations. This path dependency helps explain why even well-resourced platforms have been slow to add support for varieties such as Indian English or Caribbean Spanish despite the obvious size of the relevant speaker populations.

Consequences for Learners and Speakers

The consequences of this dialectal prioritization fall into three broad categories. First, learners are presented with a distorted picture of the languages they study. A student of French who completes a Duolingo course may be entirely unaware that the variety they have learned would be regarded as marked or formal in much of the Francophone world, or that millions of French speakers in Africa use grammatical and lexical forms not introduced anywhere in the curriculum. This narrowing of representation is pedagogically problematic because it leaves learners ill-equipped to communicate in the diverse contexts where the target language is actually used.

Second, speakers of nonstandard varieties experience their own languages as being treated as deficient by the very tools designed to teach those languages. A heritage speaker of Quebecois French who attempts to use the platform's pronunciation feedback may find their natural pronunciation scored as incorrect; a speaker of African American English who interacts with an LLM-based tutor may find their grammatical patterns identified as errors requiring correction. These experiences reproduce, at the level of individual interaction, the hierarchies that sociolinguists have long identified as drivers of linguistic insecurity (Rosa & Flores, 2017).

Third, and most consequentially, the systematic exclusion of low-resource

languages and minoritized varieties from AI-driven language platforms accelerates the broader process of language endangerment. As digital tools become the dominant infrastructure for language learning, languages that lack representation in those tools become invisible to the next generation of learners. Crystal (2014) argued that one of the most important factors in language survival is whether speakers perceive their language as having a future. AI platforms that exclude a language send a powerful signal that the language has no place in the digital economy, and that signal can hasten intergenerational shift toward dominant varieties.

Discussion

The findings outlined above suggest that AI-driven language apps are not neutral instruments of linguistic democratization but rather active participants in the global reproduction of linguistic hierarchy. This conclusion does not entail that the platforms are malicious or that their developers are indifferent to questions of equity. Many companies in the sector have publicly acknowledged the importance of linguistic diversity and have undertaken efforts to expand their language inventories. The argument advanced here is structural rather than intentional: the combination of training-data availability, economic incentive, and technical path dependency produces outcomes that systematically favour prestige varieties even in the absence of any explicit decision to do so.

Recognizing this structural dimension has important implications for how the problem might be addressed. Individual product decisions—adding a new language course, recording a new accent for the pronunciation engine—are valuable but insufficient. What is required is a reorientation of the underlying processes by which training data is collected, models are evaluated, and product priorities are set. Several promising directions emerge from the literature.

Participatory data collection offers one such direction. Initiatives such as Mozilla's Common Voice project have demonstrated that crowdsourced speech recording can produce ASR-quality datasets for languages and varieties that would otherwise lack representation, provided that communities of speakers are recognized as collaborators rather than as data sources to be extracted from. Embedding such participatory mechanisms within commercial language-app development would enable platforms to draw on the linguistic expertise of speaker communities while distributing the benefits of representation more equitably.

Researchers are also working to diversify training datasets and ensure transparency by reporting exactly which language varieties each model covers. Bender, Friedman, and McMillan-Major (2021) proposed the concept of "data statements" that document the

demographic composition of training data, allowing downstream users to assess what populations a model has and has not been exposed to. Adopting analogous transparency standards in the language-learning industry would enable learners, educators, and regulators to evaluate platform claims of comprehensiveness and to identify gaps requiring remediation.

A third direction concerns the design of evaluation metrics. Standard ASR benchmarks reward systems that perform well on prestige varieties because those varieties dominate the test sets used to compute accuracy scores. Reweighting evaluation toward variety-balanced or equity-weighted benchmarks would create incentives for developers to invest in underserved varieties, much as fairness benchmarks in other domains of machine learning have prompted methodological innovation. Without such reweighting, the apparent technical superiority of systems trained on dominant varieties will continue to justify continued investment in those same varieties.

Finally, the role of public policy deserves attention. National and international institutions concerned with linguistic and cultural heritage, including UNESCO, ministries of education, and regional language commissions, have a legitimate interest in ensuring that AI-driven language platforms do not accelerate the erosion of linguistic diversity. Policy instruments ranging from procurement standards for educational institutions to disclosure requirements for AI products marketed within particular jurisdictions could meaningfully shift the incentive landscape facing developers.

Limitations and Future Research

This study has several limitations that bear acknowledgement. As a qualitative meta-analysis, it relies on the body of empirical work that has been published in academic venues and on platform disclosures that are voluntary and incomplete. The actual training-data composition and dialectal coverage of leading platforms are proprietary and have not been independently audited at the level of granularity that a comprehensive evaluation would require. Future research employing systematic auditing methodologies, including the kinds of probe-based evaluations developed by Koenecke et al. (2020), would significantly strengthen the empirical foundation for the claims advanced here.

Additionally, the rapid pace of development in AI language technology means that any snapshot of platform offerings becomes outdated quickly. Subsequent research should track changes in coverage and quality over time, paying particular attention to whether the introduction of large multilingual models such as those built on transformer architectures meaningfully reduces or merely repackages existing dialectal hierarchies. Comparative work examining platforms developed outside Western markets,including language-learning

tools developed in China, India, and Africawould also enrich understanding of how regional contexts shape design choices.

Finally, ethnographic and learner-centred research is needed to document how speakers and learners of nonstandard varieties actually experience interactions with AI-driven language platforms. Quantitative measures of system performance capture only part of the issue; understanding the affective and identity-related dimensions of being told by a machine that one's heritage variety is incorrect requires methodological tools that this study did not employ.

Conclusion

AI-driven language apps occupy an increasingly central position in how the world's population encounters and acquires additional languages. The promise of these tools to extend linguistic access beyond the limits of traditional educational infrastructure is real and substantial. Yet the present study has argued that this promise is undermined by the systematic prioritization of Western prestige dialects across the major platforms in the sector, a prioritization rooted in the structural conditions under which AI language systems are developed and deployed. The consequences of this dynamic include the narrowing of pedagogical representation, the reinforcement of linguistic hierarchies that disadvantage speakers of nonstandard varieties, and the acceleration of language endangerment for the thousands of varieties that fall outside platform coverage.

Addressing these consequences will require coordinated action across the technical, commercial, educational, and policy dimensions of the language-learning ecosystem. Developers must embrace participatory data collection and transparent reporting; researchers must build evaluation frameworks that reward dialectal equity; educators must cultivate critical awareness in learners about the variety they are being taught; and policymakers must recognize linguistic diversity as a public good worthy of regulatory protection. Without such coordinated action, the dominant trajectory of AI-driven language education will continue to encode, at unprecedented scale, the same hierarchies that have long constrained who is heard and whose language is valued. The opportunity to chart a different course remains open, but only if the structural sources of dialectal hegemony are recognized and confronted as such.

References

Bender, E. M., Friedman, B., & McMillan-Major, A. (2021). A guide for writing data statements for natural language processing. *University of Washington*. Retrieved from https://techpolicylab.uw.edu/data-statements

Bender, E. M., Gebru, T., McMillan-Major, A., &Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). New York, NY: Association for Computing Machinery. doi:10.1145/3442188.3445922

Bibauw, S., François, T., & Desmet, P. (2019). Discussing with a computer to practice a foreign language: Research synthesis and conceptual framework of dialogue-based CALL. *Computer Assisted Language Learning, 32*(8), 827–877. doi:10.1080/09588221.2018.1535508

Blodgett, S. L., Barocas, S., Daumé, H., III, & Wallach, H. (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* (pp. 5454–5476). Stroudsburg, PA: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.485

Bourdieu, P. (1991). *Language and symbolic power* (G. Raymond & M. Adamson, Trans.). Cambridge, MA: Harvard University Press.

Crystal, D. (2014). *Language death* (2nd ed.). Cambridge, England: Cambridge University Press.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Stroudsburg, PA: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.560

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., … Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*(14), 7684–7689. doi:10.1073/pnas.1915768117

Phillipson, R. (1992). *Linguistic imperialism*. Oxford, England: Oxford University Press.

Rosa, J., & Flores, N. (2017). Unsettling race and language: Toward a raciolinguistic perspective. *Language in Society, 46*(5), 621–647. doi:10.1017/S0047404517000562

Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (pp. 53–59). Stroudsburg, PA: Association for Computational Linguistics. doi:10.18653/v1/W17-1606