

Language data for African languages

Andrea Lösch
andrea.loesch@dfki.de
www.dfki.de

AI4D AFRICA WEBINAR SERIES

MAKING NLP WORK IN AFRICA
WITH AN INTRODUCTION TO THE GIZ AI4D
AFRICAN LANGUAGE DATASET CHALLENGE

3 July 2020 from 14:00 to 16:00 pm CAT/CEST/UTC+2



german
cooperation

giz

FAIR FORWARD
Artificial Intelligence for all.

DFK

IDRC | CRDI
International Development Research Centre
Centre de recherches pour le développement international

moz://a

*“Those who know nothing of
foreign languages know nothing
of their own.”*

Johann Wolfgang von Goethe (1749 – 1832)



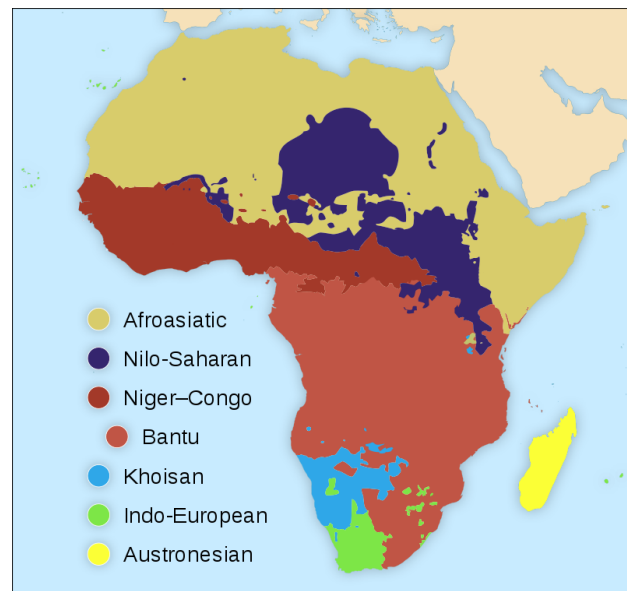
Introduction



- Language data as fundamental pre-requisite for MT and other LT
- DFKI is currently involved in the European Language Resource Coordination ([ELRC](#)) → Goal: make public services across Europe multilingual
- Support for 24 EU-official languages plus Norwegian and Icelandic
- A snapshot of resources across various domains: [ELRC-SHARE](#)
- Strong focus on under-resourced languages (Croatian, Maltese, Irish, Icelandic)

Languages in Africa

- 1.500 – 2.000 languages in Africa...
- Different language families...
- Arabic, Somali, Berber, Amharic, Oromo, Igbo, Swahili, Hausa, Manding, Fulani and Yoruba are spoken by tens of millions of people...



Source: https://en.wikipedia.org/wiki/Languages_of_Africa

A snapshot of data for African languages

- A first, non-exhaustive overview of language data and/or sources for African languages is available [here](#)
- Languages found:
 - Hausa
 - Igbo
 - Luganda
 - Luo
 - Northern Sotho
 - Setswana
 - Swahili
 - Twi
 - Xhosa
 - Xitsonga
 - Yorùbá
 - Zulu
- Central question: How and where to get data from?

Data collection approaches

- Important aspects of language data collection:
 - Identification of and collaboration with relevant language data holders
 - Identification and use of sources of language data
 - Making language data reusable!

Language data holders

- Identification of and collaboration with relevant language data holders
- Language data holders include any organisations and/or people that may create language data
- Examples of language data holders:
 - 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 and/or governmental agencies (e.g. Language Unit of the Department of Cultural Affairs and Sport ([DCAS](#)) of Western Cape Government, South African Centre for Digital Language Resources ([SADiLaR](#)))

Language data holders

- Examples of language data holders (cont.):
 - 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](#))

Working with language data holders

- 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.
- It's a community-building effort!

Sources of language data

- Sources of language data can be any bi- or multilingual websites in the languages sought
- Examples:
 - governmental websites in the target countries/languages
 - websites of public services and academic institutions in the target countries/languages
 - websites of international, national or local organisations in the target countries/languages
- Web crawling to identify and retrieve mono-, bi- or multilingual language data from the Internet and to turn them into MT-ready language resources
➔ language profile

Making language data (re-)usable

- Two aspects:
 - Technical usability of language data
 - Legal usability of language data



Making language data (re-)usable

- Quick-check: Ensuring technical usability of language data
 - Is the format readable?
 - Is the source / are the sources copyrighted? (also see below, legal usability)
 - Have the source and target language(s) be identified correctly?
 - Is the alignment ok?
 - Are there any tokenization errors (no separator between words)?
 - Is the content machine-translated?

Making language data (re-)usable

- Quick-check: Ensuring legal usability of language data
 - Is the data protected by copyright?
 - If the data is protected by copyright can I identify the owner of the copyright or the author of the work?
 - Is the data available under a public license?
 - 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...

Data collection in and for Africa...

... voices and perspectives
from our experts:



Orevaoghene Ahia
Instadeep



Stephen E. Moore
Ghana NLP



Tobias Schonwetter
University of Cape Town