

Work Package 2

Report on Webinar “AI4D Language Challenge”



|  |  |
| --- | --- |
| **Author(s):** | Cristina España i Bonet (DFKI)  Andrea Lösch (DFKI)  Eileen Schnur (DFKI) |
| **Dissemination Level:** | Confidential |
| **Version No.:**  **Date:** | <V1.1>  2020-07-09 |
|  |  |

 

**© 2020 DFKI**

**Table of Contents**

[1 Executive Summary 3](#_Toc45180329)

[2 Agenda Webinar 4](#_Toc45180330)

[3 Research in Natural Language Processing (NLP) 5](#_Toc45180331)

[3.1 State-of-the-art NLP research 5](#_Toc45180332)

[3.2 NLP research in low-resourced languages 6](#_Toc45180333)

[4 Data for African languages 8](#_Toc45180334)

[4.1 Overview of existing data sets for African languages 8](#_Toc45180335)

[4.1 Data collection approaches 9](#_Toc45180336)

[4.1.1 Identification of and collaboration with language data holders 9](#_Toc45180337)

[4.1.2 Identification and use of sources of language data 10](#_Toc45180338)

[4.1.3 Making language data re-usable 10](#_Toc45180339)

[4.2 Important references 11](#_Toc45180340)

[4.3 Status quo: Survey results 12](#_Toc45180341)

[5 Annex 15](#_Toc45180342)

[5.1 Participants List 15](#_Toc45180343)

[5.2 Presentations 17](#_Toc45180344)

[5.3 Summary of African Language Data Sets 18](#_Toc45180345)

[5.4 Registration List 34](#_Toc45180346)

[6 References 39](#_Toc45180347)

**List of Figures**

[Figure 1: Semantic relations of embeddings as described by Mikolov et al. (2003) 5](#_Toc45179536)

[Figure 2: Taxonomy for transfer learning in natural language processing as defined by Ruder (2019) 7](#_Toc45179537)

[Figure 3: Distribution of responses to Question 1 13](#_Toc45179538)

[Figure 4: Distribution of responses to Question 2 13](#_Toc45179539)

[Figure 5: Distribution of responses to Question 3 14](#_Toc45179540)

[Figure 6: Distribution of responses to Question 4 14](#_Toc45179541)

[Figure 7: Participants List 17](#_Toc45179542)

[Figure 8: Non-extensive summary of African language data sets 33](#_Toc45179543)

[Figure 9: List of Registrations 38](#_Toc45179544)

# Executive Summary

The webinar AI4D Network Knowledge Webinar “Making NLP work in Africa” took place on 3 July from 14:00 to 16:00 pm CAT/CEST/UTC+2. It was the first webinar of the AI4D Africa Webinar Series on Natural Language Processing in low-resourced languages and collecting language data in African languages.

The webinar featured speakers from Masakhane and Saarland University, Ghana NLP, University of Cape Town, Instadeep and the German Research Centre for Artificial Intelligence (DFKI). All information about the webinar was made available through the event website: <https://ai4d.ai/event/ai4dnetwork-webinar-nlp/>.

Key questions addressed included the status quo of NLP research with regard to low-resourced languages, how to ensure good quality translations for African languages, and how to approach data collection in and for Africa.

Also, a short quiz testing the participants’ knowledge was included in the registration and promotion of the webinar: https://forms.gle/2hGivcA1WbbrtEZu9

Overall, 99 participants followed the webinar. 53 of them answered the quiz.

# Agenda Webinar

|  |  |
| --- | --- |
| **Time** | **Description** |
| **14:00 – 14:10** | **Welcome Address** (*Kathleen Siminyu, Regional Coordinator AI4D)* |
| **14:10 – 14:30** | **NLP research in low-resources languages** *(Cristina España i Bonet, DFKI)* |
| **14:30 – 14:45** | **Yorùbá and beyond: NLP from an African perspective** *(Jesujoba Alabi, DFKI)* |
| **14:45 – 15:00** | **Ensuring good text quality in African language datasets** *(David Adelani, Saarland University & Masakhaneop)* |
| **15:00 – 15:15** | **Break** |
| **15:15 – 15:25** | **Existing language datasets in African languages** *(Andrea Lösch, DFKI)* |
| **15:25 – 15:55** | **Discussion round: Data collection approaches in Europe and Africa** *(Moderator: Andrea Lösch, DFKI, Experts: Orevaoghene Ahia - Instadeep, Stephen E. Moore - Ghana NLP, Tobias Schonwetter - University of Cape Town)* |
| **15:55 – 16:00** | **Closing Statement** *(Andrea Lösch, DFKI)* |

# Research in Natural Language Processing (NLP)

## State-of-the-art NLP research

Language technologies in general have experimented a boost in quality in correlation with the boost in computing power and the consequent use of deep learning models. The current hardware has the capacity to train models that were unimaginable just a few years ago. More neural network layers, more neurons, and ultimately more parameters are used every day to model languages.

In the deep learning era, linguistic units are represented by embeddings, semantic representations at word, sentence or even document level. An embedding is a dense representation in a low-dimensional vector space. That is, a vector with 50, 100 or 700 components for instance that represents mathematically the semantics of a linguistic unit. This mapping from linguistic units into mathematical entities allows to perform mathematical operations such as sum or subtraction of words (or sentences) to get new meanings and relations: king−man+woman≈queen.

Ein Bild, das Uhr enthält.

Automatisch generierte Beschreibung

Figure 1: Semantic relations of embeddings as described by Mikolov et al. (2003)

However, static word embeddings [1,2,3], that is, word2vec-like embeddings that allow to perform these simple operations, are powerful but still show some limitations. Since neural networks learn one vector per word, they are not able to capture polysemy and more fine grained details such as long-term dependencies, agreements, anaphora or negation are lost in the representation. To overcome this limitation, language models or sentence representations [4,5], that is, BERT-like embeddings, provide contextual embeddings which assign each word a representation based on its context. The representation of *bank* in a sentence that talks about economics will be different to the representation of *bank* in a sentence that talks about fishing. These embeddings are even more powerful, and achieve state-of-the-art results for several natural language processing tasks, but need large amounts of data to be estimated. The usage is also different. Word vectors allow for a dictionary look-up of words and their corresponding vectors, they are static entities and are good to initialise input word embeddings in several NLP tasks. Contextual word vectors are vectors obtained on-the-fly by passing text through a deep learning model. These representations are good for transfer learning into several NLP tasks.

Embeddings, both static and contextual, are key components for machine translation, question answering, chatbots, search engines, named entity recognition, or text classification to give some examples. All these tasks facilitate communication but only machine translation is intrinsically multilingual, the others need multilingual components to facilitate communication in and across different languages. In these cases, multilinguality is mostly achieved with machine translation or bilingual embeddings, a hot topic in current NLP research. Different deep learning architectures such as multilingual BERT [5], LASER [6] and XLM [7] have proved successful in the multilingual setting.

## NLP research in low-resourced languages

All these architectures learn the semantic representations from unannotated text, making them *cheap* when data exists. Wikipedia and the web in general is a huge source of unannotated text. But what happens for languages with a small presence in the online world and/or with few digital resources? Such data-hungry models might fail in modeling. In a low-resourced setting, the data quality, processing and model selection is more critical than in a high-resourced scenario. The characteristics of a language (such as word order, grammatical structure or diacritization) should be taken into account in order to choose the relevant data and model to use. All these features are usually ignored for English and deep learning models.

Also, for low-resourced languages, the evaluation is more difficult and therefore normally ignored simply because of the lack of resources. In the high-resourced setting, one has a smorgasbord of tasks and test sets to evaluate on. This is the best-case scenario, languages with tons of data for training that generate high-quality models. In the low-resourced setting, training data is scarce, and the assumption that the capability of deep learning architectures to learn (multilingual) representations in the high-resourced setting holds in the low-resourced one does not need to be true. In fact, massively generated embeddings perform poorly for low-resourced languages as compared to the performance for high-resourced ones and this is due both to the quantity but also the quality of the data used [8].

Several techniques are developed to be able to *transfer* information from the high-resource scenario to the low-resource one and mitigate the problems. In machine learning, transfer learning techniques are a collection of methods that allow to apply models obtained solving one problem to a different but related one. The original problem might have lots of data as compared to the new one, but this is not a necessary condition and two low-resourced tasks can benefit one from another through transfer. In natural language processing, four main transfer methodologies are used:

* Domain adaptation
* Cross-lingual learning
* Multi-task learning
* Sequential transfer learning

These allow to convert initial generic models so that they work in another domain or language (transduction). On the other hand, they allow to learn several tasks at the same time or adapt results from one task to another one (induction).

Ein Bild, das Screenshot enthält.

Automatisch generierte Beschreibung

Figure 2: Taxonomy for transfer learning in natural language processing as defined by Ruder (2019)

All these methods are used for languages rich in resources such as English and Chinese, but are essential for low-resourced languages. The best methodology and approach to use might depend on the specific language and task. For instance, Lauscher et al. [10] find that the transfer for multilingual transformer models is less effective for resource-lean settings and distant languages. Tran [11] suggested in these cases a fast adaptation method for obtaining a bilingual BERT model for English and a target language, where the target language could be any low-resourced language, and where only monolingual data is needed.

# Data for African languages

## Overview of existing data sets for African languages

An initial, non-exhaustive list of data sets for African datasets can be found below. A detailed overview on their language coverage, machine-readiness, URL, etc. is provided in the Annex 5.3.

* AfDB Statistical Data Portal
* African Speech Technology Corpus
* AfricArxiv
* Alákòwé
* ALLEX Corpora
* Autshumato Corpus
* BBC Yorùbá
* Benin Open Data Portal
* Bible
* bible-uedin (multilingual parallel corpus created from translations of the bible)
* CORAAL - Corpus of Regional African American Language
* CorpAfroAs
* Corpus Bambara de Référence
* Corpus Maninka
* Corpus of South African English (CoSAE)
* de Schryver and Prinsloo: The compilation of electronic corpora, with special reference to the African Languages (2000)
* Die Pharos-korpus van hedendaagse Afrikaans (PAK)
* DOBES - Documentation of endangered languages
* Doctrine $ Conventant
* Èdè Yorùbá Rewà
* ELAR - Endangered Language Archive
* English - Luganda Parallel Corpus
* English - Luo Machine Translation System
* Global Voices
* GlobalPhone Pronunciation Dictionaries
* GNOME
* haWaC: Hausa corpus from the Web
* Helsinki Corpus of Swahili
* igTenTen: Igbo corpus from the web
* Jehova Witness
* JW300 corpora
* KDE4 v2
* Kenia Open Data Portal
* Kiswahili Internet Corpus
* Lacito
* Lagos-NWU corpus
* Leibniz Corpora Collection
* Lwazi corpus for automatic speech recognition (ASR)
* memat
* NCHLT isiZulu Text Corpus
* Northern Sotho Part-of-Speech Tagger (V2) - Demo
* Onyenwe, Uchechukwu, Hepple: Part-of-speech Tagset and Corpus Development for Igbo, an African Language (2014)
* Open-access portal for data protection laws and information in Africa
* openAfrica Portal
* Òrò Yorùbá
* Pretoria Corpora
* Recalls Cilubà Corpus
* Rosetta Disk 1.0
* Sadilar Resource Catalogue and Index
* SAWA Corpus
* Scientific e-lexicography for Africa(2012-2015)
* SPC - Stockholm Parallel Corpora
* Swahili Computer corpora
* The Crúbadán Project: Corpus building for under-resourced languages
* The GlobalPhone Swahili corpus
* Treebanks in Universal Dependencies
* Ubuntu
* ukuxhumana
* Unsupervised compound splitter for Afrikaans
* Voice of Nigeria Yorùbá news
* Wikipedia
* XhosaNavy
* Yorúbà Bible
* Yorùbá Tweets
* Yorùbá Wikipedia
* Zulu Wikimedia

## Data collection approaches

In order to be successful, the collection of training data (especially for under-resourced) languages should (i) encourage local ownership and responsibility and (ii) build on established efforts and expertise, thus creating synergies on every level of the effort.

### Identification of and collaboration with language data holders

Language data can be retrieved from the corresponding sources and holders of language data. As such, the first step in the collection of any language data is the identification of relevant language data holders and suitable language data sources.

Language data holders include any organisations and/or people that may create language data, in particular

* African translators and/or translation agencies (e.g. The South African Translators’ Institute, [SATI](https://www.translators.org.za), a collection of South African translators is also available [here](http://www.exporthelp.co.za/network/translators_summary.html)),
* translation services in African national ministries, public services or governmental agencies (e.g. [Language Unit of the Department of Cultural Affairs and Sport (DCAS), Western Cape Government](https://www.westerncape.gov.za/your_gov/102), South African Centre for Digital Language Resources ([SADiLaR](https://www.sadilar.org/index.php/en/about))),
* African and/or international open data portals (e.g. [openAfrica](https://africaopendata.org)),
* African language and/or language technology researchers and members of academia (e.g. [AfricArxiv](https://info.africarxiv.org), African Academy of Languages ([ACALAN](https://www.acalan-au.org/aboutus.php))),
* African and/or international language technology and language service providers (e.g. [Translate4Africa](https://www.translate4africa.com/about-us/), [Folio Online](https://www.folio-online.co.za)).

Retrieving language data directly from the relevant language data holders can be done in various ways, including both

* face-to-face (e.g. through data collection workshops, focus group meetings, on-site assistance at the data holders’ site) and
* remote (e.g. through surveys among data holders or direct phone interviews).

Surveys or phone interviews are always helpful for the identification of new data sets or for the identification of problems of the sharing of language data.

Wherever possible, however, it is advisable to opt for face-to-face interaction with the different stakeholders: This ensures visibility, maximizes impact and creates synergies where possible. Moreover, the involvement of policymakers at national, regional and local levels in the activities is advisable to ensure highest level support.

While workshops and focus group meetings should encourage the sharing of language data and give data holders the opportunity to ask questions (e.g. on the practical technical and legal aspects of the sharing of language data), on-site assistance should be provided for the particular solving of a technical and legal issue with data (e.g. evaluation of particular data sets to estimate feasibility of further processing or sharing).

### Identification and use of sources of language data

Sources of language data can be any bi- or multilingual websites in the languages sought, ranging from governmental websites over websites of public services and academic institutions in the target countries to websites of international, national or local organisations in the target countries.

In order to identify and retrieve mono-, bi- or multilingual language data from the Internet and to turn them into MT-ready language resources, web crawling (e.g. using HTTrack - <https://www.httrack.com/>) can be very useful. Examples of language independent crawlers that are frequently used by our researchers include ParaCrawl (<https://github.com/bitextor/bitextor>) or the ILSP-FC (<http://nlp.ilsp.gr/redmine/projects/ilsp-fc>). The ILSP-FC does not only allow to identify and create language resources for particular languages, but also to undertake focussed crawls for particular domains. With regard to languages used in Africa, it currently supports Afrikaans, French, English, Dutch, and Arabic. In order to make the crawler work for a new language, a corresponding language needs to be created. Typically, a corpus of around 1 Mio. words is sufficient.

### Making language data re-usable

Once relevant data sets have been identified and retrieved, it is important to ensure that they are actually (re-) usable. The usability of a data set has two dimensions: Its technical usability and its legal usability. As you can see from Figure 8: Non-extensive summary of African language data sets, the majority of the language data that is currently available requires further technical processing to be MT ready and/or the clarification of the legal status of the particular data set (i.e. the identification whether and if so how data can be used and/or shared). Only 39 out of 86 data sets are MT ready, and less than 10 data sets have the legal status identified that would allow for re-use and sharing.

To ensure technical usability of the data for training MT systems, at least the following aspects need to be considered:

* Is the format readable? If files are provided under proprietary formats (like Trados format), the submitter of the resource should be contacted to obtain a converted version into a non-proprietary format. If data comes as a series of PDF or Word DOC(X) documents, they need to be automatically processed so that translation memories can be extracted.
* Is the source / are the sources copyrighted? Copyrighted contents need to be excluded or a corresponding usage license needs to be acquired (see below, legal usability).
* Were source and target language(s) identified correctly?
* Is the alignment ok? An automated validation and/or filtering of the data set should be conducted to check e.g. the alignment score, the length ratio, or translation unit variants. If scores indicate a bad quality of the data set or individual translation units (TUs), they should be excluded.
* Are there any tokenization errors (no separator between words)? Data sets (or the parts of the data set) that contain tokenization errors should be excluded.
* Is the content machine-translated? Machine-translated content may not be considered as high-quality language resource.

As regards the legal usability of a particular data set, the following questions need to be assessed:

* Does the data contain any personal or confidential information? If yes, personal and confidential data must be excluded.
* Is the data protected by copyright? National laws may contain rules excluding certain works from copyright protection.
* If the data is protected by copyright, can I identify the owner of the copyright or the author of the work? If yes, one should obtain corresponding usage rights / a corresponding license from the IPR holder.
* Is the data available under a public license? For example, certain datasets are made available by the owner of copyright under a license that allows reuse or redistribution free of charge (e.g. creative commons licenses).
* If no public license is clearly marked on the document, one should check the terms of use or if any documentation may help you determine the conditions of reuse of the material.

## Important references

The International Open Data Charter (<https://opendatacharter.net/principles/>) represents a comprehensive, world-wide guidance document on the technical and legal characteristics necessary for digital data to be freely used, reused and redistributed.

The Africa Data Revolution Report (<https://webfoundation.org/docs/2019/03/Africa-data-revolution-report.pdf>) on the other hand provides useful details on the status and emerging impact of open data in Africa, in particular open government data. While it recognizes that there is a huge diversity between African governments in embracing open data, it concludes that open data in Africa needs a vibrant, dynamic, open and multi-tier data ecosystem if the data are to make a real impact.

## Status quo: Survey results

As part of the invitation to the webinar, a short quiz with five questions was included in order to test the participants’ knowledge of NLP. A total of 53 participants responded to the quiz which comprised the following questions:

Question 1: Low-resourced languages are languages

* with few speakers
* with few data in electronic format
* with few experts

Question 2: For training a neural network for a natural language processing (NLP) task, you need

* Lots of training data
* Few data but very high quality
* The amount of data depends on the task

Question 3: Resources from English can be used for Ewe

* Sometimes, relying on common aspects
* Not directly, but can be adapted if data in Ewe exists
* Never, languages are completely different

Question 4: Transfer learning can be used in NLP to

* Adapt a resource from one domain to another one (e.g. politics to economics)
* Adapt a model from one language to another one (e.g. from English to Ewe)
* Both of the above
* None of the above

**Question 5: What's the main difference between Word2Vec and BERT?**

The distribution of responses is illustrated below. As can be seen, for the more general Question 1, 51 out of 53 respondents gave the right answer. For all further questions, the results were more ambiguous: For Question 2, only 28 participants selected the right answer. For Question 3, 39 respondents chose the correct response and for question 4, 30 participants made the right choice.

**Ein Bild, das Screenshot enthält.

Automatisch generierte Beschreibung**

Figure 3: Distribution of responses to Question 1

**Ein Bild, das Screenshot enthält.

Automatisch generierte Beschreibung**

Figure 4: Distribution of responses to Question 2

**Ein Bild, das Screenshot enthält.

Automatisch generierte Beschreibung**

Figure 5: Distribution of responses to Question 3

**Ein Bild, das Screenshot enthält.

Automatisch generierte Beschreibung**

Figure 6: Distribution of responses to Question 4

Regarding the answers to the open-ended Question 5 on the difference between BERT and word2vec, only 8 responses were received in total, out of which two were “I don’t know”. The remaining 6 responses are provided below:

* Word2Vec is a static word embedding while BERT is a pre-trained model that learns deeper contextualised
* word representations due to its bidirectional nature.
* Model vs. Embedding
* BERT considers the context of the words but Word2Vec does not.
* Architecture, Training task, context size
* Word2Vec is a context free word embedding while BERT takes into account the context of each word
* Word2Vec transforms words to vectors. BERT is based on the tenaformer model. I don't fully understand
* transformers though, just that they are currently SOTA.

# Annex

## Participants List

(Note: Participants marked with \* have either not registered prior to the workshop or their username could not be clearly assigned to an existing registration)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Last Name | First Name | Affiliation |
|  | Addo\* | Salomey | N/A |
|  | Adelani | David | Saarland University & Masakhane |
|  | Agyapon-Ntra | Kwadwo | Self-attending |
|  | Ahia | Orevaoghene | Instadeep |
|  | Alabi | Jesujoba | DFKI |
|  | Amanfu | Richard | Institute of ICT Professionals Ghana |
|  | Anebi | Emmanuel | DataInsight |
|  | Anzaldo | Isa | UdS |
|  | Asamoah | Eugene | Electricity Company Of Ghana |
|  | Ayami | Yasin | TsogoloTech |
|  | Bamutura | David | Mbarara University of Science and Technology |
|  | Berejena | Beatrice | NA |
|  | Betty | Betty | University of Johannesburg |
|  | Buabeng | Edwin | Huawei |
|  | Budu | Joel | Artificial Intelligence Association of Ghana |
|  | Byamugisha | Joan | IBM Research Africa |
|  | Cayralat | Christian | LCT |
|  | Chirwa | Temweka | University of Cape Town |
|  | Coffie Debrah | Emmanuel | University of Cape Coast |
|  | David | Davis | TYD Innovation Incubator |
|  | Dietrich | Klakow | Saarland University |
|  | Donner | Jonathan | Caribou digital |
|  | Dosseh | Desire | Dakar Institute of Technology |
|  | du Plessis | Liëtte | University of Johannesburg |
|  | Dube | Hloniphani | Bluemachines (Pty) Ltd |
|  | Ekem | Ivy | UCC SMS |
|  | Elmers | Mikey | University of Saarland |
|  | España-Bonet | Cristina | DFKI |
|  | Estarrona | Ainara | HiTZ zentroa (UPV/EHU) |
|  | Gaelejwe | Theodore | IBM Research |
|  | Gardent | Claire | CNRS |
|  | Gebremeskel | Gebrekirstos | CWI |
|  | Gesicho | Rose Delilah | Student |
|  | Graaf | Michael | Wikimedia-ZA |
|  | Griciūtė | Bernadeta | Saarland University |
|  | Gyekye | Kwame | Experian |
|  | Ibrahim | Omnia | University of Zürich |
|  | Ihle | Frank | private |
|  | Jarso | Guyo | University of Rwanda |
|  | Kashupi | Tokolo N. | Namibia University of Science and Technology |
|  | Kiden | Sarah | Northumbria University |
|  | Kioko | Moses |  |
|  | Konobelkina | Anna | University of Saarland |
|  | Ladipo | Taiwo | SAP |
|  | Lang | Inga | EM Double Master in Language and Communication Technologies (LCT), University of Groningen and University of Malta |
|  | Lösch | Andrea | DFKI |
|  | Luberenga | John | Flying Gravity Technologies |
|  | Lubrini | Elisa | Université de Lorraine |
|  | Mahamah\* | Rhoda | N/A |
|  | Mahlezana | Thiba | TGM MakerSpace |
|  | Makumbirofa | Hamony | Tshimologong |
|  | Malatji | Masike | University of Johannesburg |
|  | Mansour | Ayman | Sudan university of science and technology |
|  | Mbadi | Sheila | CMU-Africa |
|  | Mintz | Blanca | GIZ |
|  | Moore | Stephen | Ghana NLP |
|  | Moyo | Mlamuleli | Tmg makers space |
|  | Mthembu | Nkululeko | Private |
|  | Muhire | Remy | Mozilla |
|  | Muite | Benson | N/A |
|  | Mukiibi | Jonathan | AI Lab Makerere University |
|  | Musoya | Gael | Digitech group |
|  | N/A\* | Ari | Praekelt |
|  | N/A\* | Ajamitoure | N/A |
|  | N/A\* | Celina | N/A |
|  | N/A\* | Derick | N/A |
|  | N/A\* | Guest | N/A |
|  | N/A\* | Peter | N/A |
|  | Nakatumba | Joyce | Makerere University |
|  | Ndodana | Nwabisa | Student |
|  | Nxumalo | Thabiso | North Park Telecoms |
|  | Oduor | Clinton | clintonoduor3@outlook.com |
|  | Ogayo | Perez | African Leadership University |
|  | Oghenekevwe\* | Ajewole | N/A |
|  | Ojesina | Akolade | University of Ibadan |
|  | Oloke | Adewale | Unilag |
|  | Orlic | Davor | Knowledge 4 All Foundation |
|  | Owusu-Darko | Ama | Ensign College of Public Health |
|  | Oyewole\* | Isaiah | N/A |
|  | Phahlamohlaka | Lazarus | North Park Telecoms |
|  | Pienaar | Marne | UJ |
|  | Pratt Miles | Jennifer | Lacuna Fund: Our Voice on Data |
|  | Prenom | Daouda | Université Alioune Diop de Bambey |
|  | Quayson | Ebenezer | UENR |
|  | Resch | Christian | GIZ |
|  | Sam | Abraham | FAIR Forward, GIZ |
|  | Schnur | Eileen | DFKI |
|  | Schonwetter | Tobias | University of Cape Town |
|  | Sibal | Prateek | UNESCO |
|  | Siddharth\* | Nandan | GIZ |
|  | Siminyu | Kathleen | Artificial Intelligence for Development - Africa |
|  | Smith | Blake | University of Edinburgh |
|  | Smith | Matthew | International Development Research Centre |
|  | van Genabith | Josef | DFKI |
|  | Vorsah Amponsah\* | Irene Kafui | N/A |
|  | Wanzare | Lilian | Maseno University |
|  | Waterfield | Rebecca | n/a |
|  | Watson | Sarah | Mozilla |
|  | Wilson | Daniel | XRI |

Figure 7: Participants List

## Presentations

All presentations held during the webinar “Making NLP work in Africa – with an introduction to the GIZ AI4D African Language Dataset Challenge” are available online at <https://cloud.dfki.de/owncloud/index.php/s/tZHbPK4F4F4QEWf>

## Summary of African Language Data Sets

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Resource Name** | **URL** | **Language(s)** | **MT ready?** | **Domain** | **License** | **Right holder** | **contact person name** | **contact person email** | **Comments** |
| AfDB Statistical Data Portal | <https://dataportal.opendataforafrica.org/data#menu=topic> |  | unclear | Open Data Portal |  |  |  |  | Seems to be more in English; according to description "the largest public and open data repository in the world", but the website is only available in French and English and mainly includes statistics (i.e. figures and diagrams). |
| African Speech Technology Corpus | <https://rma.nwu.ac.za/index.php/resource-catalogue/ast-corpus-isizulu.html> | Zulu | unclear |  |  |  |  |  | Potential risk by access |
| AfricArxiv | <https://info.africarxiv.org/> |  | unclear | digital archive for African research communication |  |  |  |  | Seems to be more in English |
| Alákòwé | [alakoweyoruba.wordpress.com](http://alakoweyoruba.wordpress.com/) | Yorùbá | no |  |  |  |  |  |  |
| ALLEX - Ndebele Corpus | <http://www.edd.uio.no/allex/corpus/africanlang.html> | Ndebele | unclear |  |  |  | Daniel Ridings (Unit for Digital Documentation, Oslo University) | daniel.ridings@edd.uio.no | Tools. Need to ask the authors for the corpora |
| ALLEX- Shona Corpus | <http://www.edd.uio.no/allex/corpus/africanlang.html> | ChiShona | unclear |  |  |  | Daniel Ridings (Unit for Digital Documentation, Oslo University) | daniel.ridings@edd.uio.no | Tools. Need to ask the authors for the corpora |
| Autshumato Corpus | <https://rma.nwu.ac.za/index.php/autshumato-eng-zu-parallel-corpora.html> | English - Zulu, Setswana, Xitsonga, Northern-Sotho, Afrikaans | no |  |  |  |  |  | Potential risk by access |
| BBC Yorùbá | [bbc.com/yoruba](http://bbc.com/yoruba) | Yorùbá | yes |  |  |  |  |  | However: text needs to be made available |
| Benin Open Data Portal | [https://benin.opendataforafrica.org](https://benin.opendataforafrica.org/) |  | unclear | Open Data Portal |  |  |  |  | Seems to be more in English; economic, demographic, and social data available for download in Excel format as well as PDF. |
| Bible | [www.bible.com](http://www.bible.com/) | Twi | yes |  |  |  |  |  | However: text needs to be made available |
| bible-uedin | <http://opus.nlpl.eu/bible-uedin.php> |  | yes |  |  |  | Christos Chrstidoulopoulos, Mark Steedman |  | Multilingual parallel corpus created from translations of the bible; However: need to extract the African languages; https://link.springer.com/article/10.1007/s10579-014-9287-y |
| CORAAL - Corpus of Regional African American Language | <https://oraal.uoregon.edu/coraal> |  | no |  |  |  |  |  | Specific to American African Language (mostly speech) |
| CorpAfroAs | <https://corpafroas.huma-num.fr/Archives/corpus.php> |  | unclear | Spoken Afroasiatic languages | "You are welcome to use the CorpAfroAs Format and Tools for your data., Please quote the CorpAfroAs project when you use our annotation scheme and/or software and/or procedures.Thank you." |  | Azeb Amha, Christian Chanard | a.amha@hum.leidenuniv.nl, christian.chanard@cnrs.fr | Need to contact the research center, manual is available here: <https://corpafroas.huma-num.fr/fichiers/manual.pdf> |
| Corpus Bambara de Référence | [http://cormand.huma-num.fr](http://cormand.huma-num.fr/) | Bambara | yes |  |  |  | Valentin Vydrine | vydrine@gmail.com | Makes a good impression, but unclear how to download, link to the corpus: <http://cormande.huma-num.fr/corbama/run.cgi/first_form> |
| Corpus Maninka | <http://cormand.huma-num.fr/cormani/> | Maninka | unclear |  |  |  | Valentin Vydrin | vydrine@gmail.com | Corpus is available here: <http://cormande.huma-num.fr/cormani/run.cgi/first_form> |
| Corpus of South African English (CoSAE) | Source: <https://www.researchgate.net/figure/Corpora-of-African-languages-excluding-the-South-African-ones_tbl2_261835113> | English | unclear |  |  |  |  |  |  |
| de Schryver and Prinsloo: The compilation of electronic corpora, with special reference to the African Languages (2000) | <https://tshwanedje.com/publications/Corpora.pdf> |  | no |  |  |  |  |  | Publication only. If access to the data can be organised, this would be a good source. |
| Die Pharos-korpus van hedendaagse Afrikaans (PAK) | Source: <https://www.researchgate.net/figure/Corpora-of-African-languages-excluding-the-South-African-ones_tbl2_261835113> | Afrikaans | unclear |  |  |  |  |  |  |
| DOBES - Documentation of endangered languages | [https://dobes.mpi.nl/#](https://dobes.mpi.nl/) | Browsable collections for Africa | unclear | Archive |  |  | Project leader: Peter Wittenburg (+31-24-3521175), Archive Manager: Paul Trilsbeek (+31-24-3521203), Software: Han Sloetjes (+31-24-3521467) | dobes@mpi.nl. | The DOBES Archive contains language documentation data from a great variety of languages from around the world that are in danger of becoming extinct. The portal gives access to the material in the archive. |
| Doctrine $ Conventant | [github.com/Niger-Volta-LTI](http://github.com/Niger-Volta-LTI) | Yorùbá | yes |  |  |  |  |  |  |
| Èdè Yorùbá Rewà | [deskgram.cc/edeyorubarewaa](http://deskgram.cc/edeyorubarewaa) | Yorùbá | no |  |  |  |  |  |  |
| ELAR - Endangered Language Archive | <https://www.soas.ac.uk/elar/> | Browsable collections for Africa | unclear | Multimedia collections of endangered languages (every day language, verbal art, narratives, etc.) |  |  |  | elararchive@soas.ac.uk | The Endangered Language Archive (ELAR) is a digital repository preserving and publishing endangered language documentation materials from around the world. The materials are digital and freely available (after free registration). |
| English - Luganda Parallel Corpus | <https://www.aflat.org/node/86> | English - Luganda | no |  |  |  |  |  | Found [here](https://corplinguistics.wordpress.com/2012/02/08/african-language-corpora/) |
| English - Luo Machine Translation System | <https://www.aflat.org/luomt> | English - Luo (Dholuo) | no |  |  |  |  |  | Found [here](https://corplinguistics.wordpress.com/2012/02/08/african-language-corpora/), a system (offline now), but not a resource |
| Global Voices | [yo.globalvoices.org](http://yo.globalvoices.org/) | Yorùbá | yes |  |  |  |  |  | However: not directly re-usable. We know that some data are made available |
| GlobalPhone Hausa Pronunciation Dictionary | <http://catalog.elra.info/en-us/repository/browse/ELRA-S0353/> | Hausa | yes |  | Commercial use |  | V. Mapelli | mapelli@elda.org | Not free of charge |
| GlobalPhone Swahili Pronunciation Dictionary | <http://catalog.elra.info/en-us/repository/browse/ELRA-S0376/> | Swahili | yes |  | Commercial use |  | V. Mapelli | mapelli@elda.org | Not free of charge |
| GNOME | <http://opus.nlpl.eu/GNOME-v1.php> | English - Igbo, Afrikaans, Hausa | yes |  |  |  |  |  |  |
| haWaC: Hausa corpus from the Web | <https://www.sketchengine.eu/hawac-hausa-corpus/> | Hausa | yes |  |  |  |  |  |  |
| Helsinki Corpus of Swahili | <http://catalog.elra.info/en-us/repository/browse/ELRA-W0119/> | Swahili | yes |  | Commercial use |  | V. Mapelli | mapelli@elda.org | Not free of charge |
| Helsinki Corpus of Swahili 2.0 (HCS 2.0) | <http://metashare.csc.fi/repository/browse/helsinki-corpus-of-swahili-20-hcs-20-annotated-version/232c1910b9eb11e5915e005056be118e59fb2e920f1f4c0cafc94915fc6f5cac/> | Swahili | yes |  |  |  |  |  | Also available here: http://catalog.elda.org/en-us/repository/browse/ELRA-W0119/ |
| igTenTen: Igbo corpus from the web | <https://www.sketchengine.eu/igtenten-igbo-corpus/> | Igbo | yes |  |  |  |  |  |  |
| Jehova Witness | [www.jw.org/yo](http://www.jw.org/yo) | Yorùbá | yes |  | CC-BY-NC-SA. |  |  |  | However: text needs to be made available |
| Jehova Witness | [www.jw.org/tw](http://www.jw.org/tw) | Twi | yes |  |  |  |  |  |  |
| JW300 corpus | opus.nlpl.eu/JW300.php | Twi, Yorùbá | yes |  | CC-BY-NC-SA. |  |  |  |  |
| KDE4 v2 | <http://opus.nlpl.eu/KDE4-v2.php> | English - Afrikaans, Hausa | yes |  |  |  |  |  |  |
| Kenia Open Data Portal | [http://www.opendata.go.ke](http://www.opendata.go.ke/) |  | unclear | Open Data Portal |  |  |  |  | Seems to be more in English; makes public government datasets accessible for free to the public in easy reusable formats. |
| Kiswahili Internet Corpus | Source: <https://www.researchgate.net/figure/Corpora-of-African-languages-excluding-the-South-African-ones_tbl2_261835113> | Kiswahili | unclear |  |  |  |  |  | Tools. Need to ask the authors for the corpora |
| Lacito | <https://pangloss.cnrs.fr/corpus/index.html> | Bafia, Mankon, Tikar, Uldeme, Wuzlam, Vute, Ngazidja, Maore, Kakabe, Kam, Nyesam, Iraqw, Langi, Mbugwe, Nyilamba | unclear | Speech |  |  |  |  | "The Lacito Archive provides free access to documents of continuous, spontaneous speech, mostly in rare or endangered languages recorded in their cultural context and transcribed in consultation with native speakers." (Lüdeling, A., Kytö M: Corpus linguistics: an international handbook, Volume 2, p.466) |
| Lagos-NWU corpus | [github.com/Niger-Volta-LTI](http://github.com/Niger-Volta-LTI) | Yorùbá | yes |  |  |  |  |  |  |
| Leibniz Corpora Collection | <http://corpora.uni-leipzig.de/en?corpusId=ibo_community_2017> | Igbo, Hausa, Swahili | no |  |  |  |  |  | The Igbo corpus is available here: https://curl.corpora.uni-leipzig.de/languages/ibo, Note: was not downloadable |
| Lwazi corpus for automatic speech recognition (ASR) | N/A |  | no |  |  |  | Jaco Badenhorst, Charl van Heerden, Marelie Davel and Etienne Barnard HLT Research Group, Meraka Institute, CSIR, South Africa | jbadenhorst@csir.co.za, mdavel@csir.co.za cvheerden@csir.co.za, ebarnard@csir.co.za |  |
| memat | <http://opus.nlpl.eu/memat.php> | Xhosa - English | yes |  |  | Please cite the following article if you use any part of the corpus in your own work: J. Tiedemann, 2012, Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012) | Jörg Tiedemann |  | Xhosa-English parallel corpora, funded by EPSRC, the Medical Machine Translation project worked on machine translation between ixiXhosa and English, with a focus on the medical domain.2 languages, total number of files: 80; total number of tokens: 4.03M; total number of sentence fragments: 0.33M |
| NCHLT isiZulu Text Corpus | <https://rma.nwu.ac.za/index.php/isizulu-nchlt-text-corpora.html> | Zulu | unclear |  |  |  |  |  | Potential risk by access |
| Northern Sotho Part-of-Speech Tagger (V2) - Demo | <https://www.aflat.org/node/177> | Northern Sotho | no |  |  |  |  |  | found here, Tools rather than data |
| Onyenwe, Uchechukwu, Hepple: Part-of-speech Tagset and Corpus Development for Igbo, an African Language (2014) | <http://www.aclweb.org/anthology/W14-4914> | Igbo | no |  |  |  |  |  | Paper only. |
| Open-access portal for data protection laws and information in Africa | [https://dataprotection.africa](https://dataprotection.africa/) | Afrikaans, Chichewa, Igbo, Sesotho, Shona, Afsoomaali, Basa Sunda, Kiswahili, Xhosa, Yorùbá, Zulu | no | Legal |  |  | Justin Bryant (research and coordination) | dataprotection@altadvisory.africa | Seems to be more in English, Data Protection Africa is an ALT Advisory special project. |
| openAfrica Portal | [https://africaopendata.org](https://africaopendata.org/) |  | unclear | Open Data Portal |  |  |  |  | Seems to be more in English |
| Òrò Yorùbá | [oroyoruba.blogspot.com](http://oroyoruba.blogspot.com/) | Yorùbá | no |  |  |  |  |  |  |
| Pretoria Afrikaans Corpus | <https://www.up.ac.za/african-languages/article/17933/speakoutup> | Afrikaans | unclear |  |  |  | Prof DJ Prinsloo | danie.prinsloo@up.ac.za | Need to contact the center; General electronic corpora for all eleven official South African languages have been compiled at UP. These corpora are solely utilised for student training and academic research, specifically in the fields of lexicography, terminology, linguistics, translation practice and corpus-based translation studies (CTS). For controlled access to these corpora, which involves on-site computer processing of the corpus and downloading only the results of the analyses, Prof DJ Prinsloo (danie.prinsloo@up.ac.za) can be contacted. |
| Pretoria Chose Corpus | <https://www.up.ac.za/african-languages/article/17933/speakoutup> | Xhosa | unclear |  |  |  | Prof DJ Prinsloo | danie.prinsloo@up.ac.za | See above (Pretoria Afrikaans Corpus) |
| Pretoria Ndebele Corpus | <https://www.up.ac.za/african-languages/article/17933/speakoutup> | Ndebele | unclear |  |  |  | Prof DJ Prinsloo | danie.prinsloo@up.ac.za | See above (Pretoria Afrikaans Corpus) |
| Pretoria Sepedi Corpus | <https://www.up.ac.za/african-languages/article/17933/speakoutup> | Sepedi | unclear |  |  |  | Prof DJ Prinsloo | danie.prinsloo@up.ac.za | See above (Pretoria Afrikaans Corpus) |
| Pretoria Sesotho Corpus | <https://www.up.ac.za/african-languages/article/17933/speakoutup> | Sesotho | unclear |  |  |  | Prof DJ Prinsloo | danie.prinsloo@up.ac.za | See above (Pretoria Afrikaans Corpus) |
| Pretoria Setswana Corpus | <https://www.up.ac.za/african-languages/article/17933/speakoutup> | Setwana | unclear |  |  |  | Prof DJ Prinsloo | danie.prinsloo@up.ac.za | See above (Pretoria Afrikaans Corpus) |
| Pretoria Swati Corpus | <https://www.up.ac.za/african-languages/article/17933/speakoutup> | Swati | unclear |  |  |  | Prof DJ Prinsloo | danie.prinsloo@up.ac.za | See above (Pretoria Afrikaans Corpus) |
| Pretoria Tshiivenda Corpus | <https://www.up.ac.za/african-languages/article/17933/speakoutup> | Tshivenda | unclear |  |  |  | Prof DJ Prinsloo | danie.prinsloo@up.ac.za | See above (Pretoria Afrikaans Corpus) |
| Pretoria Xitsonga Corpus | <https://www.up.ac.za/african-languages/article/17933/speakoutup> | Xitsonga | unclear |  |  |  | Prof DJ Prinsloo | danie.prinsloo@up.ac.za | See above (Pretoria Afrikaans Corpus) |
| Pretoria Zulu Corpus | <https://www.up.ac.za/african-languages/article/17933/speakoutup> | Zulu | unclear |  |  |  | Prof DJ Prinsloo | danie.prinsloo@up.ac.za | See above (Pretoria Afrikaans Corpus) |
| Recalls Cilubà Corpus | Source: <https://www.researchgate.net/figure/Corpora-of-African-languages-excluding-the-South-African-ones_tbl2_261835113> | Cilubà | unclear |  |  |  |  |  | Tools. Need to ask the authors for the corpora |
| Rosetta Disk 1.0 | <https://rosettaproject.org/disk/interactive/> | Browsable collection | unclear | Browsable Archive |  |  |  | rosetta@longnow.org | The Disk contains 13,000 pages of documentation on over 1,500 human languages – a collection of information that attests to a richness of human cultural and linguistic diversity in the year 02008. The materials in the collection were gathered from archives around the world and include different kinds of language data: descriptions of the community of speakers, maps of their location, and information on writing systems and literacy. The viewer for the digital version of the Rosetta Disk on this DVD was built by Kurt Bollacker using the OpenLayers 2.5 map visualization framework. |
| Sadilar Resource Catalogue | https://repo.sadilar.org/handle/20.500.12185/7 |  | yes |  |  |  |  |  | However: a catalogue, contains what seems to be good resources |
| Sadilar Resource Index | https://repo.sadilar.org/handle/20.500.12185/9 |  | yes |  |  |  |  |  | However: a catalogue, contains what seems to be good resources |
| SAWA Corpus | <http://www.aclweb.org/anthology/W09-0702> | English - Swahili | no |  |  |  | Guy De Pauw CNTS - Language Technology Group, University of Antwerp, Belgium School of Computing and Informatics, University of Nairobi, Kenya guy.depauw@ua.ac.be Peter Waiganjo Wagacha School of Computing and Informatics, University of Nairobi, Kenya waiganjo@uonbi.ac.ke Gilles-Maurice de Schryver African Languages and Cultures, Ghent University, Belgium Xhosa Department, University of the Western Cape, South Africa gillesmaurice.deschryver@ugent.be |  | includes New Testament (7.9 k sentences), Quran (6.2 k), Declaration of HR (0.2k), [Kamusi.org](http://kamusi.org/) (5.6 k), movie subtitles (9 k), investment reports (3.2 k in English/3.1k in Swahili), Local Translator (1.5 k/1.6 k), Full corpus: 33.6 k sentences (English and Swahili), Note: Only found the paper describing the resource |
| Scientific e-lexicography for Africa(2012-2015) | <https://www.up.ac.za/african-languages/article/38000/research-projects> |  | no | medical |  |  |  |  | Access to the data not straightforward |
| SPC - Stockholm Parallel Corpora | <https://athena.clarin.gr/resources/browse/spc-stockholm-parallel-corpora/ccb9d510a33111e5a465aa3fc9efd4927929e983ae0f48f7943ac55fd42f7d0a/> | Afrikaans - English | yes |  |  |  |  |  | For Afrikaans |
| Swahili Computer corpora | <http://www.ling.helsinki.fi/uhlcs/readme-all/README-afro-as-nig-con-lgs.html#C91> | Swahili | unclear | Consists of two corpora: 1) Swahili corpus (Fiction and news paper) and 2) Swahili dialects (interviews) |  |  | Arvi Hurskainen Institute for Asian and African Studies, Helsinki University | Arvi.Hurskainen@helsinki.fi | Need to contact the research center, the use of the corpora located at the University of Helsinki Corpus Server is restricted to concern research and teaching. Reference to the corpora has to be done in the papers in which they are used as a source. |
| The Crúbadán Project: Corpus building for under-resourced languages | <http://crubadan.org/> | Afrikaans (1307 crawled documents), Gbaya, Ndebele, Nothern Sotho, Sango, Sotughwest Gbaya, Tsonga, Venda, Xhosa, Zulu | no |  |  |  |  |  | Needs to be parallelized |
| The GlobalPhone Swahili corpus | <http://catalog.elra.info/en-us/repository/browse/ELRA-S0375/> | Swahili | yes |  | Commercial use |  | V. Mapelli | mapelli@elda.org | Not free of charge. |
| Treebanks in Universal Dependencies | [https://universaldependencies.org](https://universaldependencies.org/) | Afrikaans | yes |  | CC BY-SA 4.0 |  | Peter Dirix, Liesbeth Augustinus, Daniel van Niekerk | peter.dirix@kuleuven.be, liesbeth.augustinus@kuleuven.be | Requires transformations |
| Treebanks in Universal Dependencies | [https://universaldependencies.org](https://universaldependencies.org/) | Amharic | yes | Bible, news nonfiction | CC BY-SA 4.0 |  | Binyam Ephrem, Gashaw Arutie, Tsegay Woldemariam, Juan Ignacio Navarro Horñiacek | binephrem@gmail.com | Requires transformations, further details here: https://github.com/UniversalDependencies/UD\_Afrikaans-AfriBooms/blob/master/README.txt |
| Treebanks in Universal Dependencies | [https://universaldependencies.org](https://universaldependencies.org/) | Bambara, Wolof, Yoruba | yes | news nonfiction | CC BY-SA 4.0 |  | Katya Aplonova, Francis Tyers | zeman@ufal.mff.cuni.cz | Requires transformations, further details are provided here: <https://github.com/UniversalDependencies/UD_Bambara-CRB/blob/master/README.md> |
| Treebanks in Universal Dependencies | [https://universaldependencies.org](https://universaldependencies.org/) | Wolof | yes | bible wiki | CC BY-SA 4.0 |  | Bamba Dione | dione.bamba@uib.no | Requires transformations, UD\_Wolof-WTB is a natively manual developed treebank for Wolof. Sentences were collected from encyclopedic, fictional, biographical, religious texts and news. |
| Treebanks in Universal Dependencies | [https://universaldependencies.org](https://universaldependencies.org/) | Yoruba | yes | bible wiki | CC BY-SA 4.0 |  | Adédayọ̀ Olúòkun, Daniel Zeman, Seyi Williams, Ọlájídé Ishola | zeman@ufal.mff.cuni.cz | Requires transformations |
| Ubuntu | <http://opus.nlpl.eu/Ubuntu-v14.10.php> | English - Igbo, Afrikaans, Hausa | yes |  |  |  |  |  |  |
| ukuxhumana | <https://github.com/LauraMartinus/ukuxhumana> |  | yes |  |  |  |  |  |  |
| Unsupervised compound splitter for Afrikaans | <https://www.aclweb.org/anthology/N16-1078.pdf> | Afrikaans | unclear |  |  |  | Patrick Ziering, Lonneke van der Plas | Patrick.Ziering@ims.uni-stuttgart.de, Lonneke.vanderPlas@um.edu.mt |  |
| Voice of Nigeria Yorùbá news | [von.gov.ng/yoruba](http://von.gov.ng/yoruba) | Yorùbá | yes |  |  |  |  |  |  |
| Wikipedia | [dumps.wikimedia.org/twwiki](http://dumps.wikimedia.org/twwiki) | Twi | yes | read texts from national newspapers |  |  |  |  | Text needs to be made available |
| XhosaNavy | <http://opus.nlpl.eu/XhosaNavy.php> | Xhosa - English | yes | pronunciations of all word forms found in the transcription data of the GlobalPhone speech & text database. |  | Please cite the following article if you use any part of the corpus in your own work: J. Tiedemann, 2012, Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012) | Herman Engelbrecht, Department of E&E Engineering at Stellenbosch University |  | 2 languages, total number of files: 2; total number of tokens: 1.45M; total number of sentence fragments: 0.10M |
| Yorúbà Bible | [www.bible.com](http://www.bible.com/) | Yorùbá | yes | prose text from fiction, news media and government documents domains |  |  |  |  |  |
| Yorùbá Tweets | [twitter.com/yobamoodua](http://twitter.com/yobamoodua) | Yorùbá | yes |  |  |  |  |  |  |
| Yorùbá Wikipedia | [dumps.wikimedia.org/yowiki](http://dumps.wikimedia.org/yowiki) | Yorùbá | yes |  |  |  |  |  | However: text needs to be made available |
| Zulu Wikimedia | <https://ftp.acc.umu.se/mirror/wikimedia.org/dumps/zuwiki/> | Zulu | yes | Legal nonfiction |  |  |  |  | Data needs to be extracted from the dump |

Figure 8: Non-extensive summary of African language data sets

## Registration List

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Last Name** | **First Name** | **Affiliation** |
|  | Abbott | Jade | Retro Rabbit |
|  | Adelani | David | Saarland University & Masakhane |
|  | Agbeko | Henry | Kwame Nkrumah University of Science and Technology |
|  | Agyapon-Ntra | Kwadwo | Self-attending |
|  | Ahia | Orevaoghene | Instadeep |
|  | Ajibade | Faith | University of Ibadan |
|  | Akinkunmi | Anuoluwapo | Student |
|  | Alabi | Jesujoba | DFKI |
|  | Ali | Jamiil Touré | Takwimu Lab |
|  | Amanfu | Richard | Institute of ICT Professionals Ghana |
|  | Anebi | Emmanuel | DataInsight |
|  | Anzagira | Allan re | North Carolina A &T State University |
|  | Anzaldo | Isa | UdS |
|  | Appati | Justice Kwame | University of Ghana |
|  | Aremu | Anuoluwapo | Aremu Language Consult |
|  | Asamoah | Eugene | Electricity Company Of Ghana |
|  | Ayami | Yasin | TsogoloTech |
|  | Bamutura | David | Mbarara University of Science and Technology |
|  | Barnes | Samuel | Student |
|  | Berejena | Beatrice | NA |
|  | Betty | Betty | University of Johannesburg |
|  | Blum | Seth | Meridian Institute |
|  | Boateng | Samuel | Ajuma software |
|  | Bridgman | Grant | Uliza |
|  | Brütting | Florian | GIZ |
|  | Buabeng | Edwin | Huawei |
|  | Budu | Joel | Artificial Intelligence Association of Ghana |
|  | Byamugisha | Joan | IBM Research Africa |
|  | Cayralat | Christian | LCT |
|  | Chirwa | Temweka | University of Cape Town |
|  | Clancy | Katie | IDRC |
|  | Coffie Debrah | Emmanuel | University of Cape Coast |
|  | David | Davis | TYD Innovation Incubator |
|  | Degila | Kevin | Masakhane |
|  | Diallo | Aboubacar | Independant Consultant |
|  | Dietrich | Klakow | Saarland University |
|  | Donner | Jonathan | Caribou digital |
|  | Dosseh | Desire | Dakar Institute of Technology |
|  | du Plessis | Liëtte | University of Johannesburg |
|  | du Toit | Jaco | UNESCO |
|  | Dube | Hloniphani | Bluemachines (Pty) Ltd |
|  | Dutta | Sourav | Saarland University |
|  | Dzidzinyo | Komla Tekpo Abah | INP-HB |
|  | Ekem | Ivy | UCC SMS |
|  | Elgizouli | Mukhtar | University of khartoum |
|  | Elmers | Mikey | University of Saarland |
|  | Eneremadu | Sydney | Chatbot Africa & Conversational AI Summit |
|  | España-Bonet | Cristina | DFKI |
|  | Estarrona | Ainara | HiTZ zentroa (UPV/EHU) |
|  | Gaelejwe | Theodore | IBM Research |
|  | Gardent | Claire | CNRS |
|  | Gebremeskel | Gebrekirstos | CWI |
|  | Gesicho | Rose Delilah | Student |
|  | Gimpel | Lea | GIZ |
|  | Gova | Webster | Umuzi |
|  | Graaf | Michael | Wikimedia-ZA |
|  | Griciūtė | Bernadeta | Saarland University |
|  | Gyamfi | Nana Kwame | Kumasi Technical University |
|  | Gyekye | Kwame | Experian |
|  | Ibrahim | Omnia | University of Zürich |
|  | Ihle | Frank | private |
|  | Iminza | Diana | JKUAT |
|  | Israel | Odeajo | University of Ibadan, Nigeria |
|  | Jarso | Guyo | University of Rwanda |
|  | Jobe | Wuyeh | Carnegie Mellon University Africa |
|  | Kalmykova | Anastasiia | Uds |
|  | Kashupi | Tokolo N | Namibia University of Science and Technology |
|  | Katumba | Andrew | Makerere University |
|  | Kaye | Jofish | Mozilla |
|  | Kiden | Sarah | Northumbria University |
|  | Kintu Mwanje | Timothy | Mbarara University of Science and Technology |
|  | Kioko | Moses |  |
|  | Konobelkina | Anna | University of Saarland |
|  | Koupoh | Esaïe Alain-Emmanuel | Dakar Institute of Technology |
|  | Ladipo | Taiwo | SAP |
|  | Lang | Inga | EM Double Master in Language and Communication Technologies (LCT), University of Groningen and University of Malta |
|  | Lösch | Andrea | DFKI |
|  | Luberenga | John | Flying Gravity Technologies |
|  | Lubrini | Elisa | Université de Lorraine |
|  | Mabaso | Mbangiso | Sisanda Tech |
|  | Mahlezana | Thiba | TGM MakerSpace |
|  | Makumbirofa | Hamony | Tshimologong |
|  | Malatji | Masike | University of Johannesburg |
|  | Manishimwe | Alban | The Nelson Mandela African Institution of Science and Technology |
|  | Mansour | Ayman | Sudan university of science and technology |
|  | Manzi | Innocent | 250STARTUPS |
|  | Matana | Maduhu | Platinum Credit |
|  | Mbadi | Sheila | CMU-Africa |
|  | Mbaye | Derguene | Baamtu |
|  | Mbogho | Audrey | USIU-Africa |
|  | Merci Hategekimana | Arsène | College student |
|  | Mintz | Blanca | GIZ |
|  | Mokele | David | CS SMART |
|  | Moore | Stephen | Ghana NLP |
|  | Moruye | Lawrence | Multimedia University of kenya |
|  | Moyo | Mlamuleli | Tmg makers space |
|  | Moyo | Simon | Mafikeng digital innovation hub |
|  | Mthembu | Nkululeko | Private |
|  | Mugambi | Jonan | Volkswagen Mobility Solutions Rwanda |
|  | Muhire | Remy | Mozilla |
|  | Muite | Benson | - |
|  | Mukiibi | Jonathan | AI Lab Makerere University |
|  | Munro | Robert | Machine Learning Consulting |
|  | Musoya | Gael | Digitech group |
|  | Mwatukange | Joseph | Konga Technologies |
|  | N/A | Ari | Praekelt |
|  | Nakatumba | Joyce | Makerere University |
|  | Navisa | Salma | UIN Sunan Ampel |
|  | NDIAYE | Malick | Bambey University/ Senegal |
|  | Ndodana | Nwabisa | Student |
|  | Nixon | Samantha | IDRC |
|  | Niyongabo | Rubungo Andre | Masakhane |
|  | Nordor | Eli | BlueCrest |
|  | Nxumalo | Thabiso | North Park Telecoms |
|  | Oduor | Clinton | clintonoduor3@outlook.com |
|  | Ogayo | Perez | African Leadership University |
|  | Oguche | Victor | Daptem Engineering |
|  | Ogundepo | Odunayo | Deloitte & Touche Nigeria Limited |
|  | Ojesina | Akolade | University of Ibadan |
|  | Ọládípọ̀ | Akíntúndé | Wragby Business Solutions and Technologies Limited |
|  | Olbrich | Phillipp | GIZ |
|  | Oloke | Adewale | Unilag |
|  | Oluyori | Elkanah | Reformers of Africa |
|  | Orlic | Davor | Knowledge 4 All Foundation |
|  | Owusu-Darko | Ama | Ensign College of Public Health |
|  | Pakzad | Roya | Taraaz |
|  | Patrick | Emil | Nelson Mandela African Institution of Science and Technology |
|  | Phahlamohlaka | Lazarus | North Park Telecoms |
|  | Pienaar | Marne | UJ |
|  | Pratt Miles | Jennifer | Lacuna Fund: Our Voice on Data |
|  | Prenom | Daouda | Université Alioune Diop de Bambey |
|  | Quayson | Ebenezer | UENR |
|  | Ramos | Alex | N/A |
|  | Resch | Christian | GIZ |
|  | Rutunda | Samuel | Digital Umuganda |
|  | Sam | Abraham | FAIR Forward, GIZ |
|  | Samb | Sokhar | AIMS |
|  | Schnur | Eileen | DFKI |
|  | Schonwetter | Tobias | University of Cape Town |
|  | Schulte | Kim | GIZ |
|  | Seibold | Balthas | GIZ |
|  | Sibal | Prateek | UNESCO |
|  | Siminyu | Kathleen | Artificial Intelligence for Development - Africa |
|  | Smith | Blake | University of Edinburgh |
|  | Smith | Matthew | International Development Research Centre |
|  | Sorinolu | Babafemi | Dominican University |
|  | Tembo | David | VisionAi |
|  | Tsouapi | Rigobert | aims |
|  | Turikumwe | Jean Dela paix | A. I. S |
|  | Tyira | Loyiso | ICT SMME Chamber |
|  | Umuhire | Evelyne | WeCode Moringa School |
|  | Urosevic | Jovana | Independent |
|  | van Genabith | Josef | DFKI |
|  | Wanzare | Lilian | Maseno University |
|  | Waterfield | Rebecca | N/A |
|  | Watson | Sarah | Mozilla |
|  | Wilson | Daniel | XRI |
|  | Wilson | Joe | J-MO Global |
|  | Yeboah-Boateng | Dr. Ezer Osei | Ghana Technology University College |

Figure 9: List of Registrations

# References

[1] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean,J. (2013). *Distributed representations of words and phrases and their compositionality*. Advances in Neural Information Processing System, pages 3111–3119.

[2] Pennington, J., Socher, R., and Manning, C. D. (2014). *Glove: Global vectors for word representation*. In Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543.

[3] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). *Enriching word vectors with subword information*. Transactions of the Association for Computational Linguistics, 5:135–146.

[4] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark,C., Lee, K., and Zettlemoyer, L. (2018). *Deep contextualized word representations*. In Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana, June.

[5] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.(2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, pages 4171–4186, Minneapolis, Minnesota, June.

[6] Artetxe, M. and Schwenk, H. (2019). M*assively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond*. Volume 7, pages 597–610. MIT Press, September.

[7] Lample, G. and Conneau, A. (2019). *Cross-lingual language model pre- training*. Advances in Neural Information Processing Systems (NeurIPS).

[8] Alabi, J. O., Amponsah-Kaakyire, K., Adelani, D. A. and España-Bonet, C. (2020). *Massive vs. Curated Word Embeddings for Low-Resourced Languages. The Case of Yorùbá andTwi*. In Proceedings of the International Conference on Language Resources and Evaluation, pages 2754-2762, Marseille, France, May.

[9] Ruder, S. (2019). *Neural Transfer Learning for Natural Language Processing*. PhD thesis, National University of Ireland, Galway.

[10] Lauscher, A., Ravishankar, V., Vulic, I. and Glavas, G. (2020). *From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers*. ArXiv, abs/2005.00633.

[11] Tran, K. (2020). *From English to Foreign Languages: Transferring Pretrained Language Models*. ArXiv, abs/2002.07306.