DATA FLARING: RADIO HOUSES AS REMEDY MEASURE

BY
CLEMENT ODOJE
lekeclement2@gmail.com
DEPARTMENT OF LINGUISTICS AND AFRICAN LANGUAGES
UNIVERSITY OF IBADAN

ABSTRACT

- The dichotomy between rich and low resource languages is enshrined in the availability of data.
- The definition of low resource languages is determined not by the speakers of the languages but in the quantum of data archived.
- Is it true that there is no data for African languages?
- There are large proportion of untapped data in Africa due to many reasons.

Introduction

- Most of African languages (if not all) are categorized as low-resource languages
- They are being defined or terms as this because it can be understood that they are less studied, resource scarce, less computerized, less privileged, less commonly taught, or low density, among other denominations.

Introduction

- Chan and Rosenfeld (2012) report that resource scarce languages are languages that have small users advantage in relation to technology which could be ignored by the commercial world.
- It is emphasised that of all the languages in the world, only languages spoken by economically developed nations are available in spoken dialog system (SDS), Google voice search and speech technology. This is the actual digital divide.
- Magueresse, Carles and Heetderks (2020) explain that most of today's NLP research focuses on 20 of the 7000 languages of the world, leaving the vast majority of languages understudied.
- Does it really mean that African languages do not have language resources? Or does it mean that the resources are being "flared"? If African languages are indeed resource scarce, what is the way forward? These are the question this paper intends to interrogate.

African Languages and NLP Dataset

- In addition to being a means of communication, language has a socio-economic role similar to that of money in industrial society,. While money is used to acquire material goods, language is used to acquire knowledge and intangible goods (Osborn 2010).
- information revolution worldwide is increasingly multilingual, and as the presence of the new ICTs in Africa extends to larger areas beyond the capital cities, there is a growing need to accommodate the use of diverse African languages and greater potential to tap the linguistic wealth of the continent for development and education.

African Languages and NLP Dataset

- The fact is that European languages cannot meet all of Africa’s needs, and African languages have much to contribute.
- If African languages will contribute to ICTs Natural Language Processing will play a crucial role.
- For NLP state of the arts to yield appropriate result, there is need for large quantum of data which to this end compared to European languages are not sufficiently available hence the categorization of African languages as low-resource/resource scarce languages.

African Languages and NLP Dataset

- Yoruba language will be the focus of this paper, using it as yardstick for other African languages. Evidence abound that there are lots of Yoruba text materials. These are:

- ➢Education materials
- ➢Literary text
- ➢Religions text
- ➢Newspaper materials
- ➢Audio text materials

Education Materials

- Government policy mandates that Yoruba is taught as a school subject. In fact, before now, it was a compulsory subject. In addition, it is also a discipline. As a school subject, it is taught in primary, secondary and tertiary levels. As a discipline, it is taught at Colleges of Education, and University, up to Ph.D levels.
- Hence, there are textbooks, workbooks, journal publications, periodicals, metalanguage, students’ final year long essays, projects and thesis written in Yoruba. There are standard examination question papers for years. Prominent among them is West African Examination Council questions.

Education Materials

- Yoruba is taught in 42 Colleges of Education,
- 12 Universities as undergraduate programmes and
- 6 Universities for postgraduate programmes up to Ph.D. level.
- Aside for journal publication and periodicals, there might be no need for copyright permission for these huge materials that are available unused

Literary Text

- Since 1948 when indigenous literary text began, till date no one can ascertain the number of such publications available in the market because they are very many. Odoje (2017) reports that he was able to generate 14, 216 parallel corpus from four literary texts. With this, we would be able to generate as much sentences as possible if and only if copyright permission could be granted for all literary text. Apart from this, many of them have to be converted to modern Yoruba orthography.

Religious Text

- The impact of missionaries on making religious literature available cannot be over emphasized. Aside from Bible and Koran that has been translated to many Nigerian languages. Many Christian denominations have their Sunday School Manuals and other literature translated to Nigerian languages as well. Tracts and reports in the indigenous languages could be used too. It is a known fact that Jehovah Witness materials have been of great help in this regard.

Newspaper Material

- Yoruba enjoys early news publications. Rev. Henry Townsend published the first Yoruba newspaper on 23rd November, 1859. Till date, there is a weekly publication of Alaroye which could be of immense contribution to the quest of developing a corporal of the language.

Audio Text

- This is another area where Yoruba as a language flares her data. We call this media text. People take pleasure is listening to radio and especially, programmes in their languages. This has brought about the establishment of many radio stations.

Demography of Sample Population

- There are 27 radio stations in Ibadan alone
- 38 radio stations in Oyo state
- 149 radio stations in South-West of Nigeria
- We visited 20 radio stations in Oyo state: two radio stations in Ogbomoso, one in Oyo, one in Igboho, one at Alaga (Oke Ogun), 15 in Ibadan
- 10% of the population are school based radio stations
- 50% are privately owned
- While 40% are owned by government

Data Gathered

- 78% of the radio station have at least 10 hours of Yoruba programmes daily.
- By law, all the radio stations are expected to archive their programmes especially for the purpose of retrieval in case of any accusation of violation of rights or litigation.
- Because many of the privately owned radio stations are profit oriented and partly political trying to avoid government pressure on their presenters and stations as well refused to share their data with us. Government owned stations project government project and policies but for no known reason, some station refused as well. Only only 5% of the population are willing to share part of their archived materials with this research groups.
- More than 40 hours of audio recording was donated

Available Repository

- https://dataverse.schlarsportal.info/dataset.xhtml?persistentId=doi:10.5683/SP2/VTGWQF
- https://universaldependencies.org/treebanks/yo_ytb/index.html

Challenges

- Annotation
- Storage system
- Accessibility

Advantages of Radio Data

- Studio recording, very clear

- Public information, no need for copyright permission

- It is general domain

- Thank you
- Midawasi
- Danke sun