© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1713136

The Role of Machine Learning in Preserving Languages and Promoting Digital Inclusion

Nneoma Udeze

Abstract- ***The computer era has led to new opportunities in communication and knowledge sharing that were never had before, and it has also endangered the diversity of languages in the world and increased the disparity in technology. As more languages in the world are at risk of extinction (estimated to be 40 percent), and many communities do not meaningfully access digital tools in their native languages, the importance of machine learning in language preservation is greater. This study discussed the application of machine learning technologies to document, support, and revive endangered languages and ensure digital inclusion. Automatic speech recognition had a major effect of reducing the time spent in transcription of endangered languages to almost an instant output with just a few training samples. Machine translation systems, such as new multilingual projects, have increased their languages to cover more than 200 languages and enhanced the quality of translations done on languages that were previously ignored. In addition to documentation, digital platforms, educational technologies, and available language resources that helped minority language communities were also developed because of machine learning. Overall, this paper demonstrated that machine learning offers useful means to solve linguistic vulnerability and cut digital inequality in the ever-connected world.***

Keywords: Machine Learning, Language Preservation, Digital Inclusion, Endangered Languages, Natural Language Processing, Language Diversity, Old Data Sovereignty, Limited Resource Languages, And Digital Equity.

I. INTRODUCTION

The rapid development of digital technologies across the globe has established a paradox of language diversity. Although the internet links billions of people across the globe, it operates primarily in a few dominant languages, meaning that speakers of vulnerable and minority languages are increasingly being sidelined in the digital world. According to UNESCO, a language is lost after every two weeks and with it, invaluable cultural information, worldview, and heritage (Moseley, 2010). In addition to the problem of physical access to technology, the digital divide can include the problem of linguistic access, the possibility to meaningfully use digital tools in the native language, which is a vital aspect of fair digital participation.

Machine learning (ML), a subfield of artificial intelligence that is developing systems that can learn and adapt to data (Goodfellow et al., 2016), has become one of the main tools to deal with such issues. ML systems have already acquired unparalleled potential to document, analyze, and revitalize endangered languages, as well as open opportunities to expand digital accessibility, through natural language processing (NLP), speech recognition, machine translation, and language technology (Roll & Wylie, 2016).

Hence, this paper explores how machine learning is complex in the preservation of languages and access to digital information. It discusses recent applications, methodology, and prominent case studies, and critically evaluates the issues concerning data sparsity, structural and algorithmic bias, and the necessity of the approaches that can give more importance to grassroots involvement rather than exclusively technical solutions. The paper ends with recommendations to the researcher, policy makers, and tech developers trying to use machine learning to enhance language diversity and increase digital equity.

II. THE DIGITAL EXCLUSION AND CRISIS OF LANGUAGE ENDANGERMENT

2.1 The State of Global Language Diversity

Language diversity is one of the greatest assets of humanity (Piller, 2016), the centuries of wisdom, cultural traditions, and specific approaches to cognition. But this variety is threatened with a similar danger. The Endangered Languages Project reports that over 7,000 languages in the world are now endangered and that most of those have less than 1,000 speakers (Anderson, 2011). The reasons behind the decline in languages are complicated due to globalization, urbanization, economic forces that support major languages, loss of cultural heritage

transmission, and digital inequality (Leonard, 2017; Ruiz, 1984).

Native languages are especially susceptible. Indigenous languages are still spoken in North America; for instance, around 130 languages are spoken, but most of them are considered to be severely endangered, and the number of fluent speakers is mostly represented by the older generations (Simons & Fennig, 2018). Other countries, such as Australia, South America, and certain regions of Asia and Africa, are following the same trend since historical injustices and contemporary economic and sociopolitical pressure have enhanced language change.

2.2 The Digital Divide and Linguistic Marginalization

Although the digital transformation has increased the possibilities of communication, education, and economic engagement, it has also strengthened the linguistic hierarchies that exist. It is estimated that more than half of the internet content is in English, even though of all internet users, English speakers are only about a quarter (Internet World Statistic, 2021). The top ten languages that are the most represented take up another 30 percent of online content, with thousands of languages almost nonexistent in the digital realm.

The impact of this digital language gap is widespread (Joshi et al., 2020). It denies access to the essential information, including educational resources, health information, government services, and economic opportunities to speakers of underrepresented languages. It helps in faster language change, whereby the younger generation is getting more and more inclined to perceive global dominant languages with digital involvement. It is also a massive waste of potential: the millions of people who are potentially technologically innovative and productive in the creation of knowledge worldwide are locked out because of the language barriers that confine their digital agency. Besides, the allocation of language resources in NLP is associated with colonial biases and systematic marginalization, but not with the linguistic characteristics (Blasi et al., 2022).

2.3 The Intersection of Technology and Language Rights

Language rights are acknowledged as the basic human rights in international frameworks (Piller, 2016). The documents like The Universal Declaration of Language Rights (1996) and Atlas of the World Languages in Danger by UNESCO point out the urgency of preserving linguistic diversity and the fact that all languages, regardless of their mother tongue, speakers should enjoy fair access to education, information, and technology. Planning orientations that perceive language as a resource and not a problem are vital towards the development of technology in an equitable manner (Ruiz, 1984). Nonetheless, the implementation of these principles is expensive and necessitates both technological infrastructure and long-term institutional backing, which has never existed in most low-resource languages. With digital technologies playing a huge role in civic engagement and social life, the inaccessibility of tools that are linguistically appropriate to thousands of communities is an issue of right as well as structural inequality.

III. MACHINE LEARNING TECHNOLOGIES FOR LANGUAGE PRESERVATION

3.1 Automatic Speech Recognition and Transcription

Automatic speech recognition (ASR) systems that are based on machine learning have revolutionized the process of documenting endangered languages. The conventional linguistic fieldwork is very time-consuming, where, in most cases, trained language specialists are required to handwrite hours of recordings, which can be laborious and can take 20-30 hours per hour of audio (Himmelmann, 2006). This process can, however, be dramatically accelerated with modern ASR systems, even when the dataset of a language is small.

Recent developments in limited resource modeling have been of particular importance. Researchers have, through transfer learning, that is, models trained on popular languages are applied to limited resources (Conneau et al., 2020), produced impressive results. However, such systems have to be thoroughly considered to ensure that they do not introduce inappropriate linguistic presuppositions to target languages (Bender and Koller, 2020). As an illustration, the Building Useful Languages with Big Data project shows that it is possible to create ASR systems that work on endangered languages with as little as 40 hours of transcribed speech with multilingual pretraining (Adams et al., 2018). Such methods have been effectively used for languages

such as Yoloxochitl Mixtec with word error rates of less than 20 percent with small training sets.

IV. MACHINE LEARNING FOR DIGITAL INCLUSION

4.1 Supporting Multilingual Digital Interfaces

Effective digital inclusion requires that technological platforms be available in the native language of users (Facer & Selwyn, 2021). It is now possible to make digital interfaces localized to many languages than it was many years ago due to machine learning powered translation tools. Automated systems are able to generate first-time translations, which are then enhanced by community speakers, which saves time and cost as opposed to hiring human translators only. Nevertheless, speech recognition systems are often biased when it comes to race and language, and they have much less accuracy in recognizing speakers of non-standard dialects and accents (Koenecke et al., 2020). To resolve these inequalities, it is necessary to have various training data and community-based assessments.

Voice technologies are particularly important among those communities where illiteracy is high. The ASR and text-to-speech systems allow the use of voice control devices and voice search as well as audio access to online information in native languages. Programs like Project Harmony by Google and the like are aimed at making voice recognition systems reliable in different accents and in multilingual environments.

4.2 Educational Technology and Language Learning

Educational technologies that are enhanced with machine learning enhance digital accessibility, as well as language maintenance. The smart teaching systems have the capability of offering individualized language learning experiences that are responsive to the needs of the students (Siemens, 2005). In the case of vulnerable languages, the systems can facilitate the process of revitalization by expanding the learning facilities beyond classroom environments. Although AI has the potential to make education inclusive (Holmes et al., 2022), its usage should be connected with the ethical issues such as data privacy, algorithmic bias, and equitable access (Tojimuxammadov, 2025).

Platforms such as Duolingo, controversial in linguistic circles, have spread to endangered languages such as Hawaiian, Navajo, and Scottish Gaelic, exposing millions of users to these languages. More advanced systems based on ML also examine the behavior of learners to offer individualized, data-informed teaching help.

V. CASE STUDIES AND SUCCESS STORIES

The *No Language Left Behind (NLLB)* program of Meta is one of the most ambitious attempts at the linguistic inclusivity of artificial intelligence (Conneau et al., 2020). The project designed models of translation into over 200 languages, including a large number of low-resource African, Asian, and indigenous languages that earlier systems had mostly overlooked. The most important aspect of NLLB is that it adopts open science: through the publication of model architectures, training data, and performance benchmarks as open-source, the project has allowed other researchers and communities across the world to build on this work.

The performance has been spectacular. Initial assessments indicate that the quality of translation of low-resource languages has increased up to 44 percent relative to the former systems, and thus professional quality translation is now possible in areas that were not before. In addition to the technical success, NLLB presented new evaluation methods that were aimed at low-resource situations. This is important in that the standard translation measures, which are mainly designed to work on large language pairs such as English-French or English-Chinese, are usually not able to reflect what constitutes a good translation in a minority language setting. The work of NLLB recognizes that the varying linguistic and cultural contexts need varying measures of success.

Masakhane offers a complementary model of the way language technology can be beneficial to communities. This philosophy is isiZulu and translates to we build together, and it is applied throughout the organization (Orife et al., 2020). Masakhane brings together more than 300 researchers and practitioners in Africa to create datasets, models, and research in African languages with African leadership and ownership at the centre. The given approach is a conscious contrast to the traditional research models, in which external teams take data out of communities, without actually engaging in meaningful interaction and sharing the benefits (Joshi et al., 2020).

The difference between Masakhane and other technologies is that it understands that technology cannot save languages. The initiative is an intertwined technical project and capacity-building initiative in the form of workshops, mentorship, and joint research that address local realities and demands. Masakhane has empowered both the technical and human resources capacity to support language technology in African communities by developing avenues through which African researchers can take the lead in developing language technology within their own communities. The organization shows that the most successful language preservation work can be achieved when technical innovation develops out of the organic community associations, but not by an outside imposition.

VI. CONCLUSION AND RECOMMENDATIONS

To start with, the researchers and technology developers on the intersection of machine learning and language preservation need to radically change their direction and focus on the actual community collaboration. The research is to be done with language communities, not on them, developing early and ongoing cooperation, which would make the communities beneficiaries of the results of the project without violating cultural procedures that govern language information (Leonard, 2017; Facer & Selwyn, 2021). At the same time, this necessitates the further development of low-resource approaches such as few-shot learning, active learning, and transfer learning, and their availability in open-source software and explicit documentation. In addition, common assessment indicators do not usually reflect the linguistic or cultural specifics of endangered languages (Bender and Koller, 2020; Blasi et al., 2022), and therefore, culturally relevant measures are required, which measure the specificity of vocabulary, regionalism, and preservation of information. Besides technical factors, strong ethical data gathering should be the new standard, and strong consent measures, recognition of the community ownership of data, and technical structures to facilitate community-controlled data governance (Bender and Friedman, 2018). Outside of these technical practices, the sector should proactively hire research experts within the local language communities, fund capacity building in the various regions, and offer equitable remuneration to community partners.

Equally important, policymakers and funding agencies are significant in the future of language technology. As of today, the funding mechanisms are heavily biased towards commercial projects of high-resource languages (Joshi et al., 2020), and special funding streams of low-resource language technologies are necessary. These programs should therefore include simplified community community-friendly application procedures that can allow the participation of grassroots organizations. Moreover, the promotion of language technology needs to be reinforced by using digital infrastructure, which may be achieved by investing in internet connectivity, computing devices, and digital literacy programs in linguistically diverse areas. Also, governments should implement the use of multilingual online services, so that governmental sites, educational materials, and government information could be available in all languages of a significant number of people. Finally, it is possible to note that these policy interventions recognize that digital inclusion cannot be reduced to access to devices, but meaningful engagement in your own language.

However, while machine learning has a great potential to overcome the two-fold problem of language endangerment and digital exclusion, it is not the technology that will eliminate the underlying social and political processes that contribute to the marginalization of language (Holmes et al., 2022; Pedro et al., 2019). In fact, the most significant uses of machine learning to preserve languages and achieve digital accessibility all have similar features: they are the result of an authentic partnership with the language community, they are not concerned with commercial activity, and technical innovation is embedded in the context of the larger revitalization and access programs. Importantly, one should not disregard the ethical aspects of AI in language technology (Mittelstadt et al., 2016). The language technologies should be designed with utmost consideration of their social effects (Bender and Koller, 2020) and the data privacy (Khan, 2024), the algorithmic bias (Blodgett et al., 2020; Koenecke et al., 2020), or the cultural appropriateness.

Looking ahead, the years to come will be pivotal for the future of linguistic diversity in the world. With the further development of the machine learning possibilities, the world has a severely important decision: will the technologies strengthen the existing disparities and the same language, or will they enable

communities and protect the cultural heritage of humankind? The latter should be achieved through intentionality, shifting the resources towards low-resource languages, integrating the community voices into the technological development, tackling the structural biases of the machine learning systems, and witnessing language rights as the primary human right (Piller, 2016; Ruiz, 1984). The future of machine learning in preserving languages and enabling the digital, therefore, is conditional and real. It is important to understand that it requires long-term dedication of researchers, policymakers, educators, and technologists to make equity, community partnership, and language diversity the core values. It is by this dedication alone that machine learning can be capable of realizing its potential benefit instead of continuing to pose a threat to the most vulnerable languages in the world.

REFERENCES

[1] Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., & Michaud, A. (2018). Evaluating phonemic transcription of low-resource tonal languages for language documentation. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (pp. 3356–3365). European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2018/pdf/344.pdf

[2] Anderson, G. (2011). Language hotspots: What (applied) linguistics and education should do about language endangerment in the twenty-first century. Language and Education, 25(4), 273–289. https://doi.org/10.1080/09500782.2011.577218

[3] Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587–604. https://doi.org/10.1162/tacl_a_00041

[4] Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 5185–5198. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

[5] Blasi, D., Anastasopoulos, A., & Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5486–5505. https://doi.org/10.18653/v1/2022.acl-long.376

[6] Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

[7] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[8] Facer, K., & Selwyn, N. (2021). *Digital technology and the futures of education: Towards 'non-stupid' optimism* (ED-2020/FoE-BP/27). UNESCO.

[9] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press. http://www.deeplearningbook.org

[10] Himmelmann, N. (2006). Language documentation: What it is and what it is good for. In J. Gippert, N. Himmelmann, & U. Mosel (Eds.), Essentials of language documentation (pp. 1–30). Mouton de Gruyter.

[11] Holmes, W., Persson, J., Chounta, I.-A., Wasson, B., & Dimitrova, V. (2022). *Artificial intelligence and education: A critical view through the lens of human rights, democracy and the rule of law.* Council of Europe Publishing. https://rm.coe.int/1680a956e3

[12] 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). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

[13] Khan, W. N. (2024.). *Ethical challenges of AI in education: Balancing innovation with data privacy* (pp. 1–13) [Unpublished manuscript]. NCBA.

[14] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford,

J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

[15] Leonard, W. (2017). Producing language reclamation by decolonising "language". Language Documentation and Description, 14. https://doi.org/10.25894/ldd146

[16] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). *The ethics of algorithms: Mapping the debate. Big Data & Society, 3*(2). https://doi.org/10.1177/2053951716679679

[17] Moseley, C. (Ed.). (2010). Atlas of the world's languages in danger (3rd ed.). UNESCO Publishing. https://uploads.guim.co.uk/2024/09/30/UNESCO_Atlas_of_Languages_2010.pdf

[18] Orife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K., Martinus, L., Ali, J. T., Abbott, J., Marivate, V., Kabongo, S., Meressa, M., Murhabazi, E., Ahia, O., van Biljon, E., Ramkilowan, A., Akinfaderin, A., Öktem, A., Akin, W., Kioko, G., Degila, K., . . . Bashir, A. (2020). *Masakhane -- Machine translation for Africa* [Conference paper]. AfricaNLP Workshop, ICLR 2020. https://doi.org/10.48550/arXiv.2003.11529

[19] Pedro, F., Subosa, M., Rivas, A., & Valverde, P. (2019). *Artificial intelligence in education: Challenges and opportunities for sustainable development.* UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000366994

[20] Piller, I. (2016). *Linguistic diversity and social justice: An introduction to applied sociolinguistics.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199937240.001.0001

[21] Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education,* 26, 582–599. https://doi.org/10.1007/s40593-016-0110-3

[22] Ruíz, R. (1984). Orientations in language planning. *NABE Journal, 8*(2), 15–34. https://doi.org/10.1080/08855072.1984.10668464

[23] Siemens, G. (2005). Connectivism: A learning theory for the digital age. *International Journal of Instructional Technology and Distance Learning,* 2(1). http://www.itdl.org/Journal/Jan_05/article01.htm

[24] Simons, G. F., & Fennig, C. D. (Eds.). (2018). Ethnologue: Languages of the world (21st ed.). SIL International. http://www.ethnologue.com

[25] Tojimuxammadov, J. (2025). *Ethical challenges of artificial intelligence in education. Scientia: Technology, Science and Society*, 2, 90–96. https://doi.org/10.59324/stss.2025.2(11).09