Addressing the Bender Rule for Growth Languages

JSALT-2025

Criticism: ACL is too focused on English

https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/

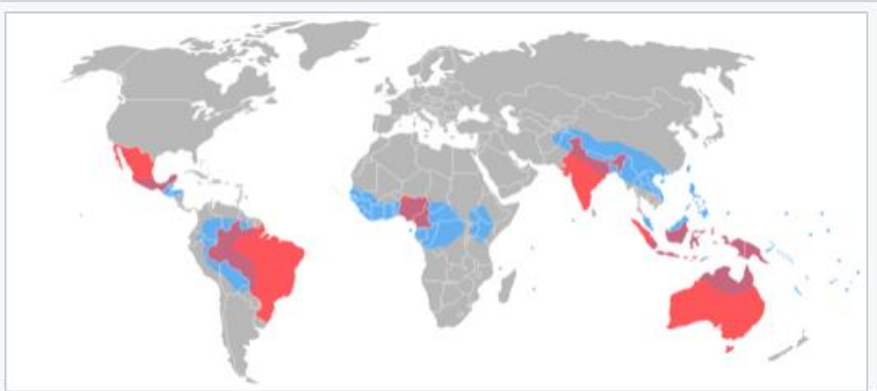
The #BenderRule: On Naming the Languages We Study and Why It Matters

14.SEP.2019 . 15 MIN READ

Emily M. Bender

Low Resources → Growth Opportunity

Endangered Languages



More than 50% of the world's endangered languages are located in just eight countries (denoted in red on the map): India, Brazil, Mexico, Australia, Indonesia, Nigeria, Papua New Guinea and Cameroon. In such countries and around them are the areas that are the most linguistically diverse in the world (denoted in blue on the map).

https://en.wikipedia.org/wiki/Endangered_language

50 Million Speakers or more

English, Chinese, Hindi, Spanish, Arabic, French, Bengali, Portuguese, Russian, Urdu, Indonesian, German, Japanese, Nigerian Pidgin, Marathi, Telugu, Turkish, Hausa, Tamil, Swahili, Tagalog, Punjabi, Korean, Persian, Javanese, Italian, Gujarati, Thai, Amharic, Kannada, Bhojpuri

https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

Good News

- We have more resources
  - for growth languages in 2020s
  - than for English in 1990s

Abstracts	Languages
$> 10^8$	en
10^6 - 10^7	zh, fr, es, id, pt, de
10^5 - 10^6	ko, ru, tr, ja, fa, sv, cs, it, uk, pl, nl, hr, ar
10^4 - 10^5	ca, ms, fi, hu, th, vi, sl, el, no, sh, sr, da, lt, af, ro, sk
10^3 - 10^4	lv, eu, hy, bg, et, kk, gl, uz, la, mk, az, hi, is, bs, eo, tg, mn, ne, he, sq, ka

Table 1: We ran fasttext-langdetect on 104M abstracts from Semantic Scholar. Many abstracts (17%) are not in English; 57 languages have 1000 abstracts or more.

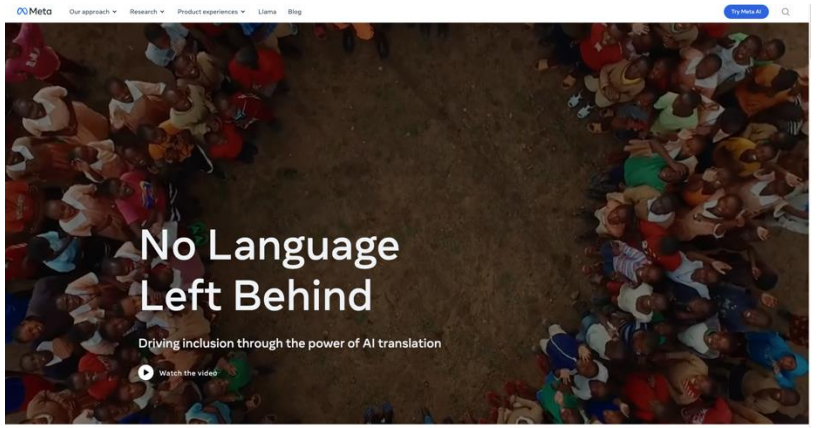
- No Language Left Behind
- Multilingual LLMs
- Google Translate
- Wikipedia (Common Crawl)
- Abstracts in Semantic Scholar
- Universal Dependencies
- WordNet
- Sketch Engine

No Language Left Behind (NLLB)

No Language Left Behind: Scaling Human-Centered Machine Translation

NLLB Team, Marta R. Costa-jussà[*], James Cross[*], Onur Çelebi[*], Maha Elbayad[*], Kenneth Heafield[*], Kevin Heffernan[*], Elahe Kalbassi[*], Janice Lam[*], Daniel Licht[*], Jean Maillard[*], Anna Sun[*], Skyler Wang[*,§], Guillaume Wenzek[*], Al Youngblood[*], Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews[†], Necip Fazil Ayan[†], Shruti Bhosale[†], Sergey Edunov[†], Angela Fan[†,‡], Cynthia Gao[†], Vedanuj Goswami[†], Francisco Guzmán[†], Philipp Koehn[†,¶], Alexandre Mourachko[†], Christophe Ropers[†], Safiyyah Saleem[†], Holger Schwenk[†], Jeff Wang[†]

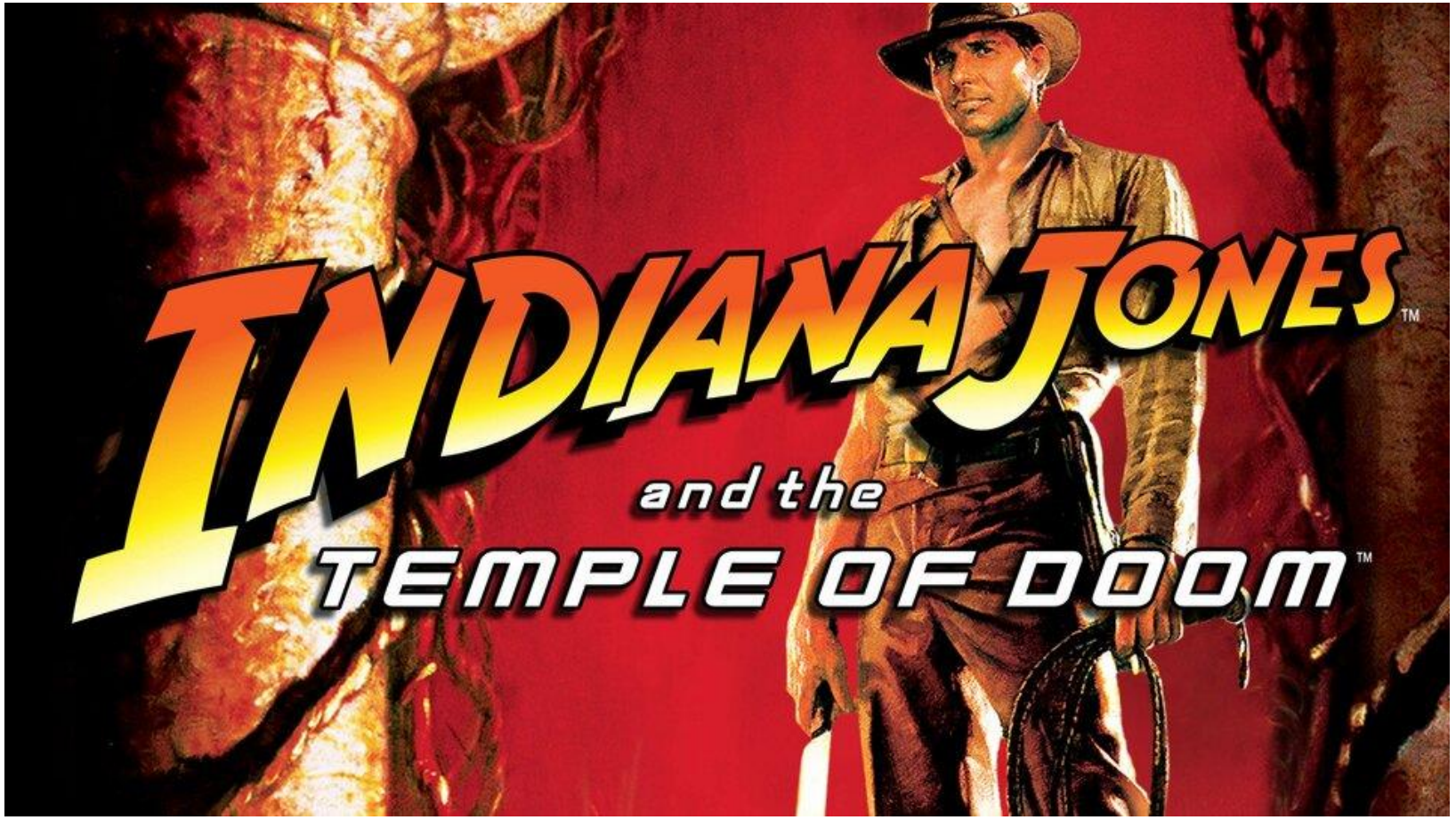
Meta AI, [§]UC Berkeley, [¶]Johns Hopkins University



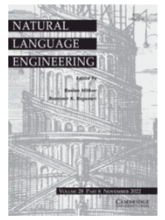
Concerns

- ➢Responsible AI (deadly and insanely profitable)
- Data quality
  - Too much machine translation
  - Too few quality controls (by people with local expertise)
- Lack of Expertise in Languages and Cultures
- Too much Emphasis on Translation
  - Is it appropriate to use English as a pivot language
- Is there a single model to rule them all?
  - Covering multiple perspectives

Move Fast and Break Things (Colonialism)



Responsible AI



Natural Language Engineering

Emerging trends: Unfair, biased, addictive, dangerous, deadly, and insanely profitable

Published online by Cambridge University Press: **19 December 2022**

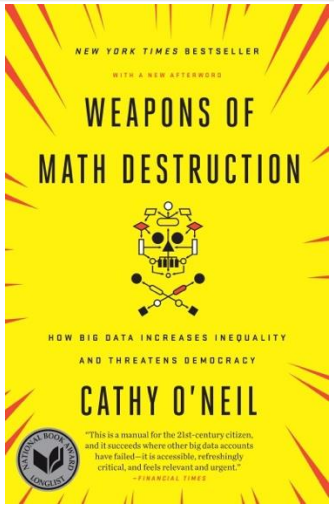
Kenneth Church, Annika Schoene, John E. Ortega, Raman Chandrasekar and Valia Kordoni

Show author details

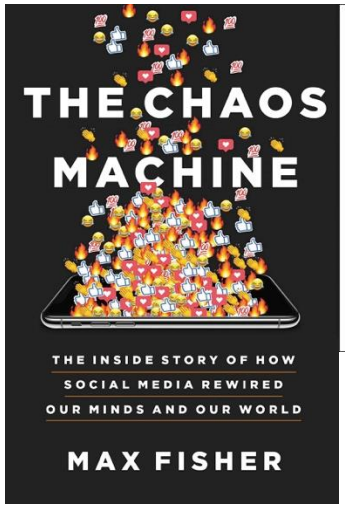
Article | Figures | Metrics

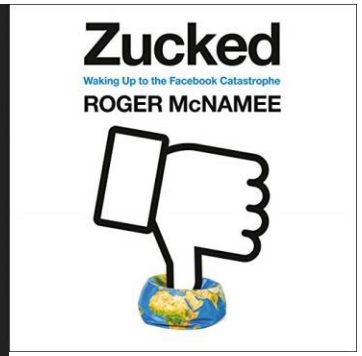
Save PDF | Share | Cite | Rights & Permissions

- Risks 1.0 (2016)
  - Unfair, Biased
- Risks 2.0 (2022)
  - Addictive, Dangerous, Deadly and Insanely Profitable



- American social media companies are not great at moderating online forums,
  - but they are worse in languages and cultures
  - that they do not understand all that well.





Concerns

✓Responsible AI (deadly and insanely profitable)

➢Data quality

- Too much machine translation
- Too few quality controls (by people with local expertise)

- Lack of Expertise in Languages and Cultures
- Too much Emphasis on Translation
  - Is it appropriate to use English as a pivot language
- Is there a single model to rule them all?
  - Covering multiple perspectives

Engage Local Expertise

- Hate speech is more toxic
  - in Hausa tweets than
  - in English tweets (in Nigeria)
- Quality control is relatively weak in 4 African Languages
  - *A significant part of the translations were suspected to have been automatically generated*

Correcting FLORES Evaluation Dataset for Four African Languages

Idris Abdulmumin[1*+], Sthembiso Mkhwanazi[2], Mahlatse S. Mbooi[2], Shamsuddeen Hassan Muhammad[3*+], Ibrahim Said Ahmad[4*+], Neo Putini[5], Miehleketo Mathebula[1], Matimba Shingange[1], Tajuddeen Gwadabe[*+], Vukosi Marivate[1,6]

[1]Data Science for Social Impact, University of Pretoria, [2]Council for Scientific and Industrial Research, South Africa, [3]Imperial College, London, [4]Northeastern University, [5]University of KwaZulu-Natal, [6]Lelapa AI, *HausaNLP, +MasakhaneNLP

correspondence: idris.abdulmumin@up.ac.za

Google Scholar

Ibrahim Said Ahmad FOLLOW

Northeastern University
Verified email at buk.edu.ng - Homepage
Natural Language Processing Big Data Data mining Artificial Intelligence

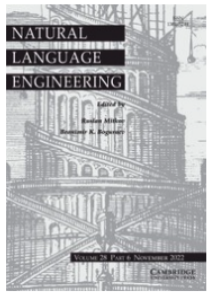
TITLE	CITED BY	YEAR
Adaption of distance learning to continue the academic year amid COVID-19 lockdown A Qazi, J Qazi, K Naseer, M Zeeshan, S Qazi, O Abayomi-Alli, IS Ahmad, ... Children and Youth Services Review 126, 106038	82	2021
Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis SH Muhammad, DI Adelani, S Ruder, IS Ahmad, I Abdulmumin, BS Bello, ... arXiv preprint arXiv:2201.08277	74	2022
A hybrid metaheuristic method in training artificial neural network for bankruptcy prediction A Ansari, IS Ahmad, AA Bakar, MR Yaakub IEEE access 8, 176640-176650	65	2020
Movie revenue prediction based on purchase intention mining using YouTube trailer reviews IS Ahmad, AA Bakar, MR Yaakub Information Processing & Management 57 (5), 102278	58	2020
Afrisenti: A twitter sentiment analysis benchmark for african languages SH Muhammad, I Abdulmumin, AA Ayele, N Ousidhoum, DI Adelani, ... arXiv preprint arXiv:2302.08956	56	2023
The role of information & communication technology in elearning environments: a systematic review A Qazi, G Hardaker, IS Ahmad, M Darwich, JZ Maitama, A Dayani Ieee Access 9, 45539-45551	55	2021
SemEval-2023 task 12: sentiment analysis for african languages (AfriSenti-SemEval) SH Muhammad, I Abdulmumin, SM Yimam, DI Adelani, IS Ahmad, ... arXiv preprint arXiv:2304.06845	45	2023

Concerns

- ✓Responsible AI (deadly and insanely profitable)
- ✓Data quality
  - ✓Too much machine translation
  - ✓Too few quality controls (by people with local expertise)
- ✓Lack of Expertise in Languages and Cultures
- ✓Too much Emphasis on Translation
  - ✓Is it appropriate to use English as a pivot language
- ✓Is there a single model to rule them all?
  - ➢Covering multiple perspectives

Lack of Perspective in Chatbots

- Homework for students in my NLP class
  - Write essays on Opium Wars
    - from six perspectives including East and West
  - Feel free to use online tools: chatbots, search, etc.
    - (but you are responsible for the content)
- Chatbots view everything
  - from a single (American) perspective
- Concern: even international students
  - handed in essays with Western perspective,
  - though they were taught another perspective in China



Natural Language Engineering

Emerging trends: When can users trust GPT, and when should they intervene?

Published online by Cambridge University Press: **16 January 2024**

Kenneth Church

Show author details

Article | Figures | Metrics

Save PDF | Share | Cite | Rights & Permissions

Adding Perspective to Vision Research: Vision is more than classifying pixels

Asirra: A CAPTCHA that Exploits Interest-Aligned Manual Image Categorization

Jeremy Elson, John R. Douceur, Jon Howell
Microsoft Research
{jelson,johndo,howell}@microsoft.com

Jared Saul
Petfinder, Inc.
jared@petfinder.com

ABSTRACT

We present Asirra (Figure 1), a CAPTCHA that asks users to identify cats out of a set of 12 photographs of both cats and dogs. Asirra is easy for users; user studies indicate it can be solved by humans 99.6% of the time in under 30 seconds. Barring a major advance in machine vision, we expect computers will have no better than a 1/54,000 chance of solving it. Asirra's image database is provided by a novel, mutually beneficial partnership with Petfinder.com. In exchange for the use of their three million images, we display an "adopt me" link beneath each one, promoting Petfinder's primary mission of finding homes for homeless animals. We describe the design of Asirra, discuss threats to its security, and report early deployment experiences. We also describe two novel algorithms for amplifying the skill gap between humans and computers that can be used on many existing CAPTCHAs.

1. INTRODUCTION

https://www.microsoft.com/en-us/research/wp-content/uploads/2007/10/CCS2007.pdf

Adding Perspective to Vision Research
COCO (Objective) → ArtEmis (Subjective)

Subjective Emotion

More room for Audience Background: Culture/Language

(a) ArtEmis: *I love everything about this painting of a mother and her two children lovingly interacting with the family pet cat.*

(b) COCO: *A man and a woman holding a little kid while sitting at a table outside*

Objective Facts

Figure 2: COCO captures the facts, and ArtEmis enhances those facts with emotion/commentary.

Adding More Perspectives to Vision (Do not pivot via English)

Figure 1: **ArtElingo-28 Benchmark:** 9 emotion labels with captions in 28 languages. The ~140 annotations per WikiArt image embrace diversity over languages and cultures.

- Input: Artworks
- Annotations:
  - 9 Emotion Labels
    - 4 Positive:
      - Contentment, Awe, Excitement, Amusement
    - 4 Negative:
      - Sadness, Anger, Fear, Disgust
    - 1 Neutral
  - Caption in 28 languages
    - English glosses added to paper
    - to help readers (but not in corpus)

JSALT-2025 Teamwork Opportunity: Zoom → Face-to-Face

No Culture Left Behind: ArtELingo-28, a Benchmark of WikiArt with Captions in 28 Languages

Youssef Mohamed[1]* Runjia Li [2] Ibrahim Said Ahmad[3] Kilichbek Haydarov [1]
Philip Torr[2] Kenneth Ward Church [3] Mohamed Elhoseiny[1]*
[1]KAUST [2] University of Oxford [3] Northeastern University

https://www.artelingo.org/

Google Scholar

Ibrahim Said Ahmad

FOLLOW

Northeastern University
Verified email at buk.edu.ng - Homepage
Natural Language Processing Big Data Data mining Artificial Intelligence

TITLE	CITED BY	YEAR
Adaption of distance learning to continue the academic year amid COVID-19 lockdown A Qazi, J Qazi, K Naseer, M Zeeshan, S Qazi, O Abayomi-Alli, IS Ahmad, ... Children and Youth Services Review 126, 106038	82	2021
Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis SH Muhammad, DI Adelani, S Ruder, IS Ahmad, I Abdulmumin, BS Bello, ... arXiv preprint arXiv:2201.08277	74	2022

Youssef Mohamed

FOLLOW

King Abdullah University of Science and Technology(KAUST)
Verified email at kaust.edu.sa
Machine Learning Vision and Language Affective Datasets Multimodality Deep Learning

TITLE	CITED BY	YEAR
ArtELingo: A Million Emotion Annotations of WikiArt with Emphasis on Diversity over Language and Culture Y Mohamed, M Abdelfattah, S Alhuwaider, F Li, X Zhang, KW Church, ... arXiv preprint arXiv:2211.10780	17	2022
It is Okay to Not Be Okay: Overcoming Emotional Bias in Affective Image Captioning by Contrastive Data Collection Y Mohamed, FF Khan, K Haydarov, M Elhoseiny Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern …	35	2022

Quality Control

Benchmark	Languages	Annotators	Annotation Hours	Coordinators	Coordinator Hours
ArteLingo	3 (en, zh, ar)	7.8K turkers	32K	3 students	??
ArteLingo-28	28	220	6.3K	32	2.5K

	ArtEmis	ArtELingo	ArtELingo-28
Image Source	WikiArt	WikiArt	WikiArt
Languages	1	3	3+ **25**
#Images	80k	80k	80k(3) , **2K (25)**
#Annotations	0.45M	1.2M	1.2M(3) + **200K (25)**
#Annot/Image	5.68	15.3	15.3(3) + **140 (25)**
Emotions	9	9	9 (3), **9 (25)**

Table 1: A Comparison of Datasets. ArtELingo-28 extends ArtELingo (Mohamed et al., 2022a) with 200K annotations from 25 additional different languages, many are low-resource.

Embrace Diversity & Multiple Perspectives

- emotion labels:
  - disgust (Burmese)
  - awe (Malay)
- captions focus
  - on chest
    - Burmese & Malay
  - on face and hair
    - Korean & Setswana

Hypothesis: Predictable

Vision: More than Pixels
Communication (Shannon): Audience Matters

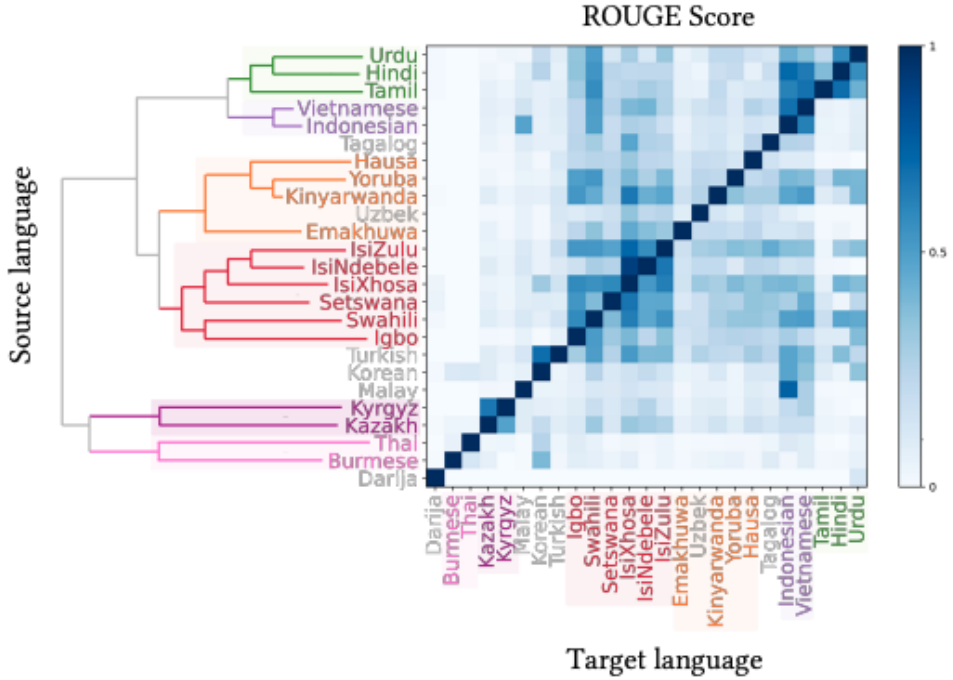


Figure 6: **One vs All Zero-Shot.** The figure shows the rouge score on the target languages. On the left the clustering reveals cultural connections. The captioning scores reveal groups that align with real world cultural connections. This clustering suggests that our trained models can capture the cultural signal.

- Art is a form of communication
  - between artist and audience
- Background matters
  - Language
  - Culture
- Task: Art, Lang → Caption
  - English pivot baseline:
    - Translate caption from English

Deliverables

- Survey current status
  - Identify available resources
  - Evaluate strengths and weaknesses
- Engage Local Expertise
  - Bring together people that would not normally work together
- Establish that we can do better than pivoting via English
- Resources
  - Re-use (or create, if necessary)
  - to encourage more research
    - on growth languages
    - embracing diversity
    - and multiple perspectives
- Develop a catalog/hub for
  - datasets, models and other resources
  - like HuggingFace
- Inclusive playground for
  - Sharing: Encourage community to upload resources
  - Ease-of –Use
    - Easy to find popular stuff
    - Easy to try things out (without downloading)