An Open Source System for Crowd Sourcing an African Language Short Story Corpus

Muite, Benson K.
Kichakato Kizito
benson_muite@emailplus.org

Abstract

Many African languages have few open access corpora for use in developing technological applications such as grammar checkers, spell checkers, speech to text, text to speech and machine translation tools. This may lead to a decline in all cultural traits associated with the peoples that speak these languages. To enable collection of textual corpora, and long term preservation of positive cultural characteristics, the design considerations and implementation of an open source online short story competition collection and evaluation system are described. The system is written in PHP, and can be relatively cheaply deployed on shared hosting servers available from many African hosting providers. This allows for the possibility of a decentralized collection of stories, as well as adaptation, and improvements of the software to different types of short story competitions. The software has been used for two short story competitions across the African continent with the aim of providing stories suitable for children. Holding the competition online has enabled participation from a wide variety of locations, but most of the submissions have came from African countries with relatively good information technology infrastructure. Preparation for a third competition is in progress.

Keywords: Crowd sourcing, African Literature, Corpus Creation

1 Introduction

Africa has a great variety of languages and language dialects. Many of these have rich oral traditions, and have only recently been used for written electronic communication, such as social media, email and internet forums. Many African countries have chosen to use English or French as the language of education, however the internet has democratized access to spaces for public expression, and with this freedom, people have begun to express themselves in their own languages - the Indigenous Tweets project (Scannell 2021) documents the use of languages that are not widespread on one social media platform. Expecting this trend to grow for at least some African languages, there will be interest and need for digital language processing tools for African languages. To develop such tools, language data is required. One way to collect this data is by crowd sourcing. Complementary multilingual short story collection efforts include StoryWeaver in India (Story Weaver 2021) and the African Story Book in South Africa (Saide 2021). Neither of these at present has openly available software for other people to use in crowd sourcing short stories, and they only focus on literacy, without much effort made for use of the collected materials to enable the development of natural language processing tools.

2 Crowd Sourcing Platform Design Considerations

While most people in Africa, uses telephones for communication rather than computers, internet access is growing. Websites enable easier communication for longer pieces of writing such as short stories, and so are a good first platform choice. Further considerations for a web based platform are:

- Many of Africa's languages are closely linked to ethnic identity. This can be a source of tension and has caused conflict. The platform should allow anonymous submissions and/or submission by a proxy.
- Many African countries have a strong focus on English and French as languages of education. Translation from African languages has also primarily focused on translating to English or French, yet many African languages are closely related and may have concepts not well expressed in English and French, so it is important to develop corpora that enable cre-

ation of translation tools between African languages.

- Many authors and translators making a submission may not know English, so the submission platform should be available in several languages, ideally every language submissions can be made in.
- Since many African languages are not used in or studied at school, the language ability of people who submit may not be very high, but they may be interested in improving it. Links to available digital resources should be made available and efforts should be made to allow for feedback on submissions.
- The software should be easy and inexpensive to deploy within Africa to allow for independent organization of short story competitions.
- The software should and be easy for software developers based in Africa to contribute to.
- The software should as much as possible respect privacy of people submitting stories and translations, and enable them to keep personally identifying information private should they choose to do so.

3 Crowd Sourcing Platform Implementation

The first consideration is the programming language to use. Python (Python Software Foundation 2021), Ruby (Ruby Community 2021), PHP (PHP Community 2021) and Javascript (ECMA International 2021) were considered. Python and Ruby have great use in web development, however they have poor support for shared hosting in Africa and typically require higher specification servers than PHP - shared hosting support can be improved, but in the time frame of this project, this was not feasible. Javascript is also used for developing the server side of a web application, in addition to making the client side of a web application interactive, however, it again has much less support for shared hosting in Africa. Thus, PHP was chosen as the programming language to use for server programming, with some amount of Javascript, in particular to create soft keyboards. Many cloud providers in Africa use the software cPanel (cPanel LLC 2021) for managed shared hosting with upto 10 GB of storage for prices between $1 and $5 per month. This has support for hosting a PHP web application and a database.

The software consists of:

- A database of original stories along with the contact details of the author or their proxy
- A database of translated stories, with the contact details of the translator or their proxy
- A database of authors, translators or their proxies and the languages in which they can vote
- A database with votes for and comments on each original story or translated story
- A server side PHP program to allow for stories and translations to be submitted
- Client side HTML and Javascript components to enable input of special characters using a soft keyboard and translation of the website, submission of stories, submission of translations, viewing of stories and voting

The website can be viewed at `https://tuvutepamoja.africa`. Figures 1 and 2 show the front page of the website in English and in Kiswahili. Figure 3 shows a soft keyboard suitable for entering Yorùbá which has a number of accented characters which many people who are not trained typists need to see to be able to use in their writing.

The software was openly developed using version control on `https://notabug.org/tuvutepamoja` and is released under the GNU general public license (Free Software Foundation 2021*b*). The repository `https://notabug.org` hosts the development of GnuSocial (GnuSocial Community 2021), an open source social networking site also written in PHP which can be translated into African languages. An attempt was made to use GnuSocial to enable communication between

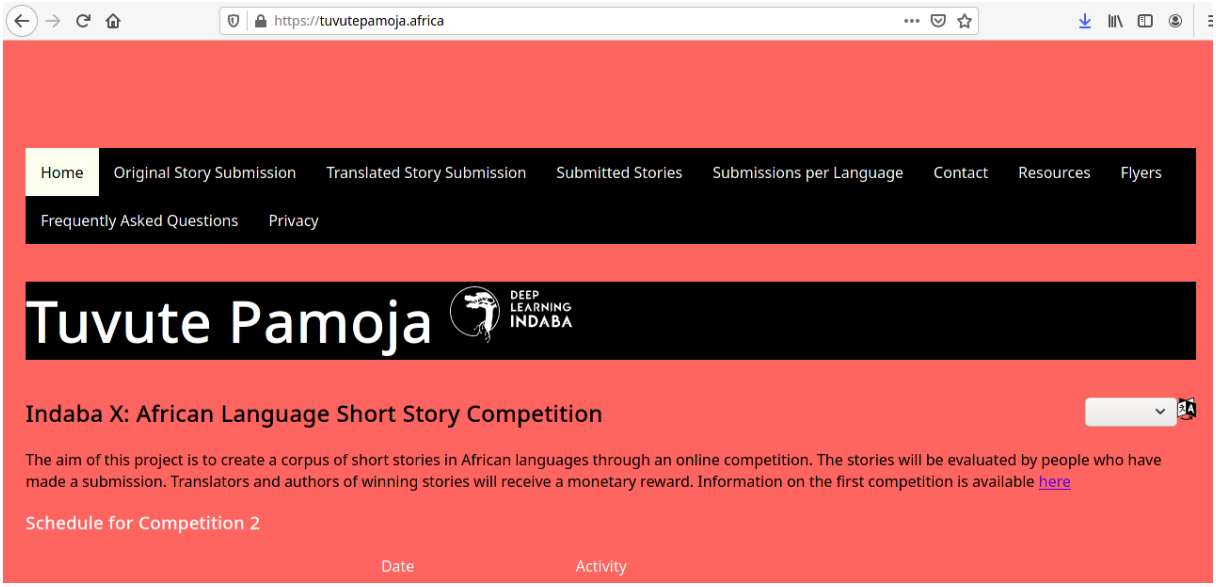

Figure 1: English front page.

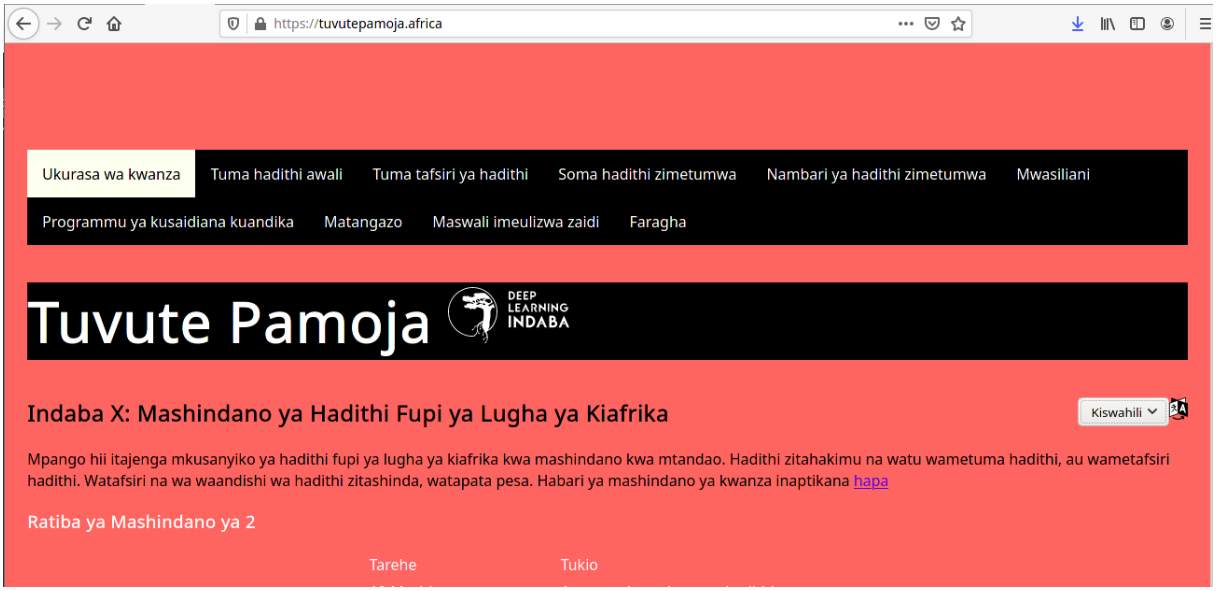

Figure 2: Kiswahili front page.

people interested in promoting the competition in African languages that they know, however engagement on this platform was low. At present an online forum using the open source software FluxBB (FluxBB Community 2021), is being used to enable public communication by people interested in the project. The forum is available at `https://ingxoxo.tuvutepamoja.africa/`.

For many possible contributors, translation of software is not as exciting as writing their own short stories, so this has not yet been successful. Translating flyers advertising the short story competition was less of a task and many people contributed to this effort, enabling use of social media platforms

Figure 3: Soft Keyboard - Yorùbá shown.

by people willing to promote their language.

4 Further Work

Word limits were used to indicate the length of acceptable submissions. However, a word in one language may correspond to many more words in another language. Further work will involve finding a reasonable measure of estimating equivalent content between languages rather than using word counts.

Voting on suitability of stories by participants can be contentious. For some African languages it is easy to identify a panel of experts. For others, consensus on the written form of the language, and on different dialects is still developing. Community models for participation are required to further enhance this. In many countries where one language dominates, such as Rwanda, language standardization has been easier to accomplish. Many African countries are multilingual, and some languages are used across borders, making intergovernmental cooperation very helpful in language standardization efforts. Cooperation with institutions pursuing African language standardization work, would also be helpful, though many such institutions, when they exist are often poorly funded. For languages with a small number of speakers who form a minority section of the population of a country, community cultural organizations or social media groups are the only available and easily accessible resource.

Voting is done by issuing story authors and translators with an id which they enter on the website to be able to cast their votes. Their votes are then stored in a separate database from the id with the aim of enabling anonymity when there are more than two voters for stories in a particular language. It would be good to improve the system so that it is encrypted and the administrator cannot determine what each participant voted for. This is technically challenging since one can also possibly use server logs and database timestamps to correlate activity. Nevertheless, electronic voting has been done for other purposes which have greater security and identity verification requirements, so it should be possible to incorporate this in the current software.

The software has been partially translated into Kiswahili. At present this is done by creating a list of strings for each user facing component on the website. Many open source software projects use files formatted for Gettext (Free Software Foundation 2021*a*) to store translations. It would be good to use such a system.

Further work on making the soft keyboards easier

to use is also needed, in particular for Ethio-Semitic languages such as Amharic. Soft keyboards were added to the software to enable use of orthography that may not be available on standard English and French computer keyboards available on computers in many African countries. Not all languages have standardized orthography, so adaptation and collaboration with linguists is required to do this well. A number of contributors indicated using keyboards on their mobile phones which have all the letters they need for their languages. Enabling creation, submission and viewing of stories through a mobile application in addition to a website may also be something to consider adding.

This note has mainly focused on the technical considerations in creating software to host an online competition for stories in African languages to collect a corpus. An additional article will analyze the collected corpora, and include viewpoints of those who helped obtain, write and translate short stories.

5 Conclusion

It is expected that many of Africa's languages will become extinct, or merged and standardized with other similar languages that are used in the same region, for example as has happened with Runyakitara. Documenting the rich cultural legacy is important for improved understanding and possible lessons that can be learned from the current linguistic diversity. Documenting the cultures is also helpful to enable cross-cultural understanding. The digital age requires digital tools, and while most of the technological developments enabling digitization have occurred in English speaking countries, these have been adapted for local conditions in countries with different cultural contexts. Notable examples of digital natural language processing tools have originated in Russia, China and Japan, some of these tools have been partly developed by commercial social media companies with a strong local presence. By providing open source software to enable short story competitions for African languages, it is hoped to stimulate further collection of short stories at relatively low cost, improve reading and writing skills in African languages, and develop natural language processing tools for African languages.

Acknowledgements

This work has benefited from a $4,000 grant provided by the International Research Development Centre administered by Knowledge for All (`https://www.k4all.org`) through the 2020 IndabaX-AI4D Innovation Grant program. The author thanks William Agbo, Liané Van Den Bergh, Audrey Mbogho, Cascious Mofokeng, Itaru Ohta, Kevin Scannell, Juan Steyn, Benito Trollip, Lilian Wanzare, and Constantine Yuka for many helpful discussions and assistance in running the competitions. The author also wishes to acknowledge all the people who participated in the competitions, by either submitting a story, submitting a translation, soliciting contributions from others, translating a flyer or checking submitted original and translated stories.

References

cPanel LLC (2021), 'cpanel website'.
URL: *https://www.cpanel.net*

ECMA International (2021), 'Javascript specification'.
URL: *https://www.ecma-international.org/publications-and-standards/standards/ecma-262/*

FluxBB Community (2021), 'Fluxbb website'.
URL: *https://fluxbb.org*

Free Software Foundation (2021*a*), 'Gettext website'.
URL: *https://www.gnu.org/software/gettext/*

Free Software Foundation (2021*b*), 'Gnu general public license'.
URL: *https://www.gnu.org/licenses/gpl-3.0.en.html*

GnuSocial Community (2021), 'Gnusocial website'.
URL: *https://www.gnusocial.rocks/*

PHP Community (2021), 'Php website'.
URL: *https://www.php.net/*

Python Software Foundation (2021), 'Python website'.
URL: *https://www.python.org/*

Ruby Community (2021), 'Ruby website'.
URL: *https://www.ruby-lang.org/*

Saide (2021), 'African story book website'.
URL: *https://www.africanstorybook.org/*

Scannell, K. (2021), 'Indigenous tweets'.
URL: *http://indigenoustweets.com/*

Story Weaver (2021), 'Story weaver website'.
URL: *https://storyweaver.org.in/weave_a_story_campaign*