Lacuna Fund
Our voice on data

Lacuna Fund: Our Voice on Data

Resources for Proposals in African Languages

This document represents a collection of resources from the Technical Advisory Panel as an addition to those referenced in the RFP document. These are intended to provide assistance in obtaining relevant background information, preparing a competitive proposal, and completing quality work.

These resources are not intended to be exhaustive nor authoritative. This document does not represent an endorsement of work by the Lacuna Fund Secretariat, the TAP, or individual members.

PREVIOUS WORK AND RELEVANT BACKGROUND

Relevant recent challenges and other efforts:

- Papers from the recent International Conference on Learning Representations (ICLR) AfricaNLP workshop. (There may also be upcoming workshops at additional conferences)
- Widening ML Workshop at ACL.
- Session notes from the Africa NLP Unconferences.
- AACL-IJCNLP Workshop on Technologies for MT of Low Resource Languages
- Both rounds of the AI4D-Zindi African Languages Challenge, including submitted datasets.
- Common Voice, including ongoing efforts to create datasets for Luganda and Kinyarwanda.
- Masakhane, a grassroots African initiative to improve NLP in African languages. The group is undertaking many efforts related to African NLP.

COMPILATIONS OF RESOURCES AND EXISTING DATASETS

- See papers from the recent International Conference on Learning Representations (ICLR) AfricaNLP workshop for information on active efforts and key considerations in a variety of languages. (There may also be upcoming workshops at additional conferences)
- Maskhane's website (maskhane.io) has a strong listing of resources and existing efforts in many languages.
- Search LREC, OPUS, and other existing repositories for datasets in languages of interest.

GENERAL CONSIDERATIONS AND THE STATE OF THE FIELD

- Martinus, Laura, and Jade Z. Abbott. "A Focus on Neural Machine Translation for African Languages." *ArXiv:1906.05685 [Cs, Stat]*, June 14, 2019. http://arxiv.org/abs/1906.05685.
- Nekoto, Wilhelmina, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, et al. "Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages." *ArXiv:2010.02353 [Cs]*, October 5, 2020. http://arxiv.org/abs/2010.02353.

- Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. "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*, 6282–93. Online: Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.acl-main.560.
- Tracey, Jennifer, Stephanie Strassel, Ann Bies, Zhiyi Song, Michael Arrigo, Kira Griffitt, Dana Delgado, et al. "Corpus Building for Low Resource Languages in the DARPA LORELEI Program." In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, 48–55. Dublin, Ireland: European Association for Machine Translation, 2019. https://www.aclweb.org/anthology/W19-6808.
- Ruder, Sebastian. "Why You Should Do NLP Beyond English,"2020. https://ruder.io/nlp-beyond-english/.
- Neubig, Graham. "The Low Resource NLP Toolkit: 2020 Edition" http://www.phontron.com/slides/neubig20africanlp.pdf. Presented at the Second AfricaNLP Workshop at ICLR 2020.

This is a rapidly evolving field, and new datasets and models are published almost weekly.

PRIVACY AND ETHICS

- See Open Data Institute's (ODI) Data Ethics Canvas as a helpful resource to consider ethical issues in a proposed project.
- ODI's privacy and openness principles for personal data.

OTHER RESOURCES ON OPEN DATA

- The Beijing Declaration on Research Data
- The Big Secret in the Academy Is That Most Research Is Secret: The dangerous rift between open and classified research, Spring 2020
- "Legal and Ethical Issues around Incorporating Traditional Knowledge in Polar Data Infrasrtuctures" Data Science Journal 16(1)pp1-14.
- The work of Africa Digital Rights Hub.