SOCIAL ENTREPRENEURSHIP PROGRAM

PROJECT SUMMARY

1. **Country:** Paraguay
2. **Project number**: PR-T1373
3. **Name of Project:** GuaranIA: Integrating the Guaraní Language Into The Digital Sphere For The Inclusion Of Rural And Vulnerable Populations
4. **Executing Agency and borrower:** Executing agency: Engineering Center for Technology Research, Development, and Innovation (CIDIT).
5. **IDB División**: Multilateral Investment Fund (MIF) – IDB Lab

6. **Financing amounts**

	IDB US$	Local US$	Total US$
Nonreimbursable Technical Cooperation:	632,400	303,600	936,000

7. **Objective and purpose of the project:**

The objective of the project is to promote the development of generative AI-based services in the Guaraní language for rural and vulnerable populations through data development, model training, and open innovation.

8. **Project Components:**

The project has four components:

Component 1: Diagnostic assessment to identify use cases and priority areas (IDB Lab: The objective of this component is to identify potential use cases in which enabling technologies for communication in Guaraní will allow vulnerable populations to access products or services that improve their quality of life.

The project will set out to conduct a diagnostic assessment using various methods for gathering field data, such as interviews, surveys, focus groups, and workshops where local community experience is used to identify issues where the lack of technological tools with Guaraní language support is a barrier to opportunities that can improve living conditions for people in these communities. In addition, efforts will be made to ensure that these sectors are aligned with the thematic areas of other Bank operations related with digital transformation processes that benefit the vulnerable population in Paraguay.16 The outcome of the diagnostic assessment process will be to identify the sectors with the greatest impact on the population's quality of life, where implementing use cases with computational linguistic resources in the Guaraní language can improve access to products and services in these high-impact sectors. Preliminary opportunities have been identified in the areas of education, health, and financial inclusion.

The diagnostic assessment will be carried out in three of the departments with the highest rates of multidimensional poverty in the country: Caazapá, Caaguazú, and Alto Paraná. It will also aim to include nearby areas where Guaraní-speaking Indigenous communities live, such as in Guairá. The outcome of this intervention will be to contribute knowledge to understand the context of rural populations, Indigenous groups, and other stakeholders, with a view to developing the proposed technological solution.

This component will also aim to connect the beneficiary population to AI tools in sectors directly related to their quality of life in order to familiarize them with this technology and thereby improve their perception of AI and the usability of the tools that will be developed in Component 2. To that end, partnerships will be created with: (i) education centers, to train teachers on how to use AI tools; (ii) family health clinics and other healthcare centers, to train healthcare staff and users; (iii) financial entities; (iv) local governments; (v) academic institutions; and (vi) companies that can facilitate the adoption and deployment of tools and training programs.

The expected outputs of the component are: (i) diagnostic assessment identifying priority sectors/areas for the development of AI pilots, completed; (ii) project communication strategy, implemented; (iii) at least 12 partnerships, formalized, with entities in the public sector, private sector, academia, and civil society for collaboration on project implementation; (iv) project documentation and analysis of results, completed.

Component 2: Development of AI-based technology tools for Guaraní language processing: The objective of this component is to develop computational linguistic resources that facilitate the inclusion of the Guaraní-speaking population in the digital economy.

The first step will be to develop a digital corpus of text in Guaraní that will be documented and published using tools such as datasheets for datasets, as proposed by several researchers.17 To that end, the following actions will be taken: (i) create partnerships with public, academic, and private entities to address the linguistic/Guaraní issue; (ii) compile text documents written in Guaraní using web scraping techniques, queries to digital platform APIs, and written digital material published online; (iii) digitize written bibliographies and transcribe audio in Guaraní; (iv) implement participatory AI techniques18 to compile data using best practices from similar experiences,19 with the involvement of the same community members from the diagnostic assessment and survey work under Component 1, in order to validate and clean up the data collected in Guaraní; (v) consider, as needed, the possibility of translating texts from other languages, generating synthetic data from real texts, and reusing existing text databases in the public domain, such as Jojajovai;20 and (vi) perform ongoing review with expert Guaraní linguists to record and validate the processed materials

The digital corpus will serve as the basis for training the natural language processing models adapted to Guaraní. It will use advanced architectures and AI foundation models21 that already have strong comprehension of Spanish and other languages, which will benefit Paraguayan Guaraní. The models will be modified to follow instructions in Guaraní and Jopará (a variant of Guaraní mixed with Spanish), thereby optimizing their ability to perform specific tasks such as translation, text generation, and classification, as well as their interactions in different applications, like chatbots or virtual assistants.

The model’s performance will also be evaluated in mixed-source settings (Guaraní-Spanish), with a view to handling Jopará more efficiently. Lastly, all of the models and tools that are developed will be released in open-source repositories under the GPLv3 license,22 which will facilitate access for local and international developers and researchers committed to the digital inclusion of the Guaraní-speaking community. The models will be published using common practices in the field, such as model cards23 for reporting documentation

During the development of the corpus and the training and evaluation phases, state-of-the-arttechniques,24,25 recommendations, and best practices26,27 will be implemented to promote responsible use of the tools and help prevent inappropriate results from the models, including hallucinations and the production of racist, misogynist, discriminatory, or offensive content that amplifies or reinforces social biases. Efforts will be made to mitigate the environmental impact of the development and operation of technology, incorporating, whenever possible, energy-efficient computational practices, the use of renewable energy sources, and modular development, preventing unnecessary energy consumption. The expected outputs of the component are: (i) digital corpus in Guaraní, validated, documented, and published under license GPLv3; (ii) natural language processing models trained on the corpus developed; (iii) report on responsible development of language processing models and corpus, documented and validated.

Component 3: Implementation of the AI tool through pilot tests in communities: The objective of this component is to implement the computational linguistic resources developed under Component 2 in software products and apply demonstration use cases that improve the quality of life of the vulnerable populations and Indigenous groups identified in Component 1. On the basis of the general corpus created in Component 2, specific corpora of text will be developed, documented, and published for each of the identified use cases. They will incorporate texts in Guaraní related to the topics and, more specifically, the use cases. As a result, specialized language generation models will be created in the areas related to the use cases identified in Component 1.

These specialized models will be published under the same license and will follow the same practices used to document the model created in Component 2. The specialized models will be incorporated into software modules to facilitate their application by third parties and especially their use in the use cases identified in Component 1. These modules will be deployed, maintained, and updated on online servers with high computing and storage capacity. The models' source code will be made available in open-source models under the GPLv3 license.

During implementation of this component, the software modules will also be used to develop mobile applications for low-end devices (in terms of processing capability, storage, resolution, and operating system version). In addition, these applications will be required to operate with minimal reliance on the Internet and minimal training for lay technology users. The functional requirements of the applications will depend on the use cases identified in Component 1. Development will follow an agile approach such as lean software development, and the source code will be made available in open-source repositories under the MIT license, which allows for use, improvement, and evaluation, free of charge, by the technical and scientific community, as well as proprietary use by the private sector.

The applications will be rolled out gradually in three phases. The first phase will consist of performing technical and functional tests in controlled environments within the development team. The applications will be corrected and improved based on the results of this phase. In the second phase, usability tests will be performed in real environments following a hybrid moderation approach that includes in-person and remote sessions. Appropriate usability improvements will be made at the end of this phase.

The third phase will focus on limited pilot studies to understand the solution's potential impact on the target population, which will be measured by experimental studies. To that end, collaboration may take place with actors from the priority areas identified in Component 1 who are present on the ground supporting potential target populations. The elements of the impact study, such as the experimental group, control or baseline group,28 hypotheses, scope, experimental conditions, dependent and independent variables, and types of measurement will be determined by the type of intervention performed for each of the use cases and beneficiary populations.

The expected outputs of the component are: (i) at least three agreements for the implementation of use cases, formalized; (ii) three specialized language generation models in the domains related to the use cases identified in the component, developed; and (iii) three pilot use cases, implemented and evaluated.

Component 4: Strategies for open innovation and training to develop products and/or services (IDB Lab: US$75,000; Counterpart: US$130,000). The objective of this component is to identify additional public and private institutions that can adopt the tools developed by the project, which would expand the scope of the products, services, and applications using this technology and increase the number of people and communities accessing GuaranIA linguistic resources.

As part of execution of this component, one or more calls for proposals will be held to incorporate the technological solutions developed by GuaranIA into services and applications where inclusion of the Guaraní language provides an opportunity for the Guaraní-speaking population to enjoy equal access to products, services, and new use cases offered by the public and private sectors.

These calls for proposals for open innovation will consider three lines of work for the adoption of GuaranIA tools: (i) Line A: development of new use cases or new applications, execution of experiments to adopt these cases or applications, and feedback from the target population; (ii) Line B: integration of an existing GuaranIA application that has been adapted to provide a service, execution of experiments to adopt this application, and feedback from the target population; and (iii) Line C: other use cases, i.e. proposals that do not fall under Lines A or B but are relevant to the scope of the call for proposals. The calls for proposals will be aimed at startups, companies, the public sector, academia, research centers, and civil society organizations. Some preliminary criteria have been identified for the selection of participating entities, such as feasibility, percentage of cofinancing from the proposing parties, potential for social impact, outlook for sustainability and scale, and relevance of the product or service to the quality of life of the vulnerable, Guaraní-speaking population.

Under this component, a scaling strategy will be developed to ensure the

sustainability of the technological products that arise from the GuaranIA project over time.

Lastly, the component will help boost the capacities of stakeholders in the public and private for the responsible development and adoption of the AI tool, with deep modules for agents involved in innovation, research and development, and the impact startup ecosystem. To that end, training cycles will be held on topics such as: (i) the principles of responsible development of AI projects; (ii) risk assessment in data science and AI projects; and (iii) technological integration of AI tools. As an indirect outcome, the training cycles will lead to the creation of a multidisciplinary task force for the responsible use of AI in the country.

The expected outputs of the component are: (i) call for open innovation for the inclusion of the Guaraní language in the digital economy, implemented; (ii) at least three technology integrations prepared, implemented, and evaluated; (iii) 60 people trained for the responsible adoption of tools developed and integrated; (iv) strategy for sustainability and scaling of use and integration of outputs developed, documented.

9. **Project Beneficiaries**:

The beneficiaries of the project will be at least 500 vulnerable individuals29 engaged in the use cases developed and implemented by the project in sectors that directly impact their quality of life, such as education, health, and financial services.

10. **Expected results and capture of benefits:**

The expected outcomes of this project will be: (a) at least 500 vulnerable individuals who benefit from the implementation of the use cases developed by the project; (b) a language and AI technological tool for the Guaraní language that is developed, published, and made available; (c) at least three new services offered to populations under the Multidimensional Poverty Index living in rural and periurban areas of Paraguay; and (d) at least three initiatives where the technological solutions developed by GuaranIA are incorporated into new services and applications for the inclusion of the Guaraní language in the digital economy.