

# Introduction to Language Identification Tools

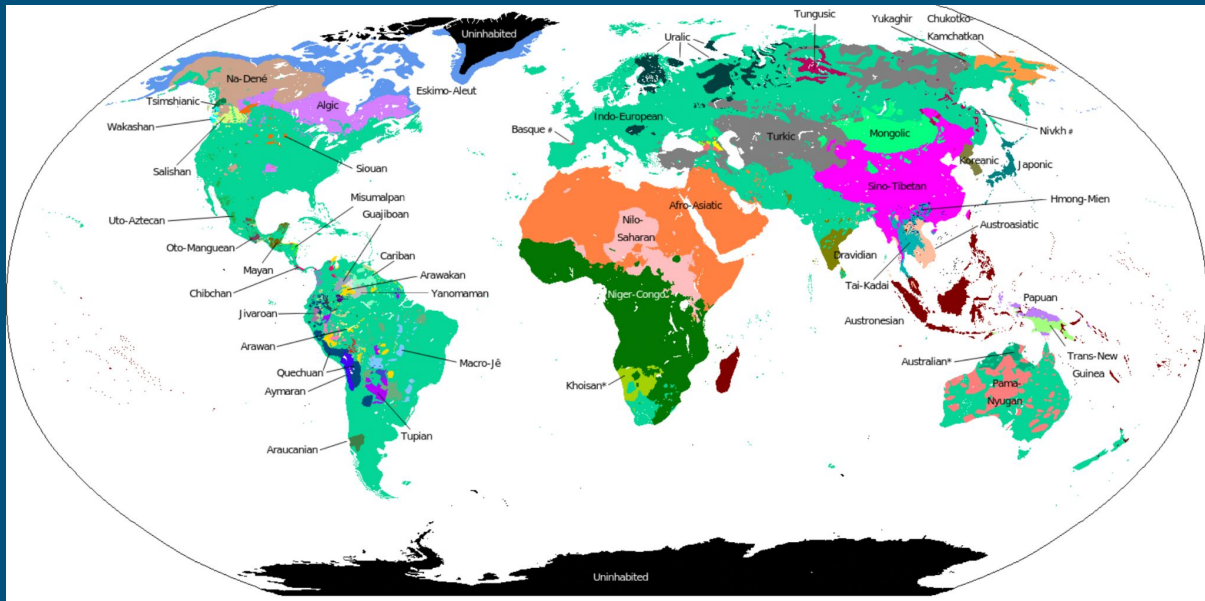
Wei-Rui Chen and Muhammad Abdul-Mageed

Deep Learning & Natural Language Processing Group  
The University of British Columbia



# Language Identification

A task to determine the language that a piece of text is written in (Jauhiainen et al. 2018)



Source: [https://en.wikipedia.org/wiki/List\\_of\\_language\\_families](https://en.wikipedia.org/wiki/List_of_language_families)

# What are these languages?

---

How are you?

English

Comment ça va ?

French

¿Cómo estás?

Spanish



# What are these languages?

---

こんにちは

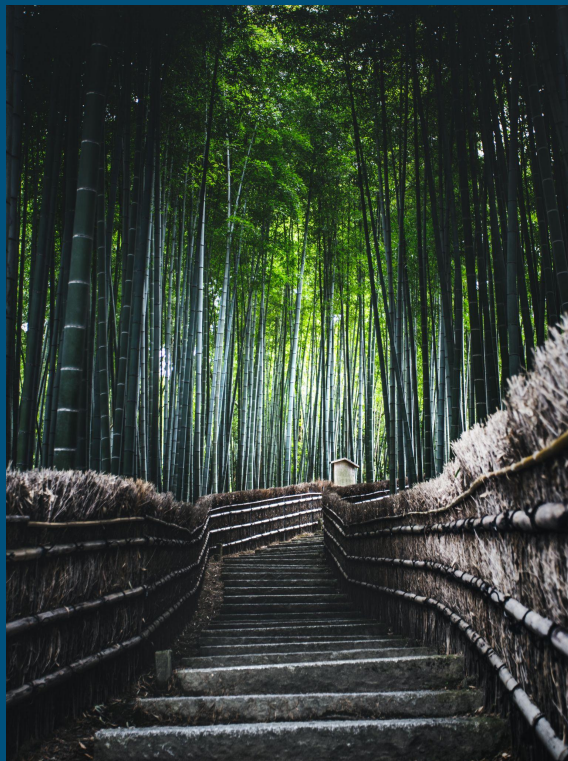
Japanese

你好

Chinese

안녕하세요

Korean



# What are these languages?

---

Ngiyajabula ukuthi kunabantu abazimisele ukungisekela.

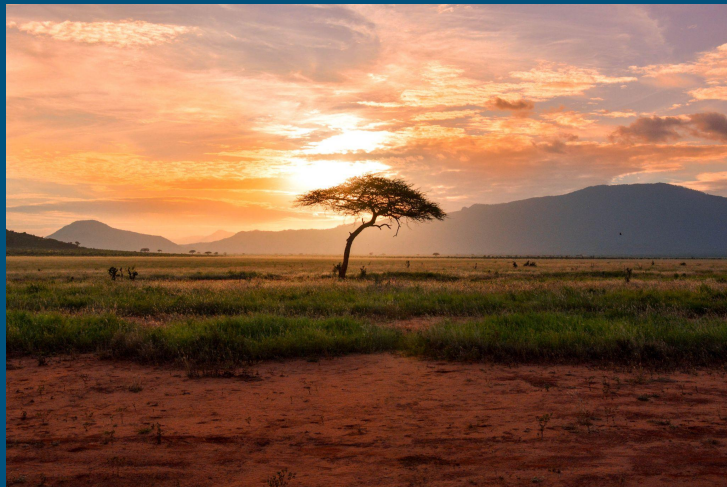
Zulu

Ndiyavuya kukho abantu abafuna ukundixhasa.

Xhosa

Inu mi dun pe awon eyan wa to fe durotimi.

Yoruba



# Potential Applications to Archival Science

---

Automatic classification of a text by language

Identifying languages at sentence-level

Automatic creation of metadata

# Language Identification Tools

---

# Langid.py

---

Lui et al. 2012

Naive Bayes

#Langs	#Babel-670 Common Langs	Acc	F1
97	78	92.39	88.80

Paper link: <https://aclanthology.org/P12-3005/>

Tool link: <https://github.com/saffsd/langid.py>



# CLD2 (Compact Language Detector 2)

---

By Google on 2013

Naive Bayes

#Langs	#Babel-670 Common Langs	Acc	F1
83	66	96.03	91.22

Tool link: <https://github.com/CLD2Owners/cld2>

# CLD3 (Compact Language Detector 3)

---

By Google on 2016

Neural Network

#Langs	#Babel-670 Common Langs	Acc	F1
106	83	96.02	89.53

Tool link: <https://github.com/google/cld3>

# Fasttext

---

Joulin et al. 2016

Neural Network

#Langs	#Babel-670 Common Langs	Acc	F1
176	101	83.77	74.02

Paper link: <https://arxiv.org/abs/1607.01759>, <https://arxiv.org/abs/1612.03651>

Tool link: <https://fasttext.cc/docs/en/language-identification.html>

# Franc

---

By Wormer from 2016 to now

model architecture unavailable

#Langs	#Babel-670 Common Langs	Acc	F1
414	216	81.05	66.28

Tool link: <https://github.com/woorm/franc>

# AfroLID

---

Adebara et al. 2022

Transformer neural network

#Langs	#Babel-670 Common Langs	Acc	F1
517	517	92.90	98.04

Paper link: <https://arxiv.org/abs/2210.11744>

Tool link: <https://github.com/UBC-NLP/afrolid>



Thank you!  
weirui.chen@ubc.ca

---