RISE UP

REVITALISING LANGUAGES AND SAFEGUARDING CULTURAL DIVERSITY

Grant Agreement: 101095048

Guidelines for discussing Generative AI for minoritised languages

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

1 Introduction

Generative AI (GenAI) is rapidly reshaping the way people use language in their everyday and professional lives. It powers tools for guidance, administrative support, translation, idea structuring, language learning, and many other services. Foremost, its responsible use requires freedom of choice in whether and how AI is used, alongside transparency, protection of privacy and intellectual property, and clear human responsibility for the results. It also requires a precise understanding of GenAI's limitations and shortcomings.

In 2026, GenAI gives qualitatively acceptable results in only a limited set of languages. Users of minoritised languages are either excluded, encounter barriers or are forced to interact in another language. This dynamic deepens the 'digital divide' further weakening minoritised languages by restricting their use and excluding them from a broad range of communicative practices. GenAI can represent a significant opportunity for speakers and learners, promoting inclusiveness and social justice. AI can support the use and learning of minority languages by providing personalised exercises, conversation partners and safer spaces for interaction – advantages that can increase motivation and autonomy when materials and conversation partners are scarce. However, ensuring AI compatibility for minoritised languages must remain a voluntary, community-driven choice.

If communities choose to adopt AI, it is of utmost importance to build open, accessible, continuously updated and protected corpora of written and spoken language under open licenses, making minoritised languages available for digital development. To this end, public, scientific and societal efforts must be mobilised. The project RISE UP has collaborated with minoritised language users and experts in the domain of Natural Language Processing (NLP) to develop a set of recommendations for key actors who engage in discussing Generative AI for minoritised languages.

2 Recommendations for researchers

- Respect and support community decision about whether the inclusion of the respective minoritised language in GenAI systems is acceptable and, if so, for which purposes.
- Serve as facilitators and bridge-builders with a community-centred perspective; collaborate across disciplines and with local linguistic expertise to align tools with actual language use and community priorities.
- Raise awareness of how existing resources can support digital development - for example, securing or updating an ISO 639-2 (or -3) code; enabling open licensing and community-driven data collection via platforms such as Common Voice; and leveraging lawful text and data mining by research organisations and cultural heritage institutions under the EU Directive 2019/790.
- Build bridges between communities and developers; prioritise task-specific models with realistic distribution pathways that can integrate feedback and be evaluated with or by community members.
- Highlight relevant principles and laws (e.g., data protection and intellectual property), and ensure accurate citation practices, including recognition of data workers; follow and promote open science and FAIR principles.

- Increase the visibility and machine-actionability of community resources by providing language tags, rich metadata, documentation, and accessible archive deposits for both academic and non-academic users.
- Involve community expertise in funded research projects; include community associations as consortium partners; adopt inclusive, diversity-oriented hiring practices.

3 Recommendations for community members

- Consult the widest possible range of community members to decide whether the inclusion of the respective minoritised language in GenAI systems is acceptable and, if so, for which purposes.
- Negotiate a shared framework for AI initiatives: agree on the language name(s), ISO 639-2/639-3 identifiers, representation of varieties/dialects, and principles for inclusion, evaluation, and governance.
- Identify solutions tailored to your language ecosystem to secure copyright-free or rights-cleared audio sources (e.g., podcasts, community radio, videos), and document permissions to enable text-to-speech, forced alignment, and automatic speech recognition pipelines.
- Collect and curate resources (text, audio, metadata); prioritise lawful, consent-based, and contextually annotated data that reflects real-life use, and release it under open terms where feasible with clear licenses and provenance.
- Build collaborative research partnerships with local and international NLP researchers and linguists, including data-sharing agreements, transparent evaluation protocols, and capacity-building opportunities for community members.

4 Recommendations for stakeholders in language policy

- Respect and support community decision about whether the inclusion of the respective minoritised language in GenAI systems is acceptable and, if so, for which purposes.
- Support applications or updates for ISO 639-2 language codes to improve findability and inclusion in platforms and datasets.
- Invest in small task-specific engineering projects (e.g., language recognisers, baseline text-to-speech/automatic speech recognition/aligners) to unlock practical functionalities with limited data and demonstrate early value.
- Enable participative and open science methodologies; favour and promote NLP literacy of local language experts and linguists.
- Provide training and governance to promote responsible sharing and machine-actionable metadata with trusted repositories (e.g., deposit campaigns on Zenodo).
- Provide and facilitate training on responsible generative AI use (verification, privacy, bias mitigation), and encourage transparent disclosure without adverse effects.
- Where possible, implement governed, locally hosted or secured cloud-based AI tools to protect data sovereignty.

The inclusion of minoritised languages in GenAI must remain a voluntary decision, allowing communities to decide for themselves whether and how AI is used. Ensuring AI compatibility for minoritised languages requires collaboration: communities, researchers and stakeholders develop resources together in interdisciplinary, participatory processes and on the basis of sustainable partnerships. FAIR data follows best practices – transparent disclosure, protection of privacy and intellectual property, clear human accountability, training, and careful attention to limitations. When these conditions are met, GenAI contributes to the learning and use of minority languages, supporting revitalisation and reclamation.

Links

RISE UP Online Event on AI for minoritised languages: https://www.riseupproject.eu/events/panel-discussion-ai-for-minoritised-languages/

ACL Anthology: https://aclanthology.org/