Games for Learning Endangered Languages: A Pilot Project on Four Languages from the Tupian Family

Gábor Bella[1], Gessiane Picanço[2], Alethia Hume[3], Adriano Clayton da Silva[4], and Yannis Haralambous[1]
[1] UMR CNRS 6285 Lab-STICC, IMT Atlantique, France
[2] Faculdade de Letras, Universidade Federal do Pará, Brasil
[3] Universidad Católica 'Nuestra Señora de la Asunción', Paraguay
[4] Universidade Federal do Amazonas

The broad 'endangered' category includes languages with varied socioeconomic backgrounds, community sizes, degrees of literacy, levels of digital presence, and even attitudes towards language use. Yet the lack of supporting materials for language teaching and learning—books, exercise sheets, spoken media, interactive tools, games—is frequently seen as a major hindrance to transmitting the language to younger generations, despite the eagerness of speaker communities to do so.

Locally grounded efforts are most likely to produce learning materials that are meaningful to communities. They are, however, often curbed by economic or political difficulties due to marginalisation and minority status. Meanwhile, jubilatory promises of 'no languages left behind' [1] or 'language technology for all' [2] by researchers and market actors from the Global North fail due to market logic, lack of training corpora [3], and the misalignment of remotely-set R&D agendas with local needs [4].

Without any pretension to 'solve' the problems outlined above, our long-term aim is to help build games and exercises for language learning and teaching, destined to communities to whom such material is only scarcely available, if at all. Our goal is to provide a method, resources, and software infrastructure for the semi-automated yet locally grounded creation of such material. Automation is crucial for scaling across multiple communities and for creating a varied pool of learning materials; however, it must be aligned with local community needs, which we articulate through the following adaptation requirements.

1. *Adaptation to the local language and culture.* A language is more than a lexicon and a grammar: it is tied to and is shaped by the local environment, culture, social structures, etc. The learning material must reflect this embeddedness and remain coherent with the local context.
2. *Adaptation to technical limitations.* Firstly, automation cannot rely on generative AI due to the lack of training corpora; instead, symbolic resources such as semantic and lexico-semantic networks, generative grammars, or grapheme-phoneme mappings need to be used. Secondly, the resulting games and exercises need to be usable without network access, and must be available on multiple platforms, including smartphones and paper-based versions in the absence of electricity.
3. *Adaptation to pedagogical requirements.* The exercises should suit students' varying levels of proficiency in the locally spoken languages. Also, language teachers should be able to correct, adapt, extend, or otherwise customise the automatically generated material.

Beyond adaptation, the deployment of the process for actual communities must respect ethical principles of co-creation, as well as the rights of these communities to ownership of and unlimited access to the data they provide and to the project results [9]. The effort should take into account the experience obtained from existing projects and platforms for game-based language learning, a few examples of which are the multilingual online platform wordwall.net, desketa.bzh for Breton, research output such as [8], but also our own existing proof-of-concept demonstrator on lexical games.

We are initiating work following these principles in the context of the recently launched two-year AILAR project, which involves four communities from South America belonging to the Tupian language family: the Sateré-Mawé, the Mundurukú, and the Asuriní do Xingu from Brazil, but also the Guaraní language, a minority language with 6.5 million speakers and an official status in Paraguay. The contact points with these communities are local researchers who speak the languages and who are partners in AILAR. While the project does not fund fieldwork, it includes initial discussions with community representatives who articulate their needs and provide background information to inform the three

adaptation mechanisms above. The project also includes the curation and extension of existing language resources, game development, and multiple stages of community feedback. It benefits from prior work on Tupian language resources [5,6] and from existing, small but curated and meaning-aligned Amerindian lexicons. These lexicons also include evidence of lexical untranslatability [7], on which we rely for a deeper adaptation of learning tools to the local semantic space.

The talk will address the community needs outlined above, present first ideas of solutions towards the goal, introduce the AILAR project and the results obtained so far, and outline the research that needs to be carried out.

[1] Costa-Jussà, Marta R., James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672* (2022).

[2] Mariani, Joseph J. Language technology for all: a challenge. *UNESCO Report on Languages* (2020).

[3] Opitz, Juri, Shira Wein, and Nathan Schneider. Natural language processing relies on linguistics. *Computational Linguistics* 51, no. 3 (2025): 1009-1032.

[4] Schwartz, Lane. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 724-731. 2022.

[5] Rodríguez, Lorena Martín, Tatiana Merzhevich, Wellington Silva, Tiago Tresoldi, Carolina Aragon, and Fabrício F. Gerardi. Tupían language resources: Data, tools, analyses. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pp. 48-58. 2022.

[6] Larroyed, Aline, Adriano da Silva, and Sharon O'Brien. Unmasking Indigenous Invisibility: Empowering AI-Driven Translation with Indigenous Participation. In *The Social Impact of Automating Translation*, pp. 77-95. Routledge, 2024.

[7] Bella, Gábor, Erdenebileg Byambadorj, Yamini Chandrashekar, Khuyagbaatar Batsuren, Danish Cheema, and Fausto Giunchiglia. Language diversity: Visible to humans, exploitable by machines. In *Proceedings of the 60th annual meeting of the Association for Computational Linguistics: system demonstrations*, pp. 156-165. 2022.

[8] Aref Abedjooy, Fatemeh Hirbodvash, and Mehran Ebrahimi. Indigenous Language Revitalization with Stories and Games. *Encyclopedia of Computer Graphics and Games*, Springer, pp. 952–957, 2024.

[9] Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons et al. "The CARE principles for indigenous data governance." *Open Scholarship Press Curated Volumes: Policy*, 2023.