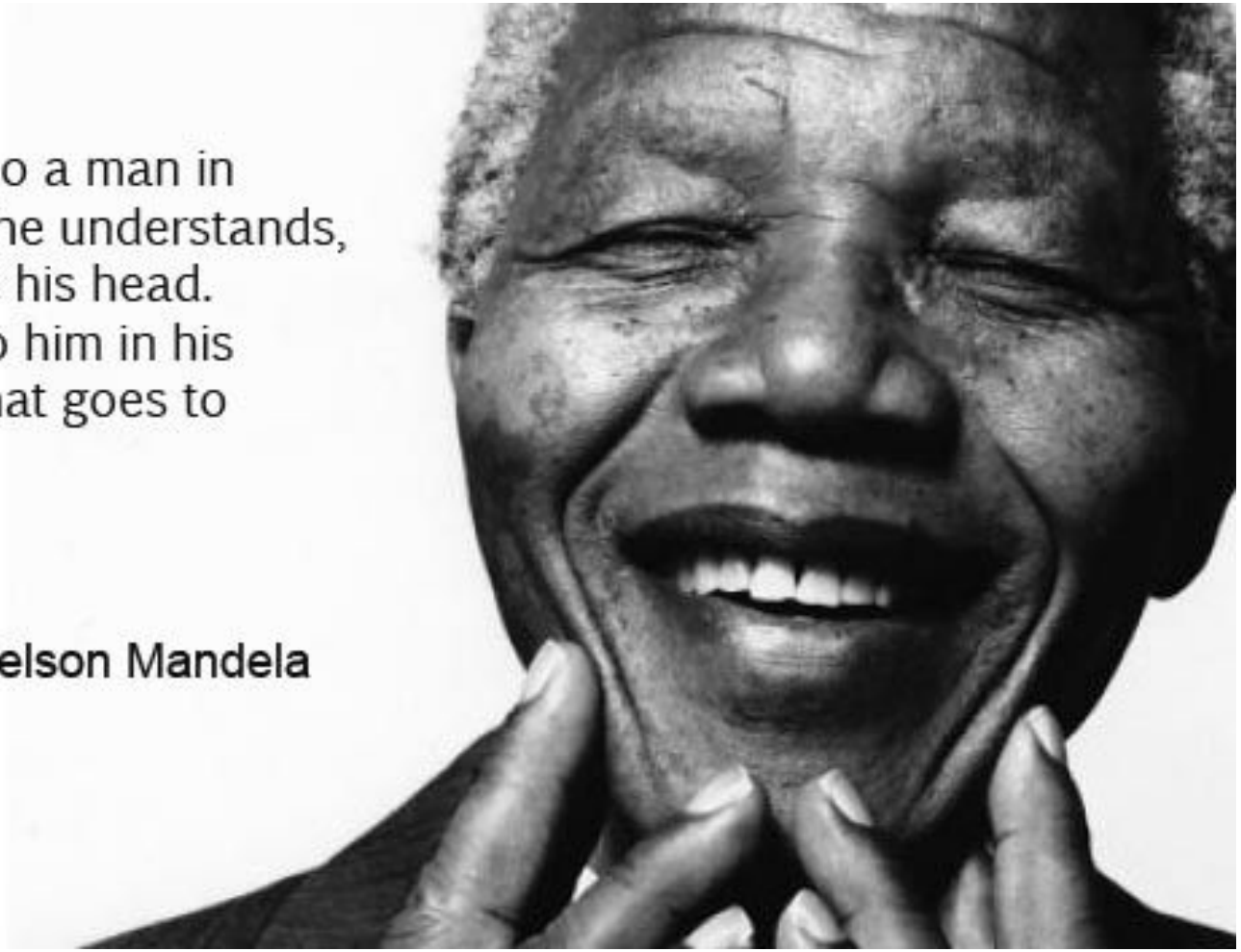
Our voice on data

Why Fund in Language?



Why Fund in Language?

Why Fund in Language?

DRAGON NATURALLYSPEAKING

SAMSUNG

NUANCE

amazon alexa

Google Voice

IBM

However,

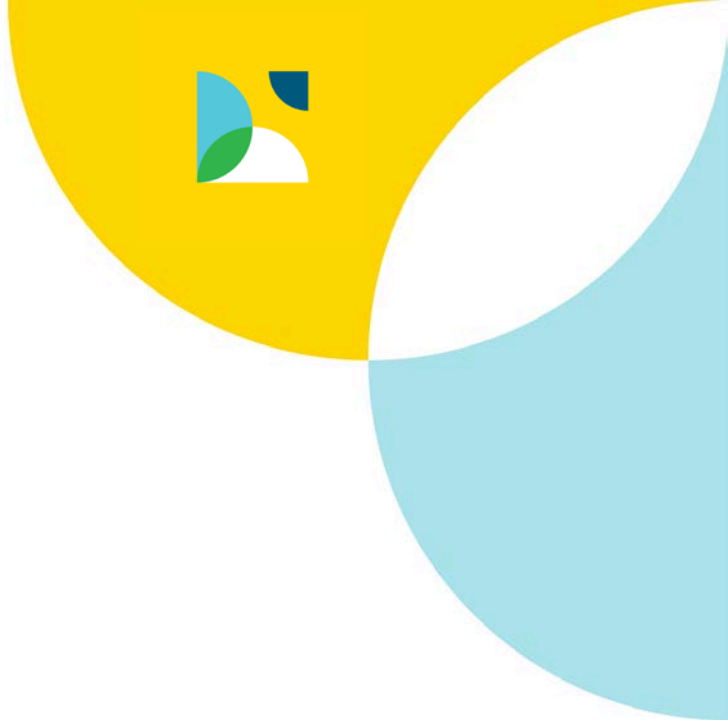
- while “language technologies contribute to promoting multilingualism and linguistic diversity around the world. However, only a very small number of the over 7000 languages of the world are represented in the rapidly evolving language technologies and applications.”

  Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, Monojit Choudhury, **”The State and Fate of Linguistic Diversity and Inclusion in the NLP World”**

- Inclusion in a digital future
- Access to enabling/assistive technology

Vision

“What do we hope to achieve through this RFP?”



Transformative datasets that address common needs and move NLP in Africa forward.

Tell **us** about the dataset **you** see a need for, why **you're the team** for the job, how you'll deliver a **quality product**, and **who is likely to benefit** from it.

We fund **datasets**, not end-to-end solutions.

Think of your proposal as a proposal on behalf of a broader community that will use the dataset you create.

Eligibility

“Can I apply?”

Applicants should:

- Be non-profits, research institutions, for-profit organizations, or teams of such organizations.
  - There can be only **one "lead" applicant** – partners should be structured as sub-contractors.
  - Mission related to social good
- Be **headquartered in Africa** or have a **substantial partnership** with an organization(s) headquartered in Africa.
- Have necessary expertise and be able to obtain national approvals.

Needs

“What are you funding?”

We see needs in:

- Speech corpora for ASR
- Labeled and unlabeled text corpora for use as training data
- Parallel corpora for machine translation
- Corpora to support fundamental NLP tasks
- Datasets for key downstream NLP tasks
- Datasets to improve the performance of NLP tasks on code-switched text or speech

We see needs in:

- Augmentation of existing datasets in all areas to decrease bias or increase the usability of NLP technology in low- and middle-income contexts
- More benchmark data for NLP tasks in underserved languages or to inform multilingual models
- Innovative datasets, such as video or audio captioning or other image-text interactions
- Domain-specific creation or augmentation of text and speech datasets that enable applications with significant social impact
- *Your idea here!*

Selection

“Should I apply?”

“What are you looking for?”

Proposals should:

- Demonstrate the team's qualifications and knowledge of data quality and standards needed for the use case.
- Clearly articulate what gap the new dataset will fill and what advances it will enable.
- Be feasible to complete within 6-12 months, considering the impacts of COVID.
- Range from $10-50k for small to medium projects or up to $250k for large, transformational projects.

Proposals should:

- Consider how to ensure collected data is representative and to the extent possible, unbiased.
- Ensure compliance with and specify relevant copyright or IP law for any source text.
- Use existing or common infrastructure.
- Consider innovative and cost-effective means to collect and maintain quality data.
- Demonstrate awareness of and attempt to be interoperable with related, existing datasets.
- Ensure your metadata is well-described.

Criteria

- **Quality** of proposed methods and team qualifications
- Potential **transformational impact** of the dataset
- **Novelty**, accounting for and/or using existing datasets and methods
- **Equity** dimension of proposed use cases
- **Feasible** in relation to budget, timeline, and scope proposed

Criteria

- Be **open source** (with very limited exceptions) and be widely accessible
- Follow **ethics and privacy** best practices
- Be **locally** developed and owned
- Plan for future **sustainability**

Application Process

“Now what?”

Application Components

- Applications are accepted only through the **portal** (access through website “Apply” page)
- Apply in either **English** or **French**
- Completed application includes
  - Information form
  - Proposal narrative
  - Budget (including budget assumptions) and deliverable timeline
  - Due diligence self-reporting
  - Letter(s) of support (if using existing materials as data source)

Timeline

23 SEPT
RFP Release
2 OCT
Webinar
7 OCT
Q&A deadline
14 OCT
Answers posted
6 NOV
Proposals due
EARLY DEC
Notifications to applicants
DEC/JAN
Work begins
6-12 MONTHS LATER
Datasets released!

See the RFP for specifics!

It is the **authoritative source of information** on the call. The document has further details on exact eligibility and selection criteria as well as specific information to include in your proposal narrative.

Upcoming Opportunities

- In addition to the open language call, the Fund anticipates releasing additional RFPs in the first half of 2021.

Join the newsletter on the Fund’s website to be notified when both are released.

Questions?