Some Aspects of Linguistic Diversity in Europe and Africa

Cristina España-Bonet
DFKI GmbH

SPARC International Symposium on Mahatma Gandhi and Linguistic Diversity
23rd September 2020
Outline

1. Language Diversity
2. Europe & the EU Council Presidency Translator
3. Low Resource African Languages
There are more than **7000 languages**
(even if the definition of language is not straightforward!)

**141 language families**
(6 of them account for 2/3 of all languages and 5/6 of the world’s population)
Language Diversity

Some Numbers

- There are more than 7000 languages
  (even if the definition of language is not straightforward!)

- 141 language families
  (6 of them account for 2/3 of all languages and 5/6 of the world’s population)

Explore:

Ethnologue  https://www.ethnologue.com/
Glottolog  http://glottolog.org/
Linguistic Maps  http://linguisticmaps.tumblr.com/
Language Diversity

Two Projects, Two Realities

Africa & Europe

language diversity

population density

endangered languages

digital richness
Language Diversity

Endangered Languages as Effect (?) of Diversity

http://www.endangeredlanguages.com
The situation is very different in different regions of the world

Three “hot spots”
- Central/South America, North Sub-Saharan Africa, South/Southeast Asia

No direct correlation with population density
Language Diversity

Two very Different Realities

Europe

- High population density
- Low language diversity
- Digital rich region

(North) Sub-Saharan Africa

- High population density
- High language diversity
- Digital poor region
Outline

1. Language Diversity
2. Europe & the EU Council Presidency Translator
3. Low Resource African Languages
~130 languages

24 official languages

More than 80 regional or minority languages
  - the number of speakers is estimated at 40 million citizens

80% of the European minority languages are endangered
  - Ladin, Rhaeto Romansh, Upper and Lower Sorbian, North Frisian, Kashubian...
Background:

- EU has a rotatory (6 months length) presidency system

- Germany assumed the presidency of the Council of the European Union on July 1st
  ... "chairs" the meetings of the EU Council
  ... is responsible for progressing EU legislation

- Official translation engines provided for the 24 official languages
EU-PT Goal:

- Improve communication in 24 official EU languages
- German as a vehicle language during the presidency
- Make (German) artificial intelligence and language technologies visible
Partners & Specific Translation Engines:

- DF
- TILDE

More (generic) engines:

- DeepL
- EU flag
- Connecting Europe eTranslation
DFKI Engines:

- Bidirectional neural systems $\{XX, \text{de}\} \leftrightarrow \{\text{de}, XX\}$
- Transformer big architecture
- Domain-adapted systems by transfer learning
  - Data selection, generation and crawling
  - Diverse domains such as covid-19, politics or energy
Europe & the EU Council Presidency Translator

The Presidency Translator, 24 Languages
Europe & the EU Council Presidency Translator

The Presidency Translator, 24 Languages

EU Council Presidency translator

From:
- French
- English
- Bulgarian

To:
- German
- English
- Spanish

Provider:
- DFKI

Enter the text or homepage address you would like to translate.

Machine translation results help to understand the meaning of a source text, but do not equal translation by a human.

Machine Translation provided by DFKI

https://presidencymt.eu
### EU Council Presidency Translator public website translation directions usage (in words, top 30, 01.08.-31.08.2020.)

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Translation Directions</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-Croatian</td>
<td>2,129,355</td>
</tr>
<tr>
<td>English-German</td>
<td>745,838</td>
</tr>
<tr>
<td>English-French</td>
<td>743,300</td>
</tr>
<tr>
<td>German-English</td>
<td>75,256</td>
</tr>
<tr>
<td>German-French</td>
<td>229,560</td>
</tr>
<tr>
<td>German-Polish</td>
<td>87,005</td>
</tr>
<tr>
<td>English-Danish</td>
<td>467,503</td>
</tr>
<tr>
<td>English-Spanish</td>
<td>360,047</td>
</tr>
<tr>
<td>German-Croatian</td>
<td>257,724</td>
</tr>
<tr>
<td>German-Portuguese</td>
<td>198,481</td>
</tr>
<tr>
<td>German-Bulgarian</td>
<td>581,623</td>
</tr>
<tr>
<td>French-Spanish</td>
<td>670,575</td>
</tr>
<tr>
<td>French-Italian</td>
<td>124,845</td>
</tr>
<tr>
<td>French-Greek</td>
<td>109,074</td>
</tr>
<tr>
<td>English-Latin</td>
<td>94,117</td>
</tr>
<tr>
<td>Spanish-English</td>
<td>197,955</td>
</tr>
<tr>
<td>English-Italian</td>
<td>128,700</td>
</tr>
<tr>
<td>English-French</td>
<td>140,855</td>
</tr>
<tr>
<td>Polish-English</td>
<td>104,784</td>
</tr>
<tr>
<td>English-Danish</td>
<td>223,398</td>
</tr>
<tr>
<td>English-German</td>
<td>223,316</td>
</tr>
<tr>
<td>English-Polish</td>
<td>173,420</td>
</tr>
<tr>
<td>English-Hungarian</td>
<td>163,180</td>
</tr>
<tr>
<td>Spanish-English</td>
<td>160,001</td>
</tr>
</tbody>
</table>
Europe & the EU Council Presidency Translator

The Presidency Translator, 24 Languages

EU Council Presidency Translator usage statistics per target language via eu2020.de (in words, 01.08.-31.08.2020.)

- Polish: 507,200
- Italian: 837,705
- Spanish: 944,821
- Bulgarian: 650,405
- Croatian: 415,498
- Latvian: 256,698
- Greek: 698,764
- Portuguese: 832,833
- Swedish: 180,873
- Dutch: 209,909
- Hungarian: 245,052
- Danish: 191,475
- Romanian: 257,889
- Finnish: 221,883
- Czech: 212,017
- Estonian: 188,489
- Slovenian: 80,581
- Irish: 26,031
- Lithuanian: 60,051
- Maltese: 16,395
- Slovakian: 22,400
Online CAT tools
Online CAT tools
Popular usage channels
(in words, 01.08.-31.08.2020.)

- Document translation: 3,189,211
- Text translation: 3,323,070
- SDL Trados plug-in: 148,912
- Web page translation: 6,882,219

Legend:
- eu2020.de
- EUCPT
Let’s move to a very different project & reality

Thanks to Jesujoba O. Alabi, Kwabena Amponsah-Kaakyire and David Adelani for sharing
Estimations talk about the existence of 1500-2000 languages.

Only 28 official languages (11 in South Africa!)

But Nigeria alone has over 500 languages.

300 languages have less than 10,000 speakers, which puts them on the UNs endangered list, and 37 are in danger of completely dying out.
Could one build an “African Presidency Translator”?

Countries as a whole have 28 official languages (~EU)

Arabic, Somali, Berber, Amharic, Oromo, Igbo, Swahili, Hausa, Manding, Fulani and Yorùbá are spoken by tens of millions of people¹

But some of them are in digital danger

- Not enough digital content for AI technologies
- No resources or low-quality resources

¹ https://afric.online/2835-languages-present-a-rich-cultural-diversity-in-africa/
In the **deep learning world** one relies on

- Word embeddings (word2vec-like)
- Contextual representations (BERT-like)
- Transfer learning to new tasks (or languages??)
Low-resource languages (lack of data or noisy data)

- Quality of pre-trained word embeddings?
- Quality of multilingual contextual representations?
- Use case: Twi and Yorùbá
## Low Resource African Languages

### Data (Text)

<table>
<thead>
<tr>
<th>Description</th>
<th>Source URL</th>
<th>#tokens</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Yorùbá</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lagos-NWU corpus</td>
<td>github.com/Niger-Volta-LTI</td>
<td>24,868</td>
<td>clean</td>
</tr>
<tr>
<td>Alákọwé</td>
<td>alakoweyoruba.wordpress.com</td>
<td>24,092</td>
<td>clean</td>
</tr>
<tr>
<td>Òrọ̀ Yorùbá</td>
<td>oroyoruba.blogspot.com</td>
<td>16,232</td>
<td>clean</td>
</tr>
<tr>
<td>Òdè Yorùbá Rëwà</td>
<td>deskgram.cc/edeyorubarewa</td>
<td>4,464</td>
<td>clean</td>
</tr>
<tr>
<td>Doctrine $ Covenants</td>
<td>github.com/Niger-Volta-LTI</td>
<td>20,447</td>
<td>clean</td>
</tr>
<tr>
<td>Yorùbá Bible</td>
<td><a href="http://www.bible.com">www.bible.com</a></td>
<td>819,101</td>
<td>clean</td>
</tr>
<tr>
<td>GlobalVoices</td>
<td>yo.globalvoices.org</td>
<td>24,617</td>
<td>clean</td>
</tr>
<tr>
<td>Jehovah Witness</td>
<td><a href="http://www.jw.org/yo">www.jw.org/yo</a></td>
<td>170,203</td>
<td>clean</td>
</tr>
<tr>
<td>Ìríinkèrìndò nínú ębó èlègbèje</td>
<td>manual</td>
<td>56,434</td>
<td>clean</td>
</tr>
<tr>
<td>Igbo Olódùmarè</td>
<td>manual</td>
<td>62,125</td>
<td>clean</td>
</tr>
<tr>
<td>JW300 Yorùbá corpus</td>
<td>opus.nlpl.eu/JW300.php</td>
<td>10,558,055</td>
<td>clean</td>
</tr>
<tr>
<td>Yorùbá Tweets</td>
<td>twitter.com/yobamoodua</td>
<td>153,716</td>
<td>clean</td>
</tr>
<tr>
<td>BBC Yorùbá</td>
<td>bbc.com/yoruba</td>
<td>330,490</td>
<td>noisy</td>
</tr>
<tr>
<td>Voice of Nigeria Yorùbá news</td>
<td>von.gov.ng/yoruba</td>
<td>380,252</td>
<td>noisy</td>
</tr>
<tr>
<td>Yorùbá Wikipedia</td>
<td>dumps.wikimedia.org/yowiki</td>
<td>129,075</td>
<td>noisy</td>
</tr>
<tr>
<td><strong>Twi</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bible</td>
<td><a href="http://www.bible.com">www.bible.com</a></td>
<td>661,229</td>
<td>clean</td>
</tr>
<tr>
<td>Jehovah’s Witness</td>
<td><a href="http://www.jw.org/tw">www.jw.org/tw</a></td>
<td>1,847,875</td>
<td>noisy</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>dumps.wikimedia.org/twwiki</td>
<td>5,820</td>
<td>noisy</td>
</tr>
<tr>
<td>JW300 Twi corpus</td>
<td>opus.nlpl.eu/JW300.php</td>
<td>13,630,514</td>
<td>noisy</td>
</tr>
</tbody>
</table>
Low Resource African Languages

Quality of Embeddings

- Experiments with **Twi** (noisy + scarce) & **Yorùbá** (noisy)

<table>
<thead>
<tr>
<th>Model</th>
<th>Vocab</th>
<th>Twi Spearman $\rho$</th>
<th>Vocab</th>
<th>Yorùbá Spearman $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1: Pre-trained Model (Wiki)</td>
<td>935</td>
<td>0.143</td>
<td>21,730</td>
<td>0.136</td>
</tr>
<tr>
<td>F2: Pre-trained Model (Common Crawl &amp; Wiki)</td>
<td>NA</td>
<td>NA</td>
<td>151,125</td>
<td>0.073</td>
</tr>
</tbody>
</table>
### Quality of Embeddings

- **Experiments with Twi (noisy+scarce) & Yorùbá (noisy)**

<table>
<thead>
<tr>
<th>Model</th>
<th>Vocab</th>
<th>Twi Spearman $\rho$</th>
<th>Vocab</th>
<th>Yorùbá Spearman $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1: Pre-trained Model (Wiki)</td>
<td>935</td>
<td>0.143</td>
<td>21,730</td>
<td>0.136</td>
</tr>
<tr>
<td>F2: Pre-trained Model (Common Crawl &amp; Wiki)</td>
<td>NA</td>
<td>NA</td>
<td>151,125</td>
<td>0.073</td>
</tr>
<tr>
<td>C1: Curated Small Dataset (Clean text)</td>
<td>9,923</td>
<td>0.354</td>
<td>12,268</td>
<td>0.322</td>
</tr>
<tr>
<td>C2: Curated Small Dataset (Clean + some noisy text)</td>
<td>18,494</td>
<td>0.388</td>
<td>17,492</td>
<td>0.302</td>
</tr>
<tr>
<td>C3: Curated Large Dataset (All Clean + Noisy texts)</td>
<td>47,134</td>
<td>0.386</td>
<td>44,560</td>
<td>0.391</td>
</tr>
</tbody>
</table>
Experiments with **Twi** (noisy+scarce) & **Yorùbá** (noisy)

<table>
<thead>
<tr>
<th>Model</th>
<th>Vocab</th>
<th>Twi Spearman $\rho$</th>
<th>Yorùbá Spearman $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1: Pre-trained Model (Wiki)</td>
<td>935</td>
<td>0.143</td>
<td>21,730</td>
</tr>
<tr>
<td>F2: Pre-trained Model (Common Crawl &amp; Wiki)</td>
<td>NA</td>
<td>NA</td>
<td>151,125</td>
</tr>
<tr>
<td>C1: Curated <em>Small</em> Dataset (Clean text)</td>
<td>9,923</td>
<td>0.354</td>
<td>12,268</td>
</tr>
<tr>
<td>C2: Curated <em>Small</em> Dataset (Clean + some noisy text)</td>
<td>18,494</td>
<td><strong>0.388</strong></td>
<td>17,492</td>
</tr>
<tr>
<td>C3: Curated <em>Large</em> Dataset (All Clean + Noisy texts)</td>
<td>47,134</td>
<td>0.386</td>
<td>44,560</td>
</tr>
<tr>
<td>D2: Curated <em>Small</em> Dataset (Clean + some noisy text)</td>
<td>22,851</td>
<td><strong>0.437</strong></td>
<td>56,086</td>
</tr>
</tbody>
</table>
**Contextual embeddings**

- NER F1 score on Global Voices Yorùbá corpus

<table>
<thead>
<tr>
<th>Embedding Type</th>
<th>Date</th>
<th>Loc</th>
<th>Org</th>
<th>Per</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-trained <em>uncased</em> Multilingual-bert (Multilingual vocab)</td>
<td>44.6</td>
<td>33.9</td>
<td>12.1</td>
<td>5.7</td>
<td>27.1 ± 0.7</td>
</tr>
<tr>
<td>Fine-tuned <em>uncased</em> Multilingual-bert (Multilingual vocab)</td>
<td>64.0</td>
<td>65.3</td>
<td>38.8</td>
<td>47.4</td>
<td>56.4 ± 2.4</td>
</tr>
<tr>
<td>Fine-tuned <em>uncased</em> Multilingual-bert (Yorùbá vocab)</td>
<td>67.0</td>
<td>71.5</td>
<td>40.4</td>
<td>49.4</td>
<td>60.1 ± 0.8</td>
</tr>
</tbody>
</table>
Language identification for African languages

- 40 African languages (+10 high resourced languages)
- Language models trained on the Bible

Parallel sentence gathering

- Manual collection/translation to create public resources

Collaboration with GIZ and their FAIR Forward Project – Artificial Intelligence for All

- Dissemination
Thanks! And...

wait!

Questions?
Some Aspects of Linguistic Diversity in Europe and Africa

Cristina España-Bonet
DFKI GmbH

SPARC International Symposium on Mahatma Gandhi and Linguistic Diversity
23rd September 2020