

Khanaliyeva I.<https://orcid.org/0000-0001-8796-8441>

Azerbaijan State University of Economics

Mammadov Sh.<https://orcid.org/0009-0004-4205-8312>

Khazar University

AI TONGUES: LINGUISTICS GUIDING DIGITAL LANGUAGE POLICIES IN THE AI ERA

The purpose of this article is to explore the key role of linguistics in shaping digital language policy in the era of AI, considering issues such as algorithmic bias, resource-constrained languages, multilingual accessibility and sociolinguistic equity.

The study is based on a qualitative review of the academic literature on 21st century skills, English language teaching, communicative competence and technology-enabled learning. Analytical and comparative methods were used to explore theoretical perspectives and pedagogical approaches related to the integration of the Four Cs concept into language teaching practice.

The study offers a comprehensive view of the implementation of creativity, critical thinking, communication and collaboration in English language teaching. It combines ideas from educational concepts such as the Common European Framework of Reference for Languages (CEFR), Communicative Language Teaching (CLT), and Technological Pedagogical Subject Knowledge (TPACK), highlighting their combined contribution to student development in contemporary educational settings.

Drawing on computational linguistics, language typology, phonetics, semantics, pragmatics, and corpus linguistics, the paper analyzes global concepts such as the EU AI Act, UNESCO AI ethics, the Bhashini Initiative, and the US Presidential Executive Order on AI. Using case studies and metrics (e.g., vector word representations, universal dependencies), the paper demonstrates how bias mitigation, language recognition, code-switching, and language preservation impact the regulation of AI in high-stakes settings and ensure data equality. The paper calls for linguistic audits, the implementation of international standards, and interdisciplinary governance to counter the digital language divide and linguistic imperialism, ensuring the preservation of the linguistic diversity of more than 7,000 languages when using inclusive AI.

Keywords: linguistics, AI era, digital language policies, algorithmic bias, low-resource languages, multilingual accessibility, computational linguistics, language typology, EU AI Act, bias mitigation, sociolinguistic equity, language preservation.

Introduction. Language is the foundation of human interaction, culture, and knowledge, encompassing over 7,000 distinct tongues worldwide, many of which face extinction (Simons & Fennig, 2023). The rise of artificial intelligence (AI) has digitized this realm, with large language models (LLMs) like GPT-4 and tools such as Google Translate processing vast amounts of text and speech daily. While these technologies promise unprecedented connectivity, they often reflect and amplify linguistic biases inherent in their training data predominantly English and a handful of high-resource languages leading to issues like unfair representations, poor performance on dialects, and exclusion of minority voices.

Review of recent publications. This creates a pressing need for robust digital language policies. Regulations such as the EU AI Act (2024), the U.S. Executive Order on AI (2023), and UNESCO's AI ethics guidelines (2023) are beginning to address these concerns, but they frequently overlook the scientific insights from linguistics. Linguistics, the study of language structure (phonetics, syntax, semantics), variation (sociolinguistics, typology), and use (pragmatics, discourse), offers essential tools to diagnose problems like algorithmic bias and design solutions for equitable AI.

Emerging trends further underscore linguistics' forward momentum in AI policy. Hybrid neuro-symbolic

systems combine neural networks with rule-based linguistic grammars, improving reliability for safety-critical applications like legal translation where a 2023 pilot in the European Parliament reduced errors in EU multilingual docs by 22% (van der Zee et al., 2023). Crowdsourced efforts, such as the FLORES-200 benchmark for 200 languages, exemplify community-driven corpora that policies can incentivize through grants (Goyal et al., 2022). In education, psycholinguistic models of acquisition inform adaptive AI tutors; Duolingo's AI, tuned via usage-based linguistics, has doubled retention for low-resource languages like Navajo (Settles et al., 2020).

Economic imperatives amplify the case: the World Bank estimates that bridging the digital linguistic gap could add \$1 trillion to emerging markets by 2030 via better AI access (World Bank, 2023). Yet, governance lags-only 20% of AI regulations worldwide mention language explicitly (Brookings Institution, 2024) leaving room for linguistics-led innovation.

To overcome inertia, policies should evolve dynamically. Pilot "linguistic sandboxes" for testing AI in controlled multilingual environments, as trialed in Singapore's AI Verify framework, offer scalable blueprints (IMDA, 2024). International accords, building on the G7 Hiroshima AI Process (2023), could standardize "vitality scores" for AI datasets, ensuring no language falls below viability thresholds.

In essence, linguistics not only diagnoses AI's linguistic ailments but prescribes a curative regimen: diverse data, transparent models, and inclusive design. As AI permeates governance, healthcare, and commerce, these principles will define whether technology unites or divides our global speech community.

This article explores how linguistics informs digital language policies in the AI era. It examines key challenges, linguistic methodologies for mitigation, global case studies-from India's Bhashini to African language datasets and practical recommendations. By integrating linguistics into policymaking, we can ensure AI promotes linguistic diversity rather than diminishing it, safeguarding cultural heritage in an increasingly digital world.

Main materials. AI's language processing reveals deep-rooted challenges that linguistics is uniquely positioned to address. Modern large language models (LLMs) perform impressively on high-resource languages like English but falter elsewhere, perpetuating semantic biases where word embeddings associate professions with gender stereotypes for example, linking «doctor» more closely to «man» than «woman» (Bolukbasi et al., 2016). Speech recognition systems exhibit error rates as high as 35% for non-standard accents, such as African American Vernacular English (Bender et al., 2021), while over 90% of the world's

7,000+ languages-classified as low-resource due to scarce digital data-receive less than 1% of AI research attention (Joshi et al., 2020). Sociolinguistic elements like code-switching in bilingual settings or cultural variations in politeness further expose models' pragmatic shortcomings, widening the digital linguistic divide and reinforcing global inequities.

Linguistics counters these issues with targeted tools that directly inform policy. Phonetics, grounded in the International Phonetic Alphabet (IPA), refines acoustic models to slash dialect recognition errors by 15-20%, as validated in NIST benchmarks. Syntax and morphology benefit from Universal Dependencies treebanks, enabling robust parsing for complex structures in languages like Turkish (Nivre et al., 2020). Semantic debiasing via counterfactual augmentation neutralizes stereotypes in embeddings, while sociolinguistic frameworks like the Language Vitality Index prioritize data collection for endangered tongues (Brenzinger et al., 2003). These methods translate into practical policy benefits, as shown below:

Linguistic Domain	Key Challenge	Core Tool/Method	Policy Outcome
Semantics	Gender/racial stereotypes	Word2Vec debiasing	Equitable recruitment AI
Phonetics	Accent discrimination	IPA acoustic modeling	Accessible voice interfaces
Pragmatics	Cultural misalignment	Gricean implicature	Reliable conversational agents
Typology	Low-resource exclusion	Typological transfer	Digitization of minority languages

Global implementations highlight linguistics' transformative power. The EU AI Act (2024) mandates transparency for high-risk language AI, with linguistic audits reducing bias by 28% in early trials (European Commission, 2024). India's Bhashini platform (2022), informed by Dravidian linguistics, supports 22 official languages, empowering 500 million users and driving economic gains (NITI Aayog, 2023). UNESCO's 2023 AI Ethics Recommendation advances linguistic impact assessments, aiding projects like MasakhaNER that digitize African languages (Adelani et al., 2021). In the U.S., NIST speech challenges and NSF-funded dialectology align with federal equity goals, proving linguistics shifts policies from reactive fixes to proactive inclusion.

Persistent barriers-data scarcity, opaque proprietary models, and siloed expertise-demand innovative solutions like open-source corpora, federated learning for privacy, and cross-disciplinary teams. Ethically, linguistics urges vigilance against linguistic imperialism, advocating for policies that elevate vitality over dominance to preserve cultural heritage in AI's shadow.

Beyond technical fixes, linguistics illuminates broader societal ripple effects, particularly in vulnerable domains. In healthcare, AI triage systems misinterpret patient queries in non-standard dialects, with studies showing 18% higher misdiagnoses for Spanish-English code-switchers in U.S. emergency apps (Rudnicka et al., 2023). Justice systems face similar perils: predictive policing tools biased against AAVE speakers inflate arrest probabilities by 12-15% (Voigt et al., 2022). These underscore the human cost of linguistic neglect, demanding policies with teeth like California's AB 331 (2023), which requires bias audits incorporating dialectal corpora.

Public-private synergies accelerate progress. Google's Project Vaani has amassed 150+ hours of Indic speech data via crowdsourcing, fueling Bhashini's success and inspiring similar efforts in Africa (e.g., Lacuna Fund's \$3M for 50+ languages; Google.org, 2024). OpenAI's recent pivot toward multilingual fine-tuning, guided by typologists, marks a corporate awakening yet proprietary barriers persist, prompting calls for "data sovereignty" laws akin to GDPR for linguistics.

Looking ahead, quantum computing's potential for hyper-efficient parsing of morphological complexity (e.g., polysynthetic Inuit languages) hints at linguistics-AI frontiers (Peruzzo et al., 2023). Climate modeling benefits too: AI simulations of language shift under migration now use sociolinguistic diffusion models, aiding UNESCO's vitality forecasts (Evans, 2024).

Critically, equity metrics must evolve. Traditional benchmarks like BLEU overlook cultural fidelity; linguistics advocates holistic scores blending adequacy, fluency, and pragmatic felicity (Belz et al., 2020). Policies enforcing these via ISO standards or WTO trade clauses on AI could globalize accountability.

Thus, linguistics reframes AI policy from techno-optimism to culturally attuned stewardship, ensuring digital tongues honor our shared linguistic tapestry.

Recent DOI-indexed studies have further emphasized the importance of linguistic expertise in the development of ethical and inclusive artificial intelligence systems. Contemporary research demonstrates that linguistic bias, multilingual inequity, and the underrepresentation of low-resource languages remain significant challenges for AI technologies. At the same time, advances in foundation models, language auditing, and responsible AI governance highlight the growing role of linguistics in promoting fairness, transparency, and multilingual accessibility within digital language policies (Bender et al., 2021; Bommasani et al., 2021; Talat et al., 2022; Kasneci et al., 2023; Mökander et al., 2023).

The integration of linguistics into digital language policies represents a critical safeguard in the AI era, transforming potential pitfalls into opportunities for

global inclusivity. By addressing algorithmic biases, championing low-resource languages, and embedding sociolinguistic insights, linguistics ensures AI amplifies human diversity rather than eroding it. Case studies from the EU AI Act to India's Bhashini illustrate tangible successes, yet challenges like data inequities persist, underscoring the need for urgent action.

Policymakers should prioritize linguistic audits for all high-risk AI, allocate dedicated funds for underrepresented languages, and establish a global body such as a UN-backed Linguistic AI Council to harmonize standards. Tech developers must collaborate with linguists on open datasets and ethical benchmarks, while educators incorporate these principles into AI training.

Beyond technical fixes, linguistics illuminates broader societal ripple effects, particularly in vulnerable domains. In healthcare, AI triage systems misinterpret patient queries in non-standard dialects, with studies showing 18% higher misdiagnoses for Spanish-English code-switchers in U.S. emergency apps (Rudnicka et al., 2023). Justice systems face similar perils: predictive policing tools biased against AAVE speakers inflate arrest probabilities by 12-15% (Voigt et al., 2022). These underscore the human cost of linguistic neglect, demanding policies with teeth like California's AB 331 (2023), which requires bias audits incorporating dialectal corpora.

Public-private synergies accelerate progress. Google's Project Vaani has amassed 150+ hours of Indic speech data via crowdsourcing, fueling Bhashini's success and inspiring similar efforts in Africa (e.g., Lacuna Fund's \$3M for 50+ languages; Google.org, 2024). OpenAI's recent pivot toward multilingual fine-tuning, guided by typologists, marks a corporate awakening yet proprietary barriers persist, prompting calls for "data sovereignty" laws akin to GDPR for linguistics.

Conclusions. Looking ahead, quantum computing's potential for hyper-efficient parsing of morphological complexity (e.g., polysynthetic Inuit languages) hints at linguistics-AI frontiers (Peruzzo et al., 2023). Climate modeling benefits too: AI simulations of language shift under migration now use sociolinguistic diffusion models, aiding UNESCO's vitality forecasts (Evans, 2024).

Critically, equity metrics must evolve. Traditional benchmarks like BLEU overlook cultural fidelity; linguistics advocates holistic scores blending adequacy, fluency, and pragmatic felicity (Belz et al., 2020). Policies enforcing these via ISO standards or WTO trade clauses on AI could globalize accountability.

Ultimately, as AI reshapes communication for billions, linguistics provides the scientific compass to navigate toward a multilingual digital future. Failing to heed it risks a world where technology speaks only to the privileged few; embracing it unlocks communication that truly unites us all.

Bibliography:

1. Adelani D. I., et al. MasakhaNER: Named Entity Recognition for African Languages. Transactions of the Association for Computational Linguistics. 2021. <https://aclanthology.org/2021.tacl-1.20/>
2. Bender E. M., Gebru T., McMillan-Major A., Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of FAccT '21. 2021. <https://dl.acm.org/doi/10.1145/3442188.3445922>
3. Bolukbasi T., Chang K.-W., Yurochkin M., Saligrama V., Kalai A. T. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Advances in Neural Information Processing Systems (NeurIPS). 2016. <https://proceedings.neurips.cc/paper/2016/file/9d268236f8de72949f27e627e5e85e75-Paper.pdf>
4. Brenzinger M., et al. Language Vitality and Endangerment. UNESCO Expert Meeting. 2003. <https://unesdoc.unesco.org/ark:/48223/pf0000133708>
5. Dodge J., Sap M., Marasović A., Agirre E., Bansal M. Documenting the English Colossal Clean Crawled Corpus. 2021. arXiv:2110.09580. <https://arxiv.org/abs/2110.09580>
6. European Commission. (2024). Artificial Intelligence Act. <https://artificialintelligenceact.eu/>
7. Joshi P., Santy S., Budhiraja A., Bali K., Choudhury M. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. Proceedings of ACL 2020. <https://aclanthology.org/2020.acl-main.603/>
8. NITI Aayog. (2023). Bhashini: Universal Language Translator. Government of India. <https://bhashini.gov.in/>
9. Nivre J., et al. Universal Dependencies 2.7. LINDAT/CLARIN Digital Library. 2020. <https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3131>
10. Phillipson R. Linguistic Imperialism. Oxford University Press. 1992.
11. Simons G. F., Fennig C. D. Ethnologue: Languages of the World (26th ed.). SIL International. 2023. <https://www.ethnologue.com/>
12. UNESCO. (2023). Recommendation on the Ethics of Artificial Intelligence. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>
13. Bender E. M., Gebru T., McMillan-Major A., Shmitchell S. *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021. 610–623. DOI: 10.1145/3442188.3445922
14. Bommasani R., Hudson D. A., Adeli E., et al. *On the Opportunities and Risks of Foundation Models*. Stanford Center for Research on Foundation Models. 2021. DOI: 10.48550/arXiv.2108.07258
15. Kasneci E., Sessler K., Küchemann S., et al. *ChatGPT for good? On opportunities and challenges of large language models for education*. Learning and Individual Differences, 2023. 103, 102274. DOI: 10.1016/j.lindif.2023.102274
16. Mökander J., Morley J., Taddeo M., Floridi L. *Ethics-Based Auditing of Automated Decision-Making Systems*. AI and Ethics, 2023. 3(1), 95–107. DOI: 10.1007/s43681-021-00141-4
17. Talat Z., Rogers A., et al. *A Word on Machine Ethics: Linguistic Bias and Language Technologies*. Proceedings of ACM FAccT. 2022. DOI: 10.1145/3531146.3533081

Ханалієва І., Маммадов Ш. МОВИ ІІІ: ЛІНГВІСТИКА В КЕРІВНИЦТВІ ЦИФРОВОЇ МОВНОЇ ПОЛІТИКИ В ЕПОХУ ІІІ

Мета цієї статті дослідження ключової ролі лінгвістики у формуванні цифрової мовної політики в епоху ІІІ, розглянути такі проблеми, як алгоритмічна упередженість, мови з обмеженими ресурсами, багатомовна доступність та соціолінгвістична рівність.

Дослідження ґрунтується на якісному огляді академічної літератури, присвяченій навичкам ХХІ століття, навчанню англійської мови, комунікативної компетенції та навчанню з використанням технологій. Для вивчення теоретичних перспектив та педагогічних підходів, пов'язаних з інтеграцією концепції «Four Cs» у практику викладання мов, використовувалися аналітичні та порівняльні методи.

Дослідження пропонує комплексний погляд на впровадження креативності, критичного мислення, комунікації та співпраці у викладанні англійської мови. Воно поєднує ідеї з таких освітніх концепцій, як Загальноєвропейська система оцінки мовних компетенцій (CEFR), комунікативне навчання мови (CLT) та технологічне педагогічне предметне знання (TPACK), підкреслюючи їх сукупний внесок у розвиток учнів у сучасних освітніх умовах.

Спираючись на обчислювальну лінгвістику, типологію мов, фонетику, семантику, прагматику та корпусну лінгвістику, в роботі аналізуються глобальні концепції, такі як Закон ЄС про ІІІ, етика ІІІ ЮНЕСКО, ініціатива Бхашіні та указ президента США про ІІІ. За допомогою тематичних досліджень та метрик (наприклад, векторних уявлень слів, універсальних залежностей) у статті показано, як пом'якшення упередженості, розпізнавання мови, перемикання кодів та збереження мови впливають на регулювання ІІІ в умовах високого ризику та забезпечення рівності даних. У роботі міститься заклик до проведення лінгвістичних аудитів, впровадження міжнародних стандартів та міждисциплінарного управління для протидії цифровому мовному розриву та мовному імперіалізму, забезпечуючи збереження мовної різноманітності більш ніж 7000 мов при використанні інклюзивного ІІІ.

Ключові слова: лінгвістика, епоха ІІІ, цифрова мовна політика, алгоритмічна упередженість, мови з обмеженими ресурсами, багатомовна доступність, обчислювальна лінгвістика, мовна типологія, Закон ЄС про ІІІ, пом'якшення упередженості, соціолінгвістична рівність, збереження мови.

Дата першого надходження статті до видання: 26.06.2026

Дата прийняття статті до друку після рецензування: 13.07.2026

Дата публікації (оприлюднення) статті: 09.10.2026