

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

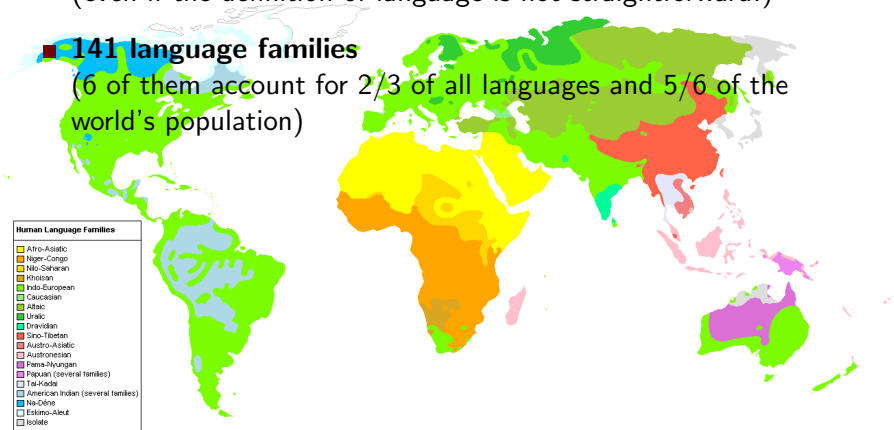
- 1 Language Diversity
- 2 Europe & the EU Council Presidency Translator
- 3 Low Resource African Languages

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)



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

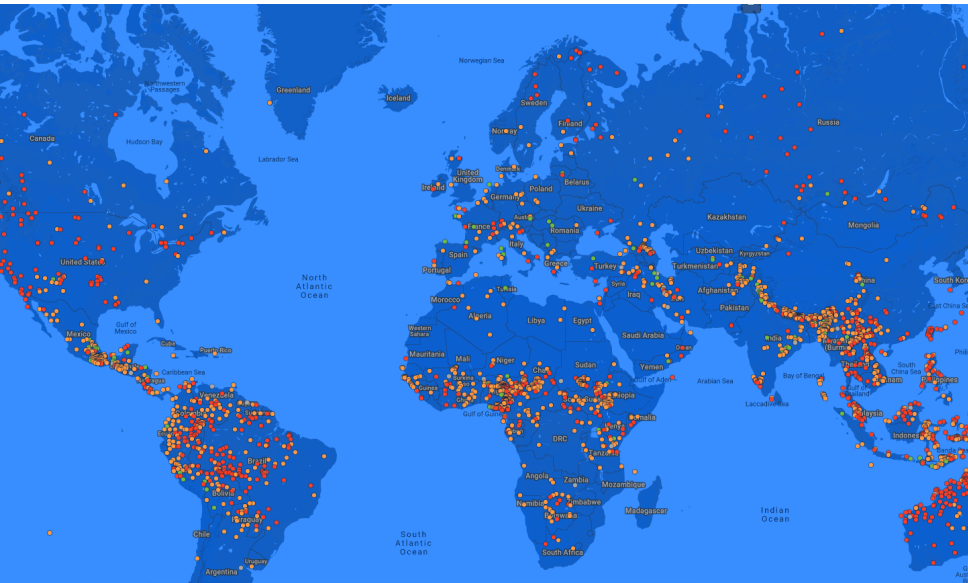
population density

endangered languages

digital richness

Language Diversity

Endangered Languages as Effect (?) of Diversity



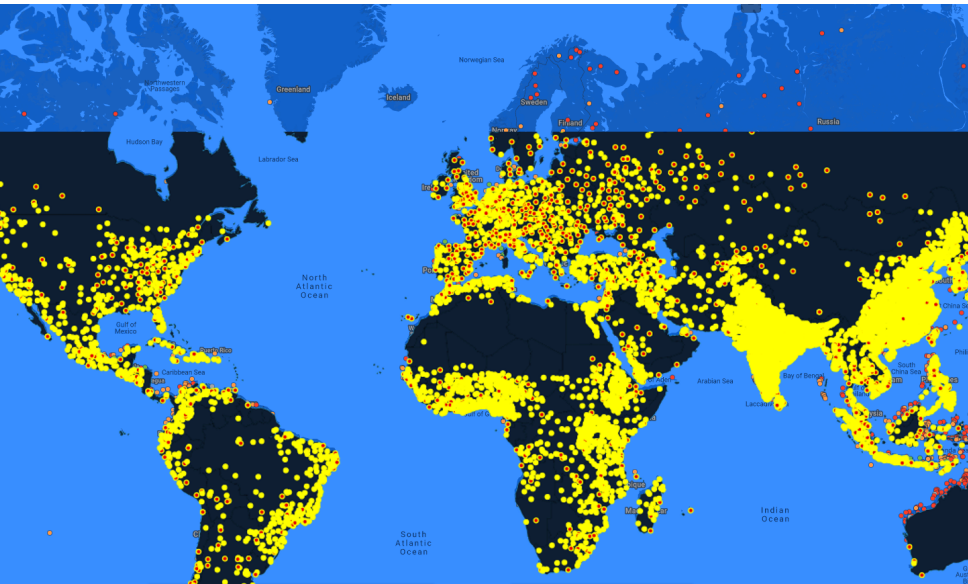
Language Diversity

Endangered Languages as Effect (?) of Diversity

- 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

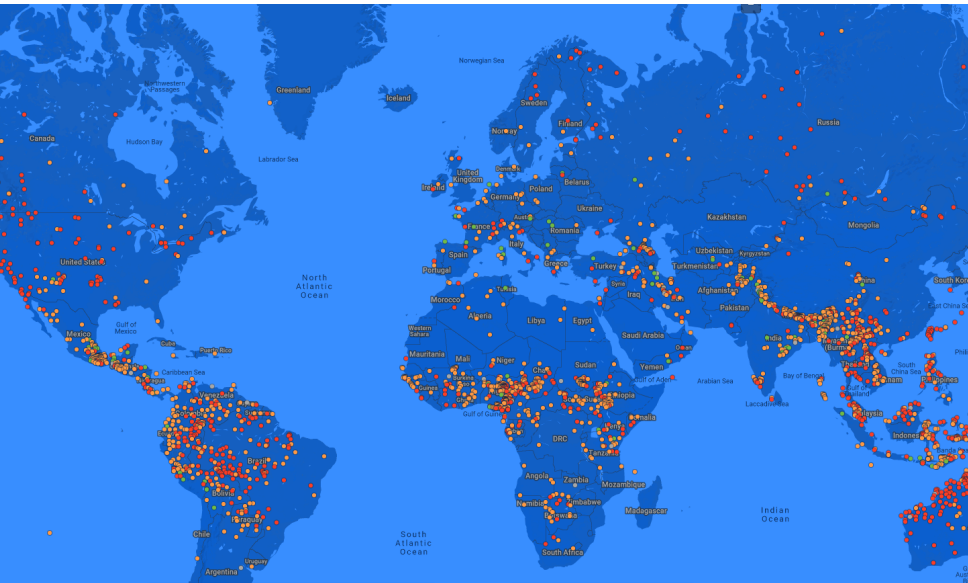
Language Diversity

Endangered Languages as Effect (?) of Diversity



Language Diversity

Endangered Languages as Effect (?) of Diversity



Language Diversity

Two very Different Realities

Europe

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

(North) Sub-Saharan Africa

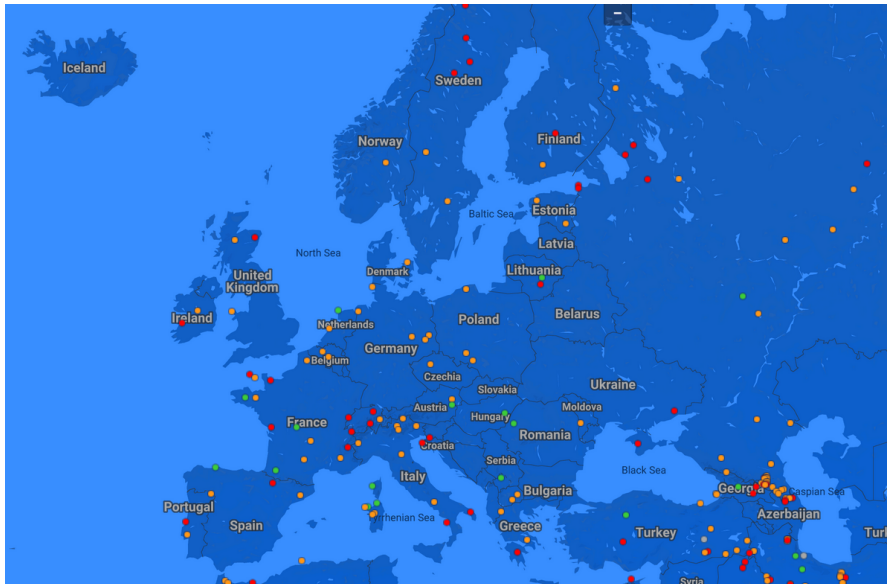
- High population density
- High language diversity
- Digital poor region

Outline

- 1 Language Diversity
- 2 Europe & the EU Council Presidency Translator
- 3 Low Resource African Languages

Europe & the EU Council Presidency Translator

Language Diversity in Europe



- ~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

Europe & the EU Council Presidency Translator

EU Presidency Translator

Partners & Specific Translation Engines:



More (generic) engines:



DFKI Engines:

- Bidirectional neural systems $\{XX, de\} \Leftrightarrow \{de, XX\}$
- Transformer big architecture
- Domain-adapted systems by transfer learning
 - Data selection, generation and crawling
 - Diverse domains such are covid-19, politics or energy

Europe & the EU Council Presidency Translator

The Presidency Translator, 24 Languages

EU Council Presidency Translator

Doc
translation



Text
translation



Web
translation



Cat
plug-in



On-line CAT
tool



Terminology
tool



Web
translation
plug-in





Europe & the EU Council Presidency Translator

The Presidency Translator, 24 Languages



EU Council Presidency translator

Log in  

From:

French English Bulgarian ▾




To:

German English Spanish ▾ DFKI ▾

Provider:

Translate

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

 Upload or drag a file

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

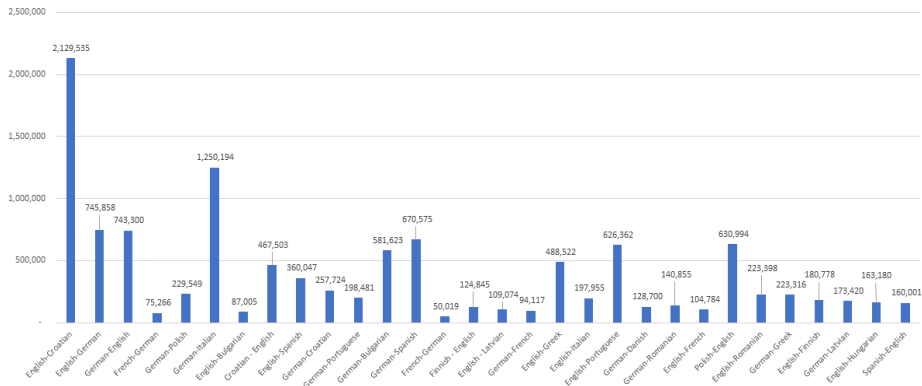


<https://presidencymt.eu>

Europe & the EU Council Presidency Translator

The Presidency Translator, 24 Languages

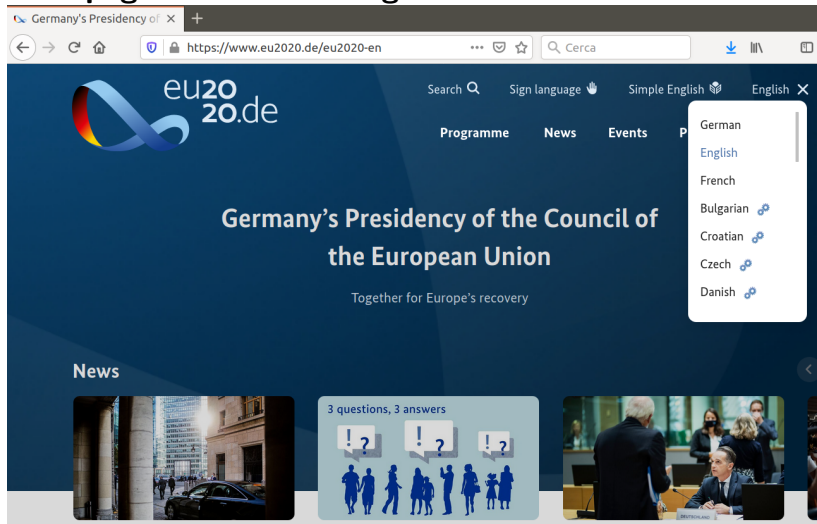
EU Council Presidency Translator public website
translation directions usage (in words, top 30, 01.08.-31.08.2020.)



Europe & the EU Council Presidency Translator

The Presidency Translator, 24 Languages

Web page translation widget



The screenshot shows a web browser window with the address bar displaying `https://www.eu2020.de/eu2020-en`. The website header features the logo `eu2020.de` and navigation links for `Programme`, `News`, and `Events`. A search bar and a 'Sign language' icon are also present. The main content area displays the title `Germany's Presidency of the Council of the European Union` and the tagline `Together for Europe's recovery`. A 'News' section is visible at the bottom, with three article thumbnails. A language translation widget is open on the right side of the page, showing a list of languages: German, English, French, Bulgarian, Croatian, Czech, and Danish. Each language is accompanied by a small icon representing the language's flag.

Germany's Presidency of the Council of the European Union

Together for Europe's recovery

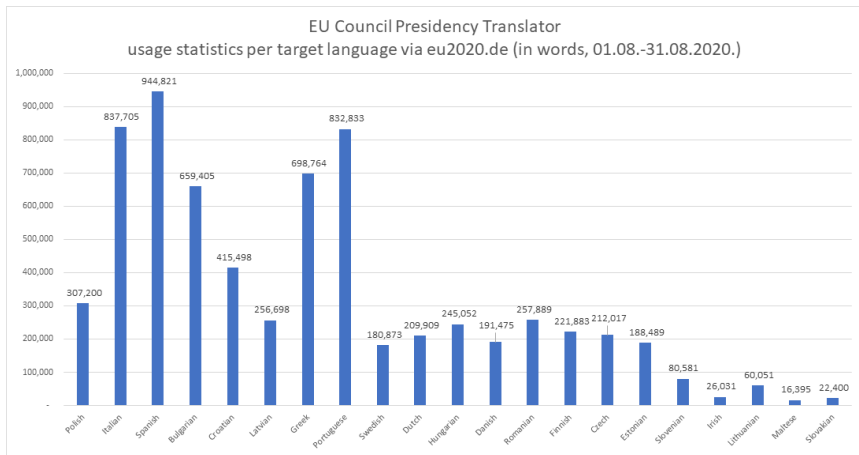
News

3 questions, 3 answers

- German
- English
- French
- Bulgarian
- Croatian
- Czech
- Danish

Europe & the EU Council Presidency Translator

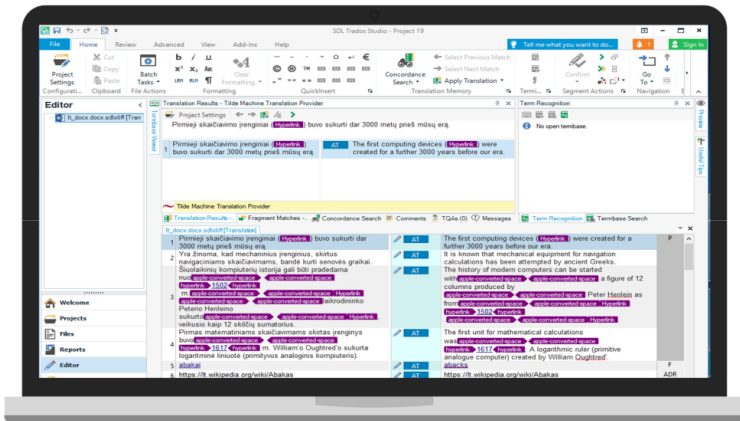
The Presidency Translator, 24 Languages



Europe & the EU Council Presidency Translator

The Presidency Translator, 24 Languages

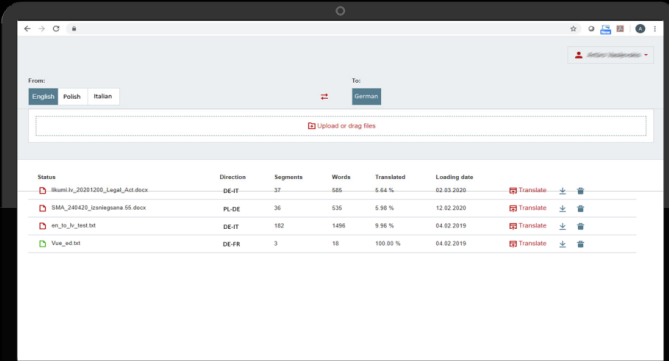
SDL Trados



















Europe & the EU Council Presidency Translator

The Presidency Translator, 24 Languages

Online CAT tools



The screenshot displays a web-based interface for a Computer-Assisted Translation (CAT) tool. At the top, there are language selection buttons for 'From:' (English, Polish, Italian) and 'To:' (German). Below this is a large text input area with a red 'Upload or drag files' prompt. The main part of the interface is a table listing translation jobs.

Status	Direction	Segments	Words	Translated	Loading date	
 ilumi_h_20201206_Legal_Act.docx	DE-IT	37	585	5.64 %	02.03.2020	  
 SMA_240420_Isriagaana 55.docx	PL-DE	36	535	5.98 %	12.02.2020	  
 en_h_1v_test.txt	DE-IT	182	1496	9.96 %	04.02.2019	  
 Voe_ed.txt	DE-FR	3	18	100.00 %	04.02.2019	  

On the right side of the laptop screen, there are three Microsoft Office icons: Word (blue 'W'), PowerPoint (orange 'P'), and Excel (green 'X').

Europe & the EU Council Presidency Translator

The Presidency Translator, 24 Languages

Online CAT tools

The screenshot displays an online CAT tool interface. The main workspace is divided into two columns: 'Source: de' and 'Translation: en'. The source text is in German, and the translation is in English. A progress bar on the right indicates the translation status for each segment. The first segment is 84% translated, the second is 100%, and the third is 81%. The interface also includes a search bar, a 'Prepare pre-translation' button, and a 'Download' button. On the right side, there are icons for Microsoft Word, PowerPoint, and Excel.

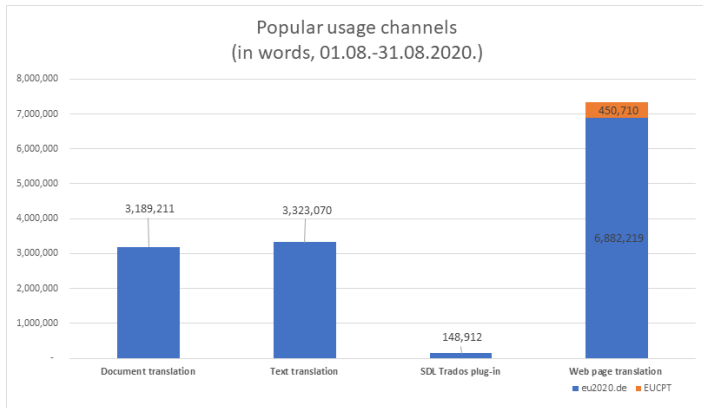
#	Source: de	Translation: en	Progress
1	Ministerkabinett mit dem Ziel, die Arbeit der Regierung zu erleichtern, werden in der Regel von den Mitgliedern des Kabinetts (Minister) und den Beamten der Regierung (Beamten) gebildet.	Initial report: Assessment Report of the staff Cabinet Regulations "Law on Requirements for Issuing Single Records and Issued Persons" (Ministerial Report)	84%
2	Die Arbeit der Regierung wird durch die Arbeit der Regierung erleichtert.	Single of the staff regulations	100%
3	Ministerkabinett mit dem Ziel, die Arbeit der Regierung zu erleichtern, werden in der Regel von den Mitgliedern des Kabinetts (Minister) und den Beamten der Regierung (Beamten) gebildet.	Objective, initial and project entry into force 20 February 2020	81%
4	Ministerkabinett mit dem Ziel, die Arbeit der Regierung zu erleichtern, werden in der Regel von den Mitgliedern des Kabinetts (Minister) und den Beamten der Regierung (Beamten) gebildet.	The staff rules have been designed to replace Cabinet regulations.	81%
5	Ministerkabinett mit dem Ziel, die Arbeit der Regierung zu erleichtern, werden in der Regel von den Mitgliedern des Kabinetts (Minister) und den Beamten der Regierung (Beamten) gebildet.		
6	Ministerkabinett mit dem Ziel, die Arbeit der Regierung zu erleichtern, werden in der Regel von den Mitgliedern des Kabinetts (Minister) und den Beamten der Regierung (Beamten) gebildet.		
7	Ministerkabinett mit dem Ziel, die Arbeit der Regierung zu erleichtern, werden in der Regel von den Mitgliedern des Kabinetts (Minister) und den Beamten der Regierung (Beamten) gebildet.		
8	Ministerkabinett mit dem Ziel, die Arbeit der Regierung zu erleichtern, werden in der Regel von den Mitgliedern des Kabinetts (Minister) und den Beamten der Regierung (Beamten) gebildet.		

Machine translation suggestions:

#	Suggestion	Progress
1	Initial report: Assessment Report of the staff Cabinet Regulations "Law on Requirements for Issuing Single Records and Issued Persons" (Ministerial Report)	MT
2	Initial report: Assessment Report of the staff Cabinet Regulations "Law on Requirements for Issuing Single Records and Issued Persons" (Ministerial Report)	88%
3	Initial report: Assessment Report of the staff Cabinet Regulations "Law on Requirements for Issuing Single Records and Issued Persons" (Ministerial Report)	84%

Europe & the EU Council Presidency Translator

The Presidency Translator, 24 Languages



Low Resource African Languages

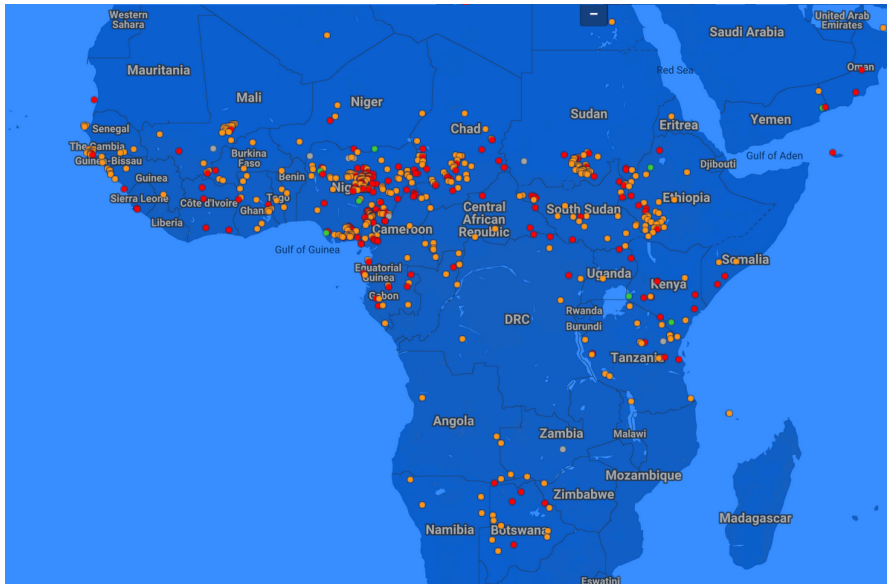
Language Diversity in Africa

Let's move to a very different project & reality

Thanks to Jesujoba O. Alabi, Kwabena Amponsah-Kaakyire
and David Adelani for sharing

Low Resource African Languages

Language Diversity in Africa



Low Resource African Languages

Language Diversity in Africa

- 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.

Low Resource African Languages

Language Diversity in Africa

- 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

Low Resource African Languages

Key NLP Resources

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 African Languages

Key NLP Resources

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)

Description	Source URL	#tokens	Status
Yorùbá			
Lagos-NWU corpus	github.com/Niger-Volta-LTI	24,868	clean
Alákòwé	alakoweyoruba.wordpress.com	24,092	clean
Òrò Yorùbá	oroyoruba.blogspot.com	16,232	clean
Èdè Yorùbá Rẹwà	deskgram.cc/edeyorubarewa	4,464	clean
Doctrine \$ Covenants	github.com/Niger-Volta-LTI	20,447	clean
Yorùbá Bible	www.bible.com	819,101	clean
GlobalVoices	yo.globalvoices.org	24,617	clean
Jehova Witness	www.jw.org/yo	170,203	clean
Ìrìnkèrindò nínú igbó elégbèje	manual	56,434	clean
Igbó Olódùmarè	manual	62,125	clean
JW300 Yorùbá corpus	opus.nlpl.eu/JW300.php	10,558,055	clean
Yorùbá Tweets	twitter.com/yobamoodua	153,716	clean
BBC Yorùbá	bbc.com/yoruba	330,490	noisy
Voice of Nigeria Yorùbá news	von.gov.ng/yoruba	380,252	noisy
Yorùbá Wikipedia	dumps.wikimedia.org/yowiki	129,075	noisy
Twi			
Bible	www.bible.com	661,229	clean
Jehovah's Witness	www.jw.org/tw	1,847,875	noisy
Wikipedia	dumps.wikimedia.org/twwiki	5,820	noisy
JW300 Twi corpus	opus.nlpl.eu/JW300.php	13,630,514	noisy

Low Resource African Languages

Quality of Embeddings

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

Model	Twi		Yorùbá	
	Vocab	Spearman ρ	Vocab	Spearman ρ
F1: Pre-trained Model (Wiki)	935	0.143	21,730	0.136
F2: Pre-trained Model (Common Crawl & Wiki)	NA	NA	151,125	0.073

Low Resource African Languages

Quality of Embeddings

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

Model	Twi		Yorùbá	
	Vocab	Spearman ρ	Vocab	Spearman ρ
F1: Pre-trained Model (Wiki)	935	0.143	21,730	0.136
F2: Pre-trained Model (Common Crawl & Wiki)	NA	NA	151,125	0.073
C1: Curated <i>Small</i> Dataset (Clean text)	9,923	0.354	12,268	0.322
C2: Curated <i>Small</i> Dataset (Clean + some noisy text)	18,494	0.388	17,492	0.302
C3: Curated <i>Large</i> Dataset (All Clean + Noisy texts)	47,134	0.386	44,560	0.391

Low Resource African Languages

Quality of Embeddings

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

Model	Twi		Yorùbá	
	Vocab	Spearman ρ	Vocab	Spearman ρ
F1: Pre-trained Model (Wiki)	935	0.143	21,730	0.136
F2: Pre-trained Model (Common Crawl & Wiki)	NA	NA	151,125	0.073
C1: Curated <i>Small</i> Dataset (Clean text)	9,923	0.354	12,268	0.322
C2: Curated <i>Small</i> Dataset (Clean + some noisy text)	18,494	0.388	17,492	0.302
C3: Curated <i>Large</i> Dataset (All Clean + Noisy texts)	47,134	0.386	44,560	0.391
D2: Curated <i>Small</i> Dataset (Clean + some noisy text)	22,851	0.437	56,086	0.345

Low Resource African Languages

Quality of Embeddings

Contextual embeddings

- NER F1 score on Global Voices Yorùbá corpus

Embedding Type	Date	Loc	Org	Per	Average
Pre-trained <i>uncased</i> Multilingual-bert (Multilingual vocab)	44.6	33.9	12.1	5.7	27.1 \pm 0.7
Fine-tuned <i>uncased</i> Multilingual-bert (Multilingual vocab)	64.0	65.3	38.8	47.4	56.4 \pm 2.4
Fine-tuned <i>uncased</i> Multilingual-bert (Yorùbá vocab)	67.0	71.5	40.4	49.4	60.1 \pm 0.8

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!

A close-up photograph of a typewriter keyboard. The focus is on a sheet of white paper with the word "Questions?" typed in a dark, monospaced font. The paper is held in place by a metal bar at the top and another at the bottom. The background is dark, and the lighting highlights the texture of the paper and the metallic surfaces of the typewriter.

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