Timeline

FAQ

Winners

Update as of June 2020: The datasets **MUST** be for languages indigenous to Uganda, Ghana, or South Africa. Any other languages will not be evaluated.

Our intention is that the datasets are kept free and open for public use under a Creative Commons license 4.0 or similar. Data already licensed under more restrictive terms will not be eligible.

The evaluation of datasets will be done by an expert committee and will take into consideration the following criteria:

Representative and Balanced (40%)

A corpus should be representative and balanced with respect to particular factors; for example, by genre—newspaper articles, literary fiction, spoken speech, blogs and diaries, and legal documents. A corpus is said to be “representative of a language variety” if the content of the corpus can be generalized to that variety (Leech 1991). Basically, if the content of the corpus, defined by specifications of linguistic phenomena examined or studied, reflects that of the larger population from which it is taken, then we can say that it "represents that language variety.”

The notion of a corpus being balanced is an idea that has been around since the 1980s, but it is still a rather fuzzy notion and difficult to define strictly. Atkins and Ostler (1992) propose a formulation of attributes that can be used to define the types of text, and thereby contribute to creating a balanced corpus.

- Two well-known corpora can be compared for their effort to balance the content of the texts. The Penn TreeBank (Marcus et al. 1993) is a 4.5-million-word corpus that contains texts from four sources: the Wall Street Journal, the Brown Corpus,

> ATIS, and the Switchboard Corpus. By contrast, the BNC is a 100-million-word corpus that contains texts from a broad range of genres, domains, and media. For your own corpus, you may find yourself wanting to cover a wide variety of text, but it is likely that you will have a more specific task domain, and so your potential corpus will not need to include the full range of human expression. The Switchboard Corpus is an example of a corpus that was collected for a very specific purpose—Speech Recognition for phone operation—and so was balanced and representative of the different sexes and all different dialects in the United States. (ref)

Proposal for building the dataset (30%)

the researcher credibly demonstrates that given a research grant of $1,500 in addition to a $500 upfront prize (for a total of $2,000), the resulting dataset will be delivered in a reasonable timeframe and perform well against the same criteria (representative and balanced, annotated for a specific downstream task, number of tokens, and underrepresentation of the language). This score will be based on the plan for building the dataset in the future as articulated in the documentation as well as other indications of the researcher's commitment and understanding of the project from the submitted dataset and documentation.

Annotated for a specific downstream task (10%)

The dataset should be designed to enable certain downstream tasks. Any downstream task will be considered and no preference will be given to one over the other. The following are examples of downstream tasks that we would be interested in seeing:

- Machine translation
- Question answering
- Sentence classification
- Sentiment analysis

Number of tokens and number of unique tokens (10%)

Is the language underrepresented on the Internet and in terms of available digital data? (10%)

Competitions

Hackathons

Data Scientists

Discussions

Jobs Board

Host competition

About Us

Our Partners

Contact Us

Terms of Use

Privacy Policy

FAQs

LinkedIn

Facebook

Twitter

Instagram