

# FAIR FORWARD

Artificial Intelligence for all.

## Work Package 2

# Report on Webinar “AI 4D Language Challenge”

Author(s): Cristina España i Bonet (DFKI)  
Andrea Lösch (DFKI)  
Eileen Schnur (DFKI)

Dissemination Level: Confidential

Version No.: <V1.1>

Date: 2020-07-09



Deutsches  
Forschungszentrum  
für Künstliche  
Intelligenz GmbH

## Table of Contents

<b>1</b>	<b>Executive Summary</b>	<b>3</b>
<b>2</b>	<b>Agenda Webinar</b>	<b>4</b>
<b>3</b>	<b>Research in Natural Language Processing (NLP)</b>	<b>5</b>
3.1	<i>State-of-the-art NLP research</i>	5
3.2	<i>NLP research in low-resourced languages</i>	6
<b>4</b>	<b>Data for African languages</b>	<b>8</b>
4.1	<i>Overview of existing data sets for African languages</i>	8
4.1	<i>Data collection approaches</i>	9
4.1.1	Identification of and collaboration with language data holders	9
4.1.2	Identification and use of sources of language data	10
4.1.3	Making language data re-usable	10
4.2	<i>Important references</i>	11
4.3	<i>Status quo: Survey results</i>	12
<b>5</b>	<b>Annex</b>	<b>15</b>
5.1	<i>Participants List</i>	15
5.2	<i>Presentations</i>	17
5.3	<i>Summary of African Language Data Sets</i>	18
5.4	<i>Registration List</i>	34
<b>6</b>	<b>References</b>	<b>39</b>

## List of Figures

Figure 1: Semantic relations of embeddings as described by Mikolov et al. (2003)	5
Figure 2: Taxonomy for transfer learning in natural language processing as defined by Ruder (2019)	7
Figure 3: Distribution of responses to Question 1	13
Figure 4: Distribution of responses to Question 2	13
Figure 5: Distribution of responses to Question 3	14
Figure 6: Distribution of responses to Question 4	14
Figure 7: Participants List	17
Figure 8: Non-extensive summary of African language data sets	33
Figure 9: List of Registrations	38

## 1 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/2hGivcA1WbbrrtEZu9>

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

## 2 Agenda Webinar

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

### 3 Research in Natural Language Processing (NLP)

#### 3.1 State-of-the-art NLP research

Language technologies in general have experienced 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.

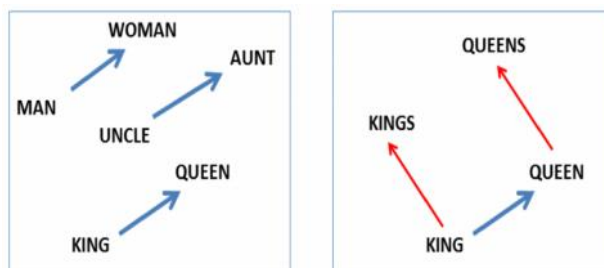


Figure 1: Semantic relations of embeddings as described by Mikolov et al. (2003)

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.

### 3.2 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).

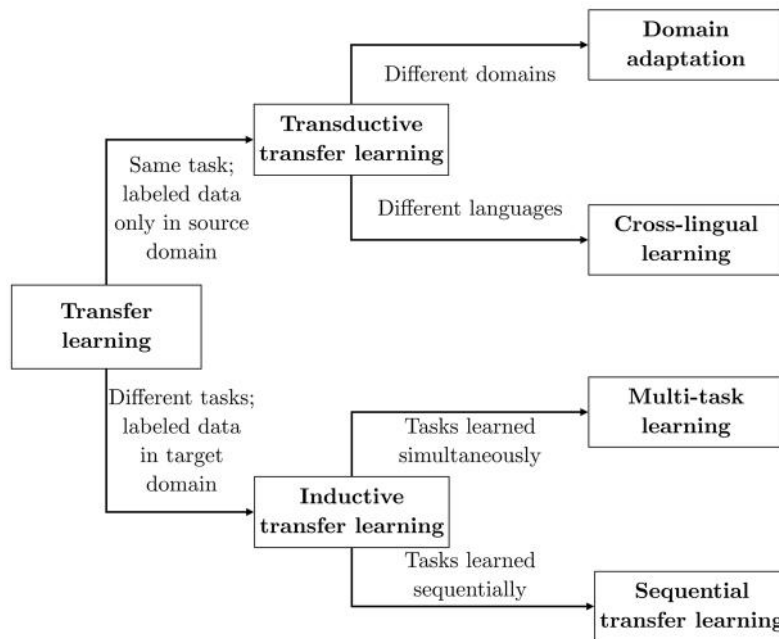


Figure 2: Taxonomy for transfer learning in natural language processing as defined by Ruder (2019)

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.

## 4 Data for African languages

### 4.1 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

#### 4.1 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.

##### 4.1.1 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, a collection of South African translators is also available here),
- 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, South African Centre for Digital Language Resources (SADiLaR)),

- African and/or international open data portals (e.g. openAfrica),
- African language and/or language technology researchers and members of academia (e.g. AfricArxiv, African Academy of Languages (ACALAN)),
- African and/or international language technology and language service providers (e.g. Translate4Africa, Folio Online).

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).

#### 4.1.2 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.

#### 4.1.3 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.

## 4.2 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.

### 4.3 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.

**Question 1: Low-resourced languages are languages**

51 / 53 correct responses

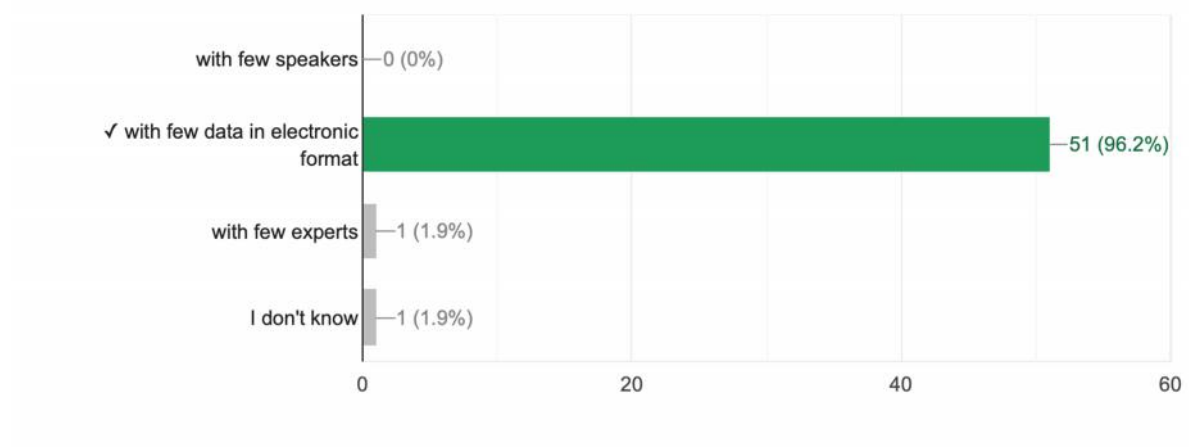


Figure 3: Distribution of responses to Question 1

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

28 / 53 correct responses

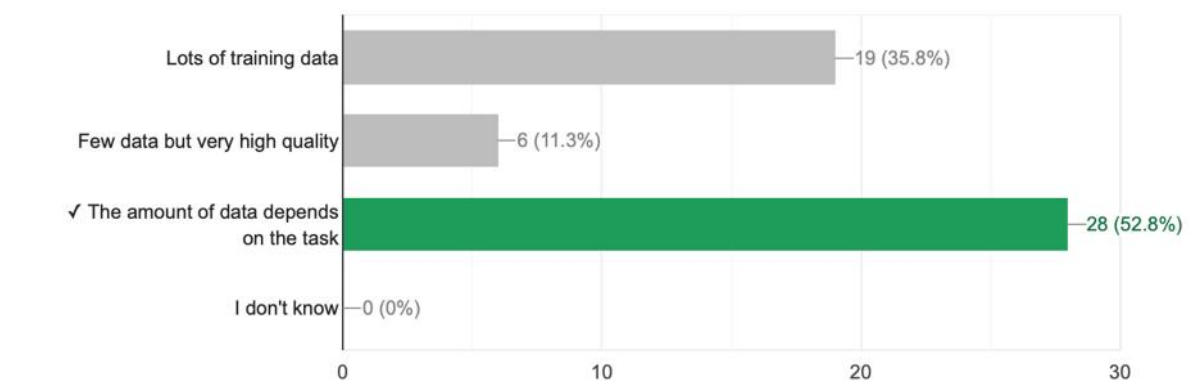


Figure 4: Distribution of responses to Question 2

**Question 3: Resources from English can be used for Ewe**

39 / 53 correct responses

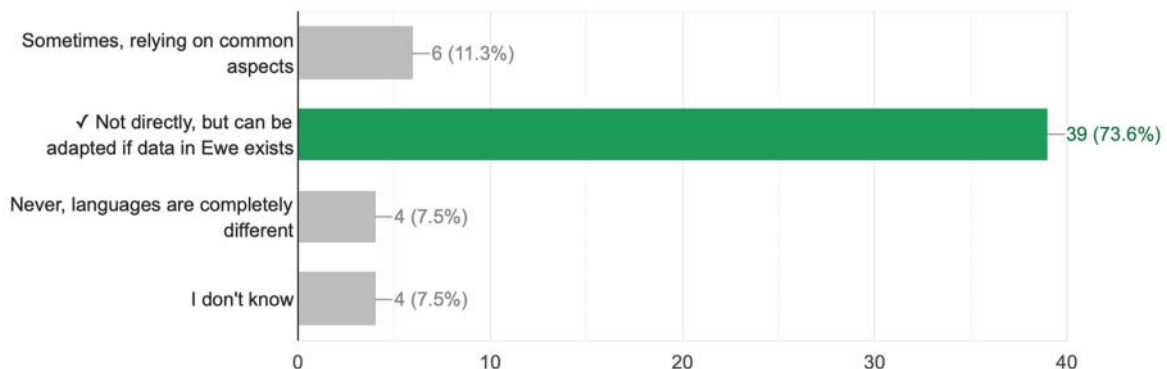


Figure 5: Distribution of responses to Question 3

**Question 4: Transfer learning can be used in NLP to**

30 / 53 correct responses

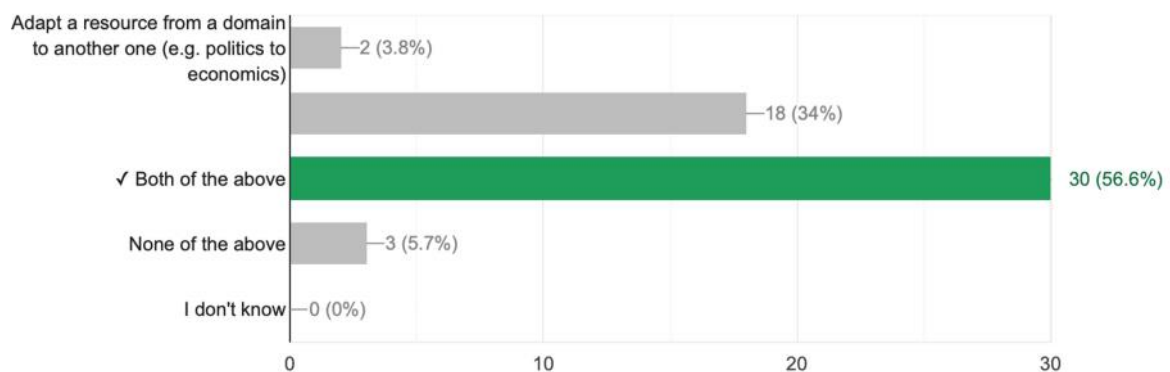


Figure 6: Distribution of responses to Question 4

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.

## 5 Annex

### 5.1 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
1	Addo*	Salomey	N/A
2	Adelani	David	Saarland University & Masakhane
3	Agyapon-Ntra	Kwadwo	Self-attending
4	Ahia	Orevaoghene	Instadeep
5	Alabi	Jesujoba	DFKI
6	Amanfu	Richard	Institute of ICT Professionals Ghana
7	Anebi	Emmanuel	DataInsight
8	Anzaldo	Isa	UdS
9	Asamoah	Eugene	Electricity Company Of Ghana
10	Ayami	Yasin	TsogoloTech
11	Bamutura	David	Mbarara University of Science and Technology
12	Berejena	Beatrice	NA
13	Betty	Betty	University of Johannesburg
14	Buabeng	Edwin	Huawei
15	Budu	Joel	Artificial Intelligence Association of Ghana
16	Byamugisha	Joan	IBM Research Africa
17	Cayralat	Christian	LCT
18	Chirwa	Temweka	University of Cape Town
19	Coffie Debrah	Emmanuel	University of Cape Coast
20	David	Davis	TYD Innovation Incubator
21	Dietrich	Klakow	Saarland University
22	Donner	Jonathan	Caribou digital
23	Dosseh	Desire	Dakar Institute of Technology
24	du Plessis	Liëtte	University of Johannesburg
25	Dube	Hlonipani	Bluemachines (Pty) Ltd
26	Ekem	Ivy	UCC SMS
27	Elmers	Mikey	University of Saarland
28	España-Bonet	Cristina	DFKI
29	Estarrona	Ainara	HiTZ zentroa (UPV/EHU)
30	Gaelejwe	Theodore	IBM Research
31	Gardent	Claire	CNRS
32	Gebremeskel	Gebrekirstos	CWI
33	Gesicho	Rose Delilah	Student
34	Graaf	Michael	Wikimedia-ZA
35	Griciūtė	Bernadeta	Saarland University

36	Gyekye	Kwame	Experian
37	Ibrahim	Omnia	University of Zürich
38	Ihle	Frank	private
39	Jarso	Guyo	University of Rwanda
40	Kashupi	Tokolo N.	Namibia University of Science and Technology
41	Kiden	Sarah	Northumbria University
42	Kioko	Moses	
43	Konobelkina	Anna	University of Saarland
44	Ladipo	Taiwo	SAP
45	Lang	Inga	EM Double Master in Language and Communication Technologies (LCT), University of Groningen and University of Malta
46	Lösch	Andrea	DFKI
47	Luberenga	John	Flying Gravity Technologies
48	Lubrini	Elisa	Université de Lorraine
49	Mahamah*	Rhoda	N/A
50	Mahlezana	Thiba	TGM MakerSpace
51	Makumbirofa	Hamony	Tshimologong
52	Malatji	Masike	University of Johannesburg
53	Mansour	Ayman	Sudan university of science and technology
54	Mbadi	Sheila	CMU-Africa
55	Mintz	Blanca	GIZ
56	Moore	Stephen	Ghana NLP
57	Moyo	Mlamuleli	Tmg makers space
58	Mthembu	Nkululeko	Private
59	Muhire	Remy	Mozilla
60	Muite	Benson	N/A
61	Mukiibi	Jonathan	AI Lab Makerere University
62	Musoya	Gael	Digitech group
63	N/A*	Ari	Praekelt
64	N/A*	Ajamitoure	N/A
65	N/A*	Celina	N/A
66	N/A*	Derick	N/A
67	N/A*	Guest	N/A
68	N/A*	Peter	N/A
69	Nakatumba	Joyce	Makerere University
70	Ndodana	Nwabisa	Student
71	Nxumalo	Thabiso	North Park Telecoms
72	Oduor	Clinton	clintonoduor3@outlook.com
73	Ogayo	Perez	African Leadership University
74	Oghenekevwe*	Ajewole	N/A



75	Ojesina	Akolade	University of Ibadan
76	Oloke	Adewale	Unilag
77	Orlic	Davor	Knowledge 4 All Foundation
78	Owusu-Darko	Ama	Ensign College of Public Health
79	Oyewole*	Isaiah	N/A
80	Phahlamohlaka	Lazarus	North Park Telecoms
81	Pienaar	Marne	UJ
82	Pratt Miles	Jennifer	Lacuna Fund: Our Voice on Data
83	Prenom	Daouda	Université Alioune Diop de Bambey
84	Quayson	Ebenezer	UENR
85	Resch	Christian	GIZ
86	Sam	Abraham	FAIR Forward, GIZ
87	Schnur	Eileen	DFKI
88	Schonwetter	Tobias	University of Cape Town
89	Sibal	Prateek	UNESCO
90	Siddharth*	Nandan	GIZ
91	Siminyu	Kathleen	Artificial Intelligence for Development - Africa
92	Smith	Blake	University of Edinburgh
93	Smith	Matthew	International Development Research Centre
94	van Genabith	Josef	DFKI
95	Vorsah Amponsah*	Irene Kafui	N/A
96	Wanzare	Lilian	Maseno University
97	Waterfield	Rebecca	n/a
98	Watson	Sarah	Mozilla
99	Wilson	Daniel	XRI

Figure 7: Participants List

## 5.2 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>

### 5.3 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	<a href="https://dataportal.opendata-forafrica.org/data#menu=topic">https://dataportal.opendata-forafrica.org/data#menu=topic</a>		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	<a href="https://rma.nwu.ac.za/index.php/resource-catalogue/ast-corpus-isizulu.html">https://rma.nwu.ac.za/index.php/resource-catalogue/ast-corpus-isizulu.html</a>	Zulu	unclear						Potential risk by access
AfricArxiv	<a href="https://info.africarxiv.org/">https://info.africarxiv.org/</a>		unclear	digital archive for African research communication					Seems to be more in English
Alákòwé	<a href="http://alakoweyoruba.wordpress.com">alakoweyoruba.wordpress.com</a>	Yorùbá	no						
ALLEX - Ndebele Corpus	<a href="http://www.edd.uio.no/allex/corpus/africanlang.html">http://www.edd.uio.no/allex/corpus/africanlang.html</a>	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	<a href="http://www.edd.uio.no/allex/corpus/africanlang.html">http://www.edd.uio.no/allex/corpus/africanlang.html</a>	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	<a href="https://rma.nwu.ac.za/index.php/autshumato-eng-zu-parallel-corpora.html">https://rma.nwu.ac.za/index.php/autshumato-eng-zu-parallel-corpora.html</a>	English - Zulu, Setswana, Xitsonga, Northern-Sotho, Afrikaans	no						Potential risk by access
BBC Yorùbá	<a href="http://bbc.com/yoruba">bbc.com/yoruba</a>	Yorùbá	yes						However: text needs to be made available
Benin Open Data Portal	<a href="https://benin.opendataforafrica.org">https://benin.opendataforafrica.org</a>		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	<a href="http://www.bible.com">www.bible.com</a>	Twi	yes						However: text needs to be made available
bible-uedin	<a href="http://opus.nlpl.eu/bible-uedin.php">http://opus.nlpl.eu/bible-uedin.php</a>		yes				Christos Christodoulopoulos, Mark Steedman		Multilingual parallel corpus created from translations of the bible; However: need to extract the African languages; <a href="https://link.springer.com/article/10.1007/s10579-014-9287-y">https://link.springer.com/article/10.1007/s10579-014-9287-y</a>
CORAAL - Corpus of Regional African American Language	<a href="https://oraal.uoregon.edu/coraal">https://oraal.uoregon.edu/coraal</a>		no						Specific to American African Language (mostly speech)
CorpAfroAs	<a href="https://corpafroas.humnum.fr/Archives/corpus.php">https://corpafroas.humnum.fr/Archives/corpus.php</a>		unclear	Spoken Afroasiatic languages	"You are welcome to use the CorpAfroAs Format and Tools for your data., Please		Azeb Amha, Christian Chanard	a.amha@hum.leidenuniv.nl, christian.chanard@cnsr.fr	Need to contact the research center, manual is available here: <a href="https://corpafroas.humnum.fr/fichiers/manual.pdf">https://corpafroas.humnum.fr/fichiers/manual.pdf</a>

					quote the CorpAfroA s project when you use our annotation scheme and/or software and/or procedures. Thank you."				
Corpus Bambara de Référence	<a href="http://cormand.huma-num.fr">http://cormand.huma-num.fr</a>	Bambara	yes				Valentin Vydrine	vydrine@gmail.com	Makes a good impression, but unclear how to download, link to the corpus: <a href="http://cormande.huma-num.fr/cor-bama/run.cgi/first_form">http://cormande.huma-num.fr/cor-bama/run.cgi/first_form</a>
Corpus Maninka	<a href="http://cormand.huma-num.fr/cormani/">http://cormand.huma-num.fr/cormani/</a>	Maninka	unclear				Valentin Vydrin	vydrine@gmail.com	Corpus is available here: <a href="http://cormande.huma-num.fr/cor-mani/run.cgi/first_form">http://cormande.huma-num.fr/cor-mani/run.cgi/first_form</a>
Corpus of South African English (CoSAE)	Source: <a href="https://www.researchgate.net/figure/Corpora-of-African-languages-excluding-the-South-African-ones_tbl2_261835113">https://www.researchgate.net/figure/Corpora-of-African-languages-excluding-the-South-African-ones_tbl2_261835113</a>	English	unclear						
de Schryver and Prinsloo: The compilation of electronic corpora, with special reference to the African Languages (2000)	<a href="https://tshwanedje.com/publications/Corpora.pdf">https://tshwanedje.com/publications/Corpora.pdf</a>		no						Publication only. If access to the data can be organised, this would be a good source.

Die Pharos-korpus van hede-ndaagse Afrikaans (PAK)	Source: <a href="https://www.researchgate.net/figure/Corpora-of-African-languages-excluding-the-South-African-ones_tbl2_261835113">https://www.researchgate.net/figure/Corpora-of-African-languages-excluding-the-South-African-ones_tbl2_261835113</a>	Afrikaans	unclear						
DOBES - Documentation of endangered languages	<a href="https://dobes.mpi.nl/#">https://dobes.mpi.nl/#</a>	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	<a href="https://github.com/Niger-Volta-LTI">github.com/Niger-Volta-LTI</a>	Yorùbá	yes						
Èdè Yorùbá Rewà	<a href="https://deskgram.cc/edeyorubarewaa">deskgram.cc/edeyorubarewaa</a>	Yorùbá	no						
ELAR - Endangered Language Archive	<a href="https://www.soas.ac.uk/elar/">https://www.soas.ac.uk/elar/</a>	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	<a href="https://www.aflat.org/node/86">https://www.aflat.org/node/86</a>	English - Luganda	no						Found <a href="#">here</a>

English - Luo Machine Translation System	<a href="https://www.aflat.org/luomt">https://www.aflat.org/luomt</a>	English - Luo (Dholuo)	no						Found <a href="#">here</a> , a system (offline now), but not a resource
Global Voices	<a href="https://yo.globalvoices.org">yo.globalvoices.org</a>	Yorùbá	yes						However: not directly re-usable. We know that some data are made available
GlobalPhone Hausa Pronunciation Dictionary	<a href="http://catalog.elra.info/en-us/repository/browse/ELRA-S0353/">http://catalog.elra.info/en-us/repository/browse/ELRA-S0353/</a>	Hausa	yes		Commercial use		V. Mapelli	mapelli@elda.org	Not free of charge
GlobalPhone Swahili Pronunciation Dictionary	<a href="http://catalog.elra.info/en-us/repository/browse/ELRA-S0376/">http://catalog.elra.info/en-us/repository/browse/ELRA-S0376/</a>	Swahili	yes		Commercial use		V. Mapelli	mapelli@elda.org	Not free of charge
GNOME	<a href="http://opus.nlpl.eu/GNOME-v1.php">http://opus.nlpl.eu/GNOME-v1.php</a>	English - Igbo, Afrikaans, Hausa	yes						
haWaC: Hausa corpus from the Web	<a href="https://www.sketchengine.eu/hawac-hausa-corpus/">https://www.sketchengine.eu/hawac-hausa-corpus/</a>	Hausa	yes						
Helsinki Corpus of Swahili	<a href="http://catalog.elra.info/en-us/repository/browse/ELRA-W0119/">http://catalog.elra.info/en-us/repository/browse/ELRA-W0119/</a>	Swahili	yes		Commercial use		V. Mapelli	mapelli@elda.org	Not free of charge
Helsinki Corpus of Swahili 2.0 (HCS 2.0)	<a href="http://metashare.csc.fi/repository/browse/helsinki-corpus-of-swahili-20-hcs-20-annotated-version/232c1910b9eb11e5915e005056be118e59fb2e920f1f4c0cafc94915fc6f5cac/">http://metashare.csc.fi/repository/browse/helsinki-corpus-of-swahili-20-hcs-20-annotated-version/232c1910b9eb11e5915e005056be118e59fb2e920f1f4c0cafc94915fc6f5cac/</a>	Swahili	yes						Also available here: <a href="http://catalog.elda.org/en-us/repository/browse/ELRA-W0119/">http://catalog.elda.org/en-us/repository/browse/ELRA-W0119/</a>
igTenTen: Igbo corpus from the web	<a href="https://www.sketchengine.eu/ig-tenten-igbo-corpus/">https://www.sketchengine.eu/ig-tenten-igbo-corpus/</a>	Igbo	yes						

Jehova Witness	<a href="http://www.jw.org/yo">www.jw.org/yo</a>	Yorùbá	yes		CC-BY-NC-SA.				However: text needs to be made available
Jehova Witness	<a href="http://www.jw.org/tw">www.jw.org/tw</a>	Twi	yes						
JW300 corpus	<a href="http://opus.nlpl.eu/JW300.php">opus.nlpl.eu/JW300.php</a>	Twi, Yorùbá	yes		CC-BY-NC-SA.				
KDE4 v2	<a href="http://opus.nlpl.eu/KDE4-v2.php">http://opus.nlpl.eu/KDE4-v2.php</a>	English - Afrikaans, Hausa	yes						
Kenia Open Data Portal	<a href="http://www.opendata.go.ke">http://www.opendata.go.ke</a>		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: <a href="https://www.researchgate.net/figure/Corpora-of-African-languages-excluding-the-South-African-ones_tbl2_261835113">https://www.researchgate.net/figure/Corpora-of-African-languages-excluding-the-South-African-ones_tbl2_261835113</a>	Kiswahili	unclear						Tools. Need to ask the authors for the corpora
Lacito	<a href="https://pangloss.cnrs.fr/corpus/index.html">https://pangloss.cnrs.fr/corpus/index.html</a>	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	<a href="https://github.com/Niger-Volta-LTI">github.com/Niger-Volta-LTI</a>	Yorùbá	yes						

Leibniz Corpora Collection	<a href="http://corpora.uni-leipzig.de/en?corpusId=ibo_community_2017">http://corpora.uni-leipzig.de/en?corpusId=ibo_community_2017</a>	Igbo, Hausa, Swahili	no						The Igbo corpus is available here: <a href="https://curl.corpora.uni-leipzig.de/languages/ibo">https://curl.corpora.uni-leipzig.de/languages/ibo</a> , 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	jba-denhorst@csir.co.za, mdavel@csir.co.za, cvheerden@csir.co.za, ebar-nard@csir.co.za	
memat	<a href="http://opus.nlpl.eu/memat.php">http://opus.nlpl.eu/memat.php</a>	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	Jörg Tiedemann		Xhosa-English parallel corpora, funded by EPSRC, the Medical Machine Translation project worked on machine translation between isiXhosa 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



						Language Resources and Evaluation (LREC 2012)			
NCHLT isiZulu Text Corpus	<a href="https://rma.nwu.ac.za/index.php/isizulu-nchlt-text-corpora.html">https://rma.nwu.ac.za/index.php/isizulu-nchlt-text-corpora.html</a>	Zulu	unclear						Potential risk by access
Northern Sotho Part-of-Speech Tagger (V2) - Demo	<a href="https://www.aflat.org/node/177">https://www.aflat.org/node/177</a>	Northern Sotho	no						found <a href="#">here</a> . Tools rather than data
Onyenwe, Uchechukwu, Hepple: Part-of-speech Tagset and Corpus Development for Igbo, an African Language (2014)	<a href="http://www.aclweb.org/anthology/W14-4914">http://www.aclweb.org/anthology/W14-4914</a>	Igbo	no						Paper only.
Open-access portal for data protection laws and information in Africa	<a href="https://dataprotection.africa">https://dataprotection.africa</a>	Afrikaans, Chichewa, Igbo, Sesotho, Shona, Afsoomaali, Basa Sunda, Kiswahili, Xhosa, Yorùbá, Zulu	no	Legal			Justin Bryant (research and coordination)	dataprotection@altaadvisory.africa	Seems to be more in English, Data Protection Africa is an ALT Advisory special project.
openAfrica Portal	<a href="https://africaopendata.org">https://africaopendata.org</a>		unclear	Open Data Portal					Seems to be more in English
Òrò Yorùbá	<a href="http://oroyoruba.blogspot.com">oroyoruba.blogspot.com</a>	Yorùbá	no						

Pretoria Afrikaans Corpus	<a href="https://www.up.ac.za/african-languages/article/17933/speakoutup">https://www.up.ac.za/african-languages/article/17933/speakoutup</a>	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	<a href="https://www.up.ac.za/african-languages/article/17933/speakoutup">https://www.up.ac.za/african-languages/article/17933/speakoutup</a>	Xhosa	unclear				Prof DJ Prinsloo	danie.prinsloo@up.ac.za	See above (Pretoria Afrikaans Corpus)
Pretoria Ndebele Corpus	<a href="https://www.up.ac.za/african-languages/article/17933/speakoutup">https://www.up.ac.za/african-languages/article/17933/speakoutup</a>	Ndebele	unclear				Prof DJ Prinsloo	danie.prinsloo@up.ac.za	See above (Pretoria Afrikaans Corpus)
Pretoria Sepedi Corpus	<a href="https://www.up.ac.za/african-languages/article/17933/speakoutup">https://www.up.ac.za/african-languages/article/17933/speakoutup</a>	Sepedi	unclear				Prof DJ Prinsloo	danie.prinsloo@up.ac.za	See above (Pretoria Afrikaans Corpus)
Pretoria Sesotho Corpus	<a href="https://www.up.ac.za/african-languages/article/17933/speakoutup">https://www.up.ac.za/african-languages/article/17933/speakoutup</a>	Sesotho	unclear				Prof DJ Prinsloo	danie.prinsloo@up.ac.za	See above (Pretoria Afrikaans Corpus)
Pretoria Setswana Corpus	<a href="https://www.up.ac.za/african-languages/article/17933/speakoutup">https://www.up.ac.za/african-languages/article/17933/speakoutup</a>	Setswana	unclear				Prof DJ Prinsloo	danie.prinsloo@up.ac.za	See above (Pretoria Afrikaans Corpus)

Pretoria Swati Corpus	<a href="https://www.up.ac.za/african-languages/arti-cle/17933/speakoutup">https://www.up.ac.za/african-languages/arti-cle/17933/speakoutup</a>	Swati	unclear				Prof DJ Prinsloo	danie.prinsloo@up.ac.za	See above (Pretoria Afrikaans Corpus)
Pretoria Tshivenda Corpus	<a href="https://www.up.ac.za/african-languages/arti-cle/17933/speakoutup">https://www.up.ac.za/african-languages/arti-cle/17933/speakoutup</a>	Tshivenda	unclear				Prof DJ Prinsloo	danie.prinsloo@up.ac.za	See above (Pretoria Afrikaans Corpus)
Pretoria Xitsonga Corpus	<a href="https://www.up.ac.za/african-languages/arti-cle/17933/speakoutup">https://www.up.ac.za/african-languages/arti-cle/17933/speakoutup</a>	Xitsonga	unclear				Prof DJ Prinsloo	danie.prinsloo@up.ac.za	See above (Pretoria Afrikaans Corpus)
Pretoria Zulu Corpus	<a href="https://www.up.ac.za/african-languages/arti-cle/17933/speakoutup">https://www.up.ac.za/african-languages/arti-cle/17933/speakoutup</a>	Zulu	unclear				Prof DJ Prinsloo	danie.prinsloo@up.ac.za	See above (Pretoria Afrikaans Corpus)
Recalls Cilubà Corpus	Source: <a href="https://www.researchgate.net/figure/Corpora-of-African-languages-excluding-the-South-African-ones_tbl2_261835113">https://www.researchgate.net/figure/Corpora-of-African-languages-excluding-the-South-African-ones_tbl2_261835113</a>	Cilubà	unclear						Tools. Need to ask the authors for the corpora
Rosetta Disk 1.0	<a href="https://rosettaproject.org/disk/interactive/">https://rosettaproject.org/disk/interactive/</a>	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	<a href="https://repo.sadilar.org/handle/20.500.12185/7">https://repo.sadilar.org/handle/20.500.12185/7</a>		yes					However: a catalogue, contains what seems to be good resources
Sadilar Resource Index	<a href="https://repo.sadilar.org/handle/20.500.12185/9">https://repo.sadilar.org/handle/20.500.12185/9</a>		yes					However: a catalogue, contains what seems to be good resources
SAWA Corpus	<a href="http://www.aclweb.org/anthology/W09-0702">http://www.aclweb.org/anthology/W09-0702</a>	English - Swahili	no				Guy De Pauw CNTS - Language Technology Group, University of Antwerp, Belgium School of Computing and Informatics, University of Nairobi, Kenya guy.de-pauw@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 De-	includes New Testament (7.9 k sentences), Quran (6.2 k), Declaration of HR (0.2k), <a href="http://www.kamusi.org">Kamusi.org</a> (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

							partment, University of the Western Cape, South Africa gillesmau- rice.deschryver @ugent.be		
Scientific e-lexicography for Africa(2012-2015)	<a href="https://www.up.ac.za/african-languages/article/38000/research-projects">https://www.up.ac.za/african-languages/article/38000/research-projects</a>		no	medical					Access to the data not straightforward
SPC - Stockholm Parallel Corpora	<a href="https://athena.clarin.gr/re-sources/browse/spc-stockholm-parallel-corpora/ccb9d510a33111e5a465aa3fc9efd4927929e983ae0f48f7943ac55fd42f7d0a/">https://athena.clarin.gr/re-sources/browse/spc-stockholm-parallel-corpora/ccb9d510a33111e5a465aa3fc9efd4927929e983ae0f48f7943ac55fd42f7d0a/</a>	Afrikaans - English	yes						For Afrikaans
Swahili Computer corpora	<a href="http://www.ling.helsinki.fi/uhlcs/readme-all/RE-ADME-afro-as-nig-conlgs.html#C91">http://www.ling.helsinki.fi/uhlcs/readme-all/RE-ADME-afro-as-nig-conlgs.html#C91</a>	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	<a href="http://crubadan.org/">http://crubadan.org/</a>	Afrikaans (1307 crawled documents), Gbaya, Ndebele,	no						Needs to be parallelized

		Nothern Sotho, Sango, Sotughwest Gbaya, Tsonga, Venda, Xhosa, Zulu							
The Global-Phone Swahili corpus	<a href="http://catalog.elra.info/en-us/repository/browse/ELRA-S0375/">http://catalog.elra.info/en-us/repository/browse/ELRA-S0375/</a>	Swahili	yes		Commercial use		V. Mapelli	mapelli@elda.org	Not free of charge.
Treebanks in Universal Dependencies	<a href="https://universaldependencies.org">https://universaldependencies.org</a>	Afrikaans	yes		CC BY-SA 4.0		Peter Dirix, Liebeth Augustinus, Daniel van Niekerk	pe-ter.dirix@kuleu-ven.be, lies-beth.augustinus@kuleuven.be	Requires transformations
Treebanks in Universal Dependencies	<a href="https://universaldependencies.org">https://universaldependencies.org</a>	Amharic	yes	Bible, news non-fiction	CC BY-SA 4.0		Binyam Ephrem, Gashaw Arutie, Tsegay Woldemariam, Juan Ignacio Navarro Horñi-acek	binephrem@gmail.com	Requires transformations, further details here: <a href="https://github.com/UniversalDependencies/UD_Afrikaans-Afri-Booms/blob/master/README.txt">https://github.com/UniversalDependencies/UD_Afrikaans-Afri-Booms/blob/master/README.txt</a>
Treebanks in Universal Dependencies	<a href="https://universaldependencies.org">https://universaldependencies.org</a>	Bambara, Wolof, Yoruba	yes	news non-fiction	CC BY-SA 4.0		Katya Aplonova, Francis Tyers	zeman@ufal.mff.cuni.cz	Requires transformations, further details are provided here: <a href="https://github.com/UniversalDependencies/UD_Bambara-CRB/blob/master/README.md">https://github.com/UniversalDependencies/UD_Bambara-CRB/blob/master/README.md</a>
Treebanks in Universal Dependencies	<a href="https://universaldependencies.org">https://universaldependencies.org</a>	Wolof	yes	bible wiki	CC BY-SA 4.0		Bamba Dione	di-one.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	<a href="https://universaldependencies.org">https://universaldependencies.org</a>	Yoruba	yes	bible wiki	CC BY-SA 4.0		Adédayò Olúòkun, Daniel Zeman, Seyi Williams, Qlájídé Ishola	zeman@ufal.mff.cuni.cz	Requires transformations
Ubuntu	<a href="http://opus.nlpl.eu/Ubuntu-v14.10.php">http://opus.nlpl.eu/Ubuntu-v14.10.php</a>	English - Igbo, Afrikaans, Hausa	yes						
ukuxhumana	<a href="https://github.com/LauraMartinius/ukuxhumana">https://github.com/LauraMartinius/ukuxhumana</a>		yes						
Unsupervised compound splitter for Afrikaans	<a href="https://www.aclweb.org/anthology/N16-1078.pdf">https://www.aclweb.org/anthology/N16-1078.pdf</a>	Afrikaans	unclear				Patrick Ziering, Lonneke van der Plas	Patrick.Ziering@ims.uni-stuttgart.de, Lonneke.vanderPlas@um.edu.mt	
Voice of Nigeria Yorùbá news	<a href="http://von.gov.ng/yoruba">von.gov.ng/yoruba</a>	Yorùbá	yes						
Wikipedia	<a href="https://dumps.wikimedia.org/twwiki">dumps.wikimedia.org/twwiki</a>	Twi	yes	read texts from national newspapers					Text needs to be made available

XhosaNavy	<a href="http://opus.nlpl.eu/XhosaNavy.php">http://opus.nlpl.eu/XhosaNavy.php</a>	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
Yorùbà Bible	<a href="http://www.bible.com">www.bible.com</a>	Yorùbá	yes	prose text from fiction, news media and government documents domains					
Yorùbá Tweets	<a href="https://twitter.com/yobamoodua">twitter.com/yobamoodua</a>	Yorùbá	yes						



Yorùbá Wikipe- dia	<a href="https://dumps.wikimedia.org/yowiki/">dumps.wikimedia.org/yowiki/</a>	Yorùbá	yes						However: text needs to be made available
Zulu Wikimedia	<a href="https://ftp.acc.umu.se/mirror/wiki-media.org/dumps/zuwiki/">https://ftp.acc.umu.se/mirror/wiki-media.org/dumps/zuwiki/</a>	Zulu	yes	Legal non-fiction					Data needs to be extracted from the dump

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

## 5.4 Registration List

	Last Name	First Name	Affiliation
1	Abbott	Jade	Retro Rabbit
2	Adelani	David	Saarland University & Masakhane
3	Agbeko	Henry	Kwame Nkrumah University of Science and Technology
4	Agyapon-Ntra	Kwadwo	Self-attending
5	Ahia	Orevaoghene	Instadeep
6	Ajibade	Faith	University of Ibadan
7	Akinkunmi	Anuoluwapo	Student
8	Alabi	Jesujoba	DFKI
9	Ali	Jamiil Touré	Takwimu Lab
10	Amanfu	Richard	Institute of ICT Professionals Ghana
11	Anebi	Emmanuel	DataInsight
12	Anzagira	Allan re	North Carolina A & T State University
13	Anzaldo	Isa	UdS
14	Appati	Justice Kwame	University of Ghana
15	Aremu	Anuoluwapo	Aremu Language Consult
16	Asamoah	Eugene	Electricity Company Of Ghana
17	Ayami	Yasin	TsogoloTech
18	Bamutura	David	Mbarara University of Science and Technology
19	Barnes	Samuel	Student
20	Berejena	Beatrice	NA
21	Betty	Betty	University of Johannesburg
22	Blum	Seth	Meridian Institute
23	Boateng	Samuel	Ajuma software
24	Bridgman	Grant	Uliza
25	Brütting	Florian	GIZ
26	Buabeng	Edwin	Huawei
27	Budu	Joel	Artificial Intelligence Association of Ghana
28	Byamugisha	Joan	IBM Research Africa
29	Cayralat	Christian	LCT
30	Chirwa	Temweka	University of Cape Town
31	Clancy	Katie	IDRC
32	Coffie Debrah	Emmanuel	University of Cape Coast

33	David	Davis	TYD Innovation Incubator
34	Degila	Kevin	Masakhane
35	Diallo	Aboubacar	Independant Consultant
36	Dietrich	Klakow	Saarland University
37	Donner	Jonathan	Caribou digital
38	Dosseh	Desire	Dakar Institute of Technology
39	du Plessis	Liëtte	University of Johannesburg
40	du Toit	Jaco	UNESCO
41	Dube	Hlonipani	Bluemachines (Pty) Ltd
42	Dutta	Sourav	Saarland University
43	Dzidzinyo	Komla Tekpo Abah	INP-HB
44	Ekem	Ivy	UCC SMS
45	Elgizouli	Mukhtar	University of khartoum
46	Elmers	Mikey	University of Saarland
47	Eneremadu	Sydney	Chatbot Africa & Conversational AI Summit
48	España-Bonet	Cristina	DFKI
49	Estarrona	Ainara	HiTZ zentroa (UPV/EHU)
50	Gaelejew	Theodore	IBM Research
51	Gardent	Claire	CNRS
52	Gebremeskel	Gebre Kirstos	CWI
53	Gesicho	Rose Delilah	Student
54	Gimpel	Lea	GIZ
55	Gova	Webster	Umuzi
56	Graaf	Michael	Wikimedia-ZA
57	Griciūtė	Bernadeta	Saarland University
58	Gyamfi	Nana Kwame	Kumasi Technical University
59	Gyekye	Kwame	Experian
60	Ibrahim	Omnia	University of Zürich
61	Ihle	Frank	private
62	Iminza	Diana	JKUAT
63	Israel	Odeajo	University of Ibadan, Nigeria
64	Jarso	Guyo	University of Rwanda
65	Jobe	Wuyeh	Carnegie Mellon University Africa
66	Kalmykova	Anastasiia	Uds
67	Kashupi	Tokolo N	Namibia University of Science and Technology

68	Katumba	Andrew	Makerere University
69	Kaye	Jofish	Mozilla
70	Kiden	Sarah	Northumbria University
71	Kintu Mwanje	Timothy	Mbarara University of Science and Technology
72	Kioko	Moses	
73	Konobelkina	Anna	University of Saarland
74	Koupoh	Esaïe Alain-Emmanuel	Dakar Institute of Technology
75	Ladipo	Taiwo	SAP
76	Lang	Inga	EM Double Master in Language and Communication Technologies (LCT), University of Groningen and University of Malta
77	Lösch	Andrea	DFKI
78	Luberenga	John	Flying Gravity Technologies
79	Lubrini	Elisa	Université de Lorraine
80	Mabaso	Mbangiso	Sisanda Tech
81	Mahlezana	Thiba	TGM MakerSpace
82	Makumbirofa	Hamony	Tshimologong
83	Malatji	Masike	University of Johannesburg
84	Manishimwe	Alban	The Nelson Mandela African Institution of Science and Technology
85	Mansour	Ayman	Sudan university of science and technology
86	Manzi	Innocent	250STARTUPS
87	Matana	Maduhu	Platinum Credit
88	Mbadi	Sheila	CMU-Africa
89	Mbaye	Derguene	Baamtu
90	Mbogho	Audrey	USIU-Africa
91	Merci Hategekimana	Arsène	College student
92	Mintz	Blanca	GIZ
93	Mokele	David	CS SMART
94	Moore	Stephen	Ghana NLP
95	Moruye	Lawrence	Multimedia University of Kenya
96	Moyo	Mlamuleli	Tmg makers space
97	Moyo	Simon	Mafikeng digital innovation hub
98	Mthembu	Nkululeko	Private
99	Mugambi	Jonan	Volkswagen Mobility Solutions Rwanda
100	Muhire	Remy	Mozilla

101	Muite	Benson	-
102	Mukiibi	Jonathan	AI Lab Makerere University
103	Munro	Robert	Machine Learning Consulting
104	Musoya	Gael	Digitech group
105	Mwatukange	Joseph	Konga Technologies
106	N/A	Ari	Prækelt
107	Nakatumba	Joyce	Makerere University
108	Navisa	Salma	UIN Sunan Ampel
109	NDIAYE	Malick	Bambey University/ Senegal
110	Ndodana	Nwabisa	Student
111	Nixon	Samantha	IDRC
112	Niyongabo	Rubungo Andre	Masakhane
113	Nordor	Eli	BlueCrest
114	Nxumalo	Thabiso	North Park Telecoms
115	Oduor	Clinton	clintonoduor3@outlook.com
116	Ogayo	Perez	African Leadership University
117	Oguche	Victor	Daptem Engineering
118	Ogundepo	Odunayo	Deloitte & Touche Nigeria Limited
119	Ojesina	Akolade	University of Ibadan
120	Oládípò	Akíntúndé	Wragby Business Solutions and Technologies Limited
121	Olbrich	Phillipp	GIZ
122	Oloke	Adewale	Unilag
123	Oluyori	Elkanah	Reformers of Africa
124	Orlic	Davor	Knowledge 4 All Foundation
125	Owusu-Darko	Ama	Ensign College of Public Health
126	Pakzad	Roya	Taraaz
127	Patrick	Emil	Nelson Mandela African Institution of Science and Technology
128	Phahlamohlaka	Lazarus	North Park Telecoms
129	Pienaar	Marne	UJ
130	Pratt Miles	Jennifer	Lacuna Fund: Our Voice on Data
131	Prenom	Daouda	Université Alioune Diop de Bambey
132	Quayson	Ebenezer	UENR
133	Ramos	Alex	N/A
134	Resch	Christian	GIZ
135	Rutunda	Samuel	Digital Umuganda

136	Sam	Abraham	FAIR Forward, GIZ
137	Samb	Sokhar	AIMS
138	Schnur	Eileen	DFKI
139	Schonwetter	Tobias	University of Cape Town
140	Schulte	Kim	GIZ
141	Seibold	Balthas	GIZ
142	Sibal	Prateek	UNESCO
143	Siminyu	Kathleen	Artificial Intelligence for Development - Africa
144	Smith	Blake	University of Edinburgh
145	Smith	Matthew	International Development Research Centre
146	Sorinolu	Babafemi	Dominican University
147	Tembo	David	VisionAi
148	Tsouapi	Rigobert	aims
149	Turikumwe	Jean Dela paix	A. I. S
150	Tyira	Loyiso	ICT SMME Chamber
151	Umuhire	Evelyne	WeCode Moringa School
152	Urosevic	Jovana	Independent
153	van Genabith	Josef	DFKI
154	Wanzare	Lilian	Maseno University
155	Waterfield	Rebecca	N/A
156	Watson	Sarah	Mozilla
157	Wilson	Daniel	XRI
158	Wilson	Joe	J-MO Global
159	Yeboah-Boateng	Dr. Ezer Osei	Ghana Technology University College

Figure 9: List of Registrations

## 6 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). *Massively 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 and Twi*. 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.