Ministry of Enterprises and Made in Italy

UNDP

AI Hub for Sustainable Development

Scaling Language Data Ecosystems to Drive Industrial Development Growth

A discussion paper authored by participants of the AI Hub for Sustainable Development's Local Language Partnership Accelerator Pilot

List of Contributors

United Nations Development Programme

Lead contributors: Alena Klatte, Barbora Bromová, Omar Jagne

Other contributors: Grace Baketa, Dwayne Carruthers, Romilly Golding, Alex Hradecky, Jennifer Louie, Keyzom Ngodup Massally, Jayant Narayan, Francesco Puggioni, Yashica Yashica

Community Members

African House of Numeric and Artificial Intelligence by ELIT (Mohamed Cissouma, Dr. Kouadio Jean-Philippe Akpoue, Moussa Bamba and Abdoulaye Doucoure)

CLEAR Global (Aimee Ansari, Christian Resch and Alp Öktem)

Cohere Labs (Sara Hooker, John Dang and Marzieh Fadaee)

Data.org (Uyi Stewart)

Equiano Institute (Jonas Kgomo)

FactCheckAfrica (Mustapha Lawal)

Afrika Digital (Grant McNulty)

IamYourClounon (Fabroni Bill Yoclounon)

iGenius (Joey Stryker)

Kamusi Project (Martin Benjamin)

Kytabu (Tonee Ndungu)

Lanfrica Labs (Chris Emezue)

Lelapa AI (Mbali Ndandani)

Masakhane Research Foundation (Tajuddeen Gwadabe)

Mauritanian Machine Intelligence Hub (Ismail Diop)

Mozilla Common Voice (EM Lewis-Jong)

MyAIFactChecker (Abideen Olasupo)

Neuravox (Gideon Abako)

Pinion Search Partners (Jennifer van Riet)

ToumAI Analytics (Naira Abdou Mohamed, Odin Demassieux, Imade Benelallam, Youcef Rahmani and Oumaima Abidi)
University of Cape Coast, Ghana (Stephen Moore)

VIOU Digital (Asabe Vincent-Otiono)

Wikinetix (Jan Goossenaerts)

The **AI Hub for Sustainable Development** was co-designed during the Italian G7 Presidency with UNDP and partners in Africa, and endorsed by the G7 Leaders in Borgo Egnazia. The work of the AI Hub tracks closely with the Italy – Africa Mattei Plan and the African Union's AI Strategy, anchored by 14 partner countries in Africa catalyzing transformation across six key sectors. Its mission is to strengthen AI foundations with Africa's innovators – in areas of green compute, data and talent – to drive private-sector-led growth and advance prosperity for all.

The **Ministry of Enterprises and Made in Italy (MIMIT)** is responsible for promoting and supporting Italian businesses and the "Made in Italy" brand. It focuses on areas such as production, economic activities, internationalization, and innovation, and aims to enhance the competitiveness of the Italian economy and its manufacturing sector.

The **United Nations Development Programme** is the leading United Nations organization fighting to end the injustice of poverty, inequality, and climate change. Working with our broad network of experts and partners in 170 countries, we help nations to build integrated, lasting solutions for people and the planet. Learn more at undp.org or follow at @UNDP and @UNDPDigital

Produced as part of the AI Hub for Sustainable Development, powered by the Ministry of Enterprises and Made in Italy and implemented by the United Nations Development Programme

AI Hub for Sustainable Development

Abstract

Most AI systems today are developed in 'high resource' languages such as English, Spanish and Mandarin. This limits access to AI services for much of the world's population and is hindering the ability to drive new industrial growth. To address this, the Local Language Partnerships Accelerator Pilot (LLAP) as part of the AI Hub for Sustainable Development is helping to digitize local languages and advance more equitable 'Language AI' tools. This paper outlines four action areas to increase Language AI: amplify awareness around the need to digitize local languages; foster collaboration among Language AI innovators; advance inclusive data collection; and scale rights-based data-sharing. The paper concludes with targeted recommendations for governments, philanthropic organizations, the private sector, academia and innovators and the role they can play in building a more inclusive AI future.

Closing the AI equity gap: Scaling language data ecosystems for inclusive digital transformation

Today's pace of artificial intelligence (AI) innovation and adoption indicates significant potential for people and our planet. AI is projected to drive as much as US$15.7 trillion in contributions to the global economy by 2030.

Across countries, AI is expected to recast entire industries and impact multiple areas critical to sustainable development, including financial inclusion, healthcare diagnostics, precision agriculture, disaster response and climate modelling. However, current projections forecast serious inequities in how this growth will be distributed globally: Currently, innovators in the United States and China hold 60 percent of all AI-related patents, positioning them to reap most of the benefits from their technological advancements. On the other hand, 95 percent of AI innovators in Africa lack access to the computing power necessary to develop and scale their innovations. This reality risks deepening the digital divide in the Global South, for as much as 70 percent of the global population.

Making AI more responsive to the needs of everyone is not only a matter of global equity -- it is a prerequisite for unlocking new avenues for creativity and long-term economic growth. More inclusive AI can remove barriers for businesses, governments and society to build, innovate and scale sustainable markets worldwide.

AI's limited linguistic diversity is one of these barriers. Many of the most transformative AI technologies of recent years—search engines, spell-checkers, recommender systems and chatbots—are built and benchmarked for just a handful of high-resource languages, such as English, Spanish and Mandarin. More than 3 billion speakers of low-resource and Indigenous languages do not have access to digital tools, (public) services and information in their native languages, limiting their opportunity to engage in digital spaces and to benefit equally from public services. Even when AI systems possess the ability to process queries in low-resource languages, the outputs tend to be slower, more costly, less culturally relevant and insufficiently covered by model safeguards. Additionally, across many AI-driven language initiatives, native speakers are often viewed solely as data providers rather than active contributors to the development, validation and governance of innovation, or as active creators of AI systems. This could lead to models that do not fully reflect linguistic and cultural nuances and reinforce the marginalization of Indigenous and local languages in the digital arena.

Voice capability for AI is also limited in these underserved languages, presenting a significant accessibility barrier. For example, many African languages, including tonal languages such as Bambara, Yoruba, Fon, and Igbo, rely on tonal variation to convey different meanings for the same word. Unique phonemes such as clicks (isiXhosa, Zulu), implosive and ejective consonants (Wolof, Hausa, Fula), and nasalized vowels (Bambara, Lingala, Fula) require specific AI adaptations that are often overlooked in standard language models.

Closing the language divide in AI is essential to catalyse locally relevant innovation, support stronger agency for marginalized communities, record and share linguistic and cultural heritage in the digital space and ensure no one is left behind—no matter the language they speak.

10 reasons to digitize low-resource languages for AI applications

1 **Expand meaningful online access** to growing populations that speak low-resource languages. Linguistic diversity in online spaces is essential to ensure that users from underrepresented communities can actively participate in and benefit from the growing digital economy.

2 **Ensure AI serves more people.** Greater availability of diverse datasets to train and refine AI systems will ensure that AI serves a wider range of use cases, narrowing the widening AI divide.

3 **Increase accessibility and inclusion** to ensure global knowledge and digital services are designed for and available to all. Translation technologies can make essential information available in multiple languages. Voice-enabled systems (including text-to-speech and sign language interfaces) can play a critical role in meeting the needs of people with disabilities or low-literacy communities.

4 **Preserve linguistic and cultural heritage in the digital age.** The digitalization of languages helps to ensure that cultural and historical artifacts, traditions and knowledge—including those that exist in oral form—are recorded, preserved, continually enriched, and available to current and future generations. When communities deem it appropriate, out-bound translation enables the inclusion of local knowledge and content within global discourse.

5 **Enable local innovation and drive economic growth.** Language datasets are needed to develop new digital products and services in people's native languages. A broader availability of datasets will allow local innovators to advance along the AI value chain and run thriving digital businesses with broader reach and improved customer service. Inbound translation can enable economic prosperity by facilitating local entrepreneurs' access to global markets, knowledge and resources that were previously mostly accessible to speakers of high-resource languages.

6 **Enable digitalization of public services.** Multilingual technology, including AI, can help governments reach remote and hard-to-access populations, increase the accessibility of digital public infrastructure (DPI), facilitate faster information-sharing, and improve monitoring and evaluation of public programmes. Multilingual, voice-enabled AI systems can also be leveraged to serve marginalized and limited literacy populations by enhancing public service delivery, online and offline.

7 **Improve digital education.** Digital resources unlock new ways of teaching Indigenous languages and create opportunities to localize digital literacy programmes and expand the teaching of other world languages, including high-resources languages such as English.

8 **Facilitate digital trust** by ensuring meaningful participation, agency and fully informed consent of underrepresented language speakers in decisions regarding how their data is used, processed and stored. This contributes to making digital infrastructures more reliable and trusted by individuals and communities.

9 **Ensure effective and inclusive public communication.** Vital communications regarding public policies and behavioural interventions are highly effective in local languages—and inclusive language technologies can scale their reach. Multilingual technologies can also improve public education and help reach linguistically diverse communities.

10 **Address gaps in human security and Trust & Safety.** Foundational natural language processing (NLP) applications (e.g., spam filters, authentication, recommenders and content moderation systems) and AI models (e.g., chatbots) are less accurate and more vulnerable to malicious misuse in low-resource languages. Linguistically diverse NLP research has the potential to close these gaps while exploring new frontiers of NLP.

Creating the foundations of long-term Language AI ecosystems for all

Limited linguistic diversity is just one of the barriers slowing inclusive and equitable AI innovation. Other challenges include limited access to green computing infrastructure, the need to develop new technical talent and difficulties in securing funding for start-ups and non-profit actors. These are all essential for building and sustaining robust ecosystems capable of leveraging the benefits of AI while managing its risks.

- Access to **green compute** (and associated physical and digital infrastructure) is needed to securely store datasets, to develop and deploy affordable language models and to sustainably build and scale practical applications.
- Access to **AI/data talent** and digital upskilling enables job prosperity within local communities, facilitates equitable distribution of innovation gains and enriches research and development with fresh perspectives.
- **Funding for start-ups and non-profits** nurtures the development of local projects active at various stages of the AI value chain, fostering community involvement and representation through and within Language AI.

The individual and collective benefits deriving from inclusive language technology—as well as the development and maintenance of a vibrant innovation ecosystem beyond Language AI—depend on investments in these foundational elements. However, more targeted interventions that specifically develop and deploy AI-enabled projects in low-resource linguistic communities are needed. This would catalyse local innovation and drive inclusive and economically sustainable digital transformation.

Approach and methodology

As part of the co-design of the AI Hub for Sustainable Development in 2024, the Local Language Partnerships Accelerator Pilot was undertaken with the goal of developing Language Ai solutions that can deliver society-wide benefits. Over a three-month learning journey, UNDP, the G7 Italian Presidency and with the support of seven learning advisor organizations collaborated on the Pilot, connecting more than 70 innovators from 17 countries on the African continent and beyond.

The Pilot included various components: 15 detailed consultations with local and global Language AI professionals and a set of workshops across three topical learning tracks. These included: digitizing local languages at scale; data governance and community ownership of language data; and developing and managing sustainable partnerships for Language AI. Through these activities, participants explored the following research questions:

1. How might we unlock and digitize African language datasets at scale, while preventing extractive data practices?
2. How might we enable community ownership models and stewardship of language data for the public good?
3. What kinds of partnerships could drive the responsible and scalable digitalization of low-resource languages? What principles should private sector Language AI technologies adopt for the public good?
4. What is the role of under-resourced languages in start-ups and industrial acceleration? What are the use cases for Language AI in key industrial areas?

The Pilot's findings were further supported by desk research into the latest developments in NLP for low-resource languages and insights from a 2024 convening in Nairobi led by the AI for Development Funders Collaborative. Discussions were also informed by practical insights from country pilots focused on low-resource language digitalization, launched by UNDP Country Offices in Ghana.

Four action areas for inclusive Language AI

The Local Language Partnerships Accelerator Pilot identified four ways in which innovative partnerships can drive meaningful progress towards inclusive and sustainable Language AI ecosystems.

1. Amplify awareness and build momentum

Government institutions, public stakeholders, language communities and infrastructure providers are rarely engaged in efforts to digitize local languages. Sociocultural challenges, such as the stigma surrounding local languages and positive bias towards non-Indigenous languages,[3] including English, French or Portuguese, can further hinder participation and adoption. This tends to result in less favourable policy environments, limited community trust and a lack of material support for Language AI initiatives and entrepreneurs.

To raise greater awareness, it is important to:

- Advocate for the digitalization of local languages to be a strategic priority within regional and national digital strategies, especially in areas related to AI and data governance.

  - South Africa has enshrined a commitment to multilingualism in its 1996 Constitution and set up the Pan South African Language Board (PanSALB) to develop and preserve 12 official languages as well as prevent linguistic discrimination.
  - Nigeria's 2022 National Language Policy includes an ambition to promote the teaching of STEM and other subjects in local Nigerian languages, as well as implement their effective utilization in the ICT sector.
  - The European Commission and European countries such as Iceland, the Netherlands and Spain have created digital programmes focused on preserving languages in the digital sphere.

- Run regional and/or global campaigns to identify and highlight government 'champions' of language data and connect leaders to exchange good practices. Campaigns, discussions and events could include unconventional actors, such as Village Development Committees (VDCs) or other cultural collectives.

  - The African Academy of Languages (ACALAN), acting as the African Union (AU) language organ and official steward of language policy within the AU, has convened linguists, technologists and stakeholders from across Africa and finalized a detailed plan for a technology platform for African languages.
  - Mozilla Common Voice collaborates with governments to support public participation in language data creation—such as in Project AINA with the Catalan Government.

- Facilitate cross- and within-sector interactions among stakeholders working with and managing language data across the data value chain, be it analog, digitized, digital-native or related to Language AI. Bring together researchers, teachers, domain experts, public broadcasters, internet SMEs, public servants and other stakeholders to facilitate understanding of community needs, preferences and resources available.

  - The UNESCO conference, Language Technologies for All (LT4All) connects policymakers, technologists, linguists and other practitioners interested in advancing language technologies for cultural preservation and empowerment.

Partnership opportunities include:

- Public–private coalitions that unite government institutions, local tech and academia to digitize non-personal or open data.
- Public–private partnerships that integrate AI-driven language tools into education, public services and administrative systems (such the Nigerian start-up Awarri partnering with the Federal Ministry of Communications, Innovation, and Digital Economy to build the first large language model (LLM) supported by the Nigerian Government.

- Platforms and channels for cross-border collaboration among policymakers and public servants to share knowledge, discuss best practices, advocate for shared causes and inspire new projects (such as the International Decade of Indigenous Languages 2022–2032 proclaimed by the United Nations General Assembly and other multistakeholder events such as the AI Action Summit in France or the Global AI Summit on Africa in Rwanda).

2. Foster collaboration among Language AI innovators

Current efforts to digitize local languages and expand access to culturally adapted Language AI remain disjointed. Spaces and fora suitable for coordination and knowledge transfer are limited, and projects often lack interoperability, restricting opportunities for collaboration. This fragmentation also makes it harder to trace data provenance and manage and maintain datasets—key areas for AI readiness.

To strengthen these exchanges, it is important to:

» Support existing opportunities for ecosystem exchange and create new spaces for knowledge-sharing, specifically on the challenges faced by under-resourced NLP communities.

- Deep Learning Indaba is a conference and community for African AI and machine learning researchers and practitioners, held annually since 2017.
- Masakhane, a grassroots organization whose mission is to strengthen and spur NLP research in African languages, for Africans, by Africans.

» Fund travel and support visa application processes to amplify the representation of Indigenous voices in regional and global debates around Language AI. Support advocates to identify joint messages and represent the diverse interests of their language communities, e.g., by funding regional workshops and research. Advocate for events in less visa-restrictive countries.

» Support the development and scaling of existing solutions and initiatives that catalogue and maintain logs of ongoing work, past research and existing datasets.

- Lanfrica is a directory of African language resources, an online platform and research community that helps mitigate the difficulty of discovering African works.
- AfricArXiv is a pan-African, community-led digital archive created to enhance the visibility and discoverability of African research and resources.

Partnership opportunities include:

- Partnerships between regional (and global) coalitions of NLP academics and practitioners to facilitate knowledge-exchange and pooling data resources
- New regional partnerships in currently underserved regions and/or with more diverse actors (e.g., language professionals). These may also be convened for a limited amount of time to solve a specific collaboration challenge or address a specific domain—such as the Data Futures Lab Showcase at MozFest Zambia 2024, featured during the Pilot project's track on 'data governance and community ownership of language data'.

3. Advance inclusive data collection and cataloguing

Some of the challenges of low-resource NLP stem from a scarcity of human, material and technical resources across the data value chain. Limited access to linguistic expertise, insufficient hardware infrastructure and devices for data collection, and underdeveloped systems for reliable data cleaning often slow progress and constrain what is possible.

To invest in innovative approaches, it is important to:

» Improve access to linguists and translators, engage local communities to contribute and validate data (with fair compensation), and develop trusted community liaison.

- Kytabu works with teachers who take on the role of linguists and translators to improve data collection and curation of digital education resources.
- The self-service platform on Mozilla Common Voice supports open-source data collection for more than 200 language communities.
- The Kamusi Project is a collaborative online dictionary using crowdsourcing strategies, including games, to collect language data.
- iAfrika employs a community model for African language speakers to record and share cultural heritage and knowledge in their own languages through mobile platforms like Diji.
- ToumAI Analytics transfers learnings from well-presented languages to low-resource languages (for example leveraging Swahili for Comorian Dialects.)
- The Aya Initiative is a global open science movement led by Cohere Labs that has compiled the largest multilingual dataset to date.
- Audiopedia.AI bridges the gap between the oral traditions of Indigenous languages and modern digital environments. It uses a keyword-spotting approach, where community members speak a keyword (e.g., "child fever") into a knowledge platform and a pre-recorded response tagged to the keyword is played.

» Identify and scale emerging and effective data collection approaches that are adapted to the needs and cultural contexts of local language communities. Many communities that would benefit the most from AI-driven language tools face limited internet access and low digital literacy, making lightweight, offline and SMS-based solutions essential.

- DVoice, developed by ToumAI Analytics, is an open-source platform for speech data collection, transcription and AI development for African languages such as Swahili, Hausa, Wolof, and Comorian. The platform is expanding into speech-to-text data collection, translation and model fine-tuning.
- Viou is a youth-driven, geolocation-powered content marketplace that transforms everyday cultural expression—videos, photos, audio, and illustration—into AI-ready datasets. The platform enables youth to become data contributors, creators, and digital entrepreneurs while preserving local languages and heritage, designed to reflect the people it serves.
- Smartly.AI's Moroccan Darija Dataset improves Language AI applications by crowdsourcing and merging pull requests on GitHub.
- First Languages AI Reality supports the revitalization of Indigenous languages with AI and immersive technology.
- IamYourClounon focuses on digitalizing and preserving Beninese languages through creative solutions that promote linguistic diversity across digital infrastructures (for example, a keyboard app that enables users to text more easily in Beninese languages and an emoji app that translates Beninese words into emojis).
- To address the need for lightweight African language models, Lelapa AI developed a small language model named InkubaLM-0.4B, trained for five African languages: IsiZulu, Yoruba, Hausa, Swahili, and IsiXhosa as well as English and French.

» Improve access to digital tools, including translation interfaces and data management systems. Compile commonly used reference content related to civic and economic life into machine-readable, Wiki-style classifications and make these available in multiple languages and formats to support consistent and reliable translations of key terms across systems.

- CERCO, a Benin-based tech company specializing in digital inclusion, has integrated 50 African languages into the operating system of its mobile devices, including a voice control system that increases connectivity among illiterate populations.
- Sign language dataset and generative AI-enabled interfaces are essential for serving the deaf community and sign language learners, particularly as sign languages gain broader official recognition (e.g., in South Africa). The AI4KSL Project led by Maseno University in Kenya develops an automated translator between spoken English and Kenyan Sign Language (KSL), enabled by visual representation of virtual signing characters.

» Invest in human and technical capacity-building to expand and enrich existing NLP ecosystems for inclusive data collection and cataloguing.

- ZINDI is an African continent-wide platform hosting data science resources and educational challenges to help develop new computer science expertise and assist enterprises with data-informed decision-making.

Partnership opportunities include:

- Start-up collaborations across continents, including exchanges among local tech actors and between African and global SMEs, as well as collaborations across and within industries.
  - During the learning journey undertaken as part of the Local Language Partnerships Accelerator Pilot, iGenius and Kytabu shared their experience of collecting, digitizing and integrating low-resource language data with participants.

- More frequent joint projects with civil society organizations, local government, and non-profit actors, facilitating broader participation of linguists and educators and access to existing technical expertise and infrastructures.
  - Clear Global, a global nonprofit specializing in making information available for underserved language communities, collaborated with Digital Umuganda to support digital inclusion for Kinyarwanda speakers.

4. Scale data-sharing and secure data rights

Many international AI efforts rely on unclear data provenance and extractive labour practices. To ensure the responsible scaling of African NLP efforts, informed consent and tracing of data origins throughout the entire data value chain are necessary. Emphasizing fair compensation for data owners and contributors and decent conditions of data work is also essential.

Within this action area, it is important to:

» Test and scale existing solutions in data governance: pilot new data frameworks, licenses or techniques that have been developed but not yet deployed.

- The Esethu Framework for reimagining sustainable dataset governance and curation of low-resource languages.
- Nwulite Obodo Open Data License developed by the University of Pretoria's Data Science Law Lab.
- New thinking on data provenance by Equiano Institute.

» Pilot new platforms and institutions for trusted data-sharing: experiment with data spaces, trusted intermediaries, community trusts or data cooperatives to provide researchers and innovators with responsible options. Adopt tailored approaches for cross-cutting and sector-specific uses.

- Datawise has developed an online repository of datasets and online resources across multiple domains, such as education, agriculture and language.
- As part of the programmatic work of Mozilla Common Voice, Mozilla works with Maseno University in Kenya to pilot the use of new data licenses on their platform.This will allow local communities to maintain more control over the digital resources they have compiled.

- Support and iterate on NLP business models (such as freemium options, crowdsourced language data collection and public–private partnerships) to drive scalable solutions across domains relevant to the Sustainable Development Goals. Sustainable business models are key to ensuring long-term impact.

Partnership opportunities include:

- Collaboration between academics, civil society and technologists to test the implementation of data licenses.
  - UNDP Ghana, University of Ghana, GhanaNLP and Mozilla Common Voice are collaborating to collect voice and text data in Twi under a new data licence.

- New spaces for research that foster open collaboration between researchers across borders and continents.
  - Cohere Labs has launched or supported cross-border collaborations such as Aya, Masakhane, SEACrowd, Singapore AI and UNDP Serbia—promising initiatives that are advancing inclusive language technologies across regions.

A call to scale, sustain, and strengthen Language AI ecosystems

To scale the development and implementation of society-wide digital transformation, it would be valuable for various organizations and institutions to engage in local Language AI ecosystems and beyond. The following section outlines key recommendations and activities categorized by objective and stakeholder group.

Accelerate the availability of resources and know-how for:

Governments

- Act as a **committed partner** to sustainably strengthen emerging Language AI ecosystems, remaining open to new and evolving approaches and emphasizing locally owned initiatives.
- Identify and support **high impact use cases of Language AI** for the public good.
- Support the often **less tangible, yet important, work of local/regional advocacy, regional/global coordination, and knowledge exchange** to enable AI innovators to build the foundations for the successful deployment of Language AI at scale.
- Invest in the **development and maintenance of key language resources** and data governance infrastructures; consider and share **rationales, insights and learnings from proprietary programmes** on preserving Indigenous languages and digitizing under-resourced languages.
- Invest in, develop and maintain **core infrastructure** (including data storage systems, computing capacity, and foundational legal frameworks) in addition to programmatic work—recognizing that the success and sustainability of individual projects depends on the availability and accessibility of these foundational elements.
- Establish and operationalize **robust data governance norms and mechanisms** to recognize, credit, and fairly compensate **local data contributors and community stewards**, especially for non-text, voice contributions and data work.
- Ensure **alignment with existing initiatives** (e.g., the Masakhane AI Hub for African Languages) and **funders** (e.g., the AI for Development Funders Collective).

Philanthropic actors

- Act as a partner for t**esting and scaling approaches** to key levers for impact outlined above, supporting **high-impact use cases of Language AI** for the public good, enabled by robust data governance and data-sharing.
- Support the often **less tangible, yet important, work of local/regional advocacy, regional/global coordination, and knowledge exchange** to enable AI innovators to build the foundations for successful deployment of Language AI at scale.
- Facilitate regional/global **collaboration and learning among grantees** in Language AI, as well as **between grantees and experts** (e.g., from academia and the established private sector) by funding travel and convening opportunities for knowledge transfer.
- Ensure **alignment with existing initiatives** (e.g., the Masakhane AI Hub for African Languages) and **funders** (e.g., the AI for Development Funders Collective).

International private sector

- Contribute to global AI equity and innovation by **sharing know-how, releasing model weights, training data and evaluation data openly** (with appropriate consent).
- Continue to invest in making **infrastructure** accessible to the Global South and in developing AI/data talent globally.
- Leverage **technology available in local languages** to enhance **AI Trust & Safety measures** in the Global South.
- **Partner with local actors** to improve the performance, safety, and applicability of private sector products in emerging markets.
- Ensure **alignment with existing initiatives** (e.g., the Masakhane AI Hub for African Languages) and funders (e.g., the AI for Development Funders Collective).

Angel investors, venture capitalists, incubators, accelerators

- **Invest in Language AI innovators in the Global South**, especially for the development of scalable, context-specific models and tools—acknowledging the potential of **local AI solutions to offer more contextually relevant, accessible and cost-effective applications** for underserved communities. Support ventures that combine sustainable business models with public interest considerations.
- Facilitate regional/global **collaboration and learning among start-ups** in Language AI, as well as **between start-ups and experts** (e.g., from academia and the established private sector).

Academia

- Encourage opportunities for **interdisciplinary knowledge-sharing** and collaboration among university departments; foster cross-sectoral engagements and data-sharing projects with external stakeholders and industry practitioners.
- Conduct, support and share **data** and **research** into issues emerging in the low-resource language digitalization space, including participatory methods, community-sensitive data sharing modalities, computational linguistics, or AI governance and Trust & Safety.
- Empower **youth participation** and agency in research and language digitalization efforts. Support opportunities for students and early-career researchers to shape research agendas and explore practical applications of research outputs.

Accelerate implementation for:

Governments aiming to strengthen their own Language AI ecosystems

- Act as an **ecosystem builder** for Language AI by strengthening the start-up ecosystem (e.g., through small business loans), prioritizing infrastructure investments (e.g., in green computing) and investing in AI/data talent (e.g., supporting university degrees in AI).
- Become **a user** of Language AI by piloting integration into systems administering public services and education. Prioritize the availability of public information and digital public services in Indigenous languages, including text-to-speech features to enable users to receive information in multiple formats.
- Provide **an ethical strategy and regulatory environment** for Language AI, such as integrating the digitization of Indigenous languages into AI and data strategies and developing appropriate data governance frameworks.

Innovators and (co-)implementors

- Continue to **innovate for the public good**, aiming for **scale and sustainability**.
- Invest in communication and **advocacy** to attract collaborators, users, and funders.
- **Join and strengthen regional networks,** including technical collaborations (e.g., Masakhane AI Hub for African Languages) and policy advocacy coalitions.

Delivered together, these actions can chart a path towards closing the language divide in AI. By supporting sustainable growth and the expansion of local and global NLP and Language AI ecosystems, these steps will expand the economic horizons of local language communities and catalyse new digital innovation to accelerate long-term development.

One United Nations Plaza, NEW YORK, NY10017, USA