Carnegie Mellon University

(WIP) GlobalBench: A Benchmark for Global Progress in Natural Language Processing

Graham Neubig

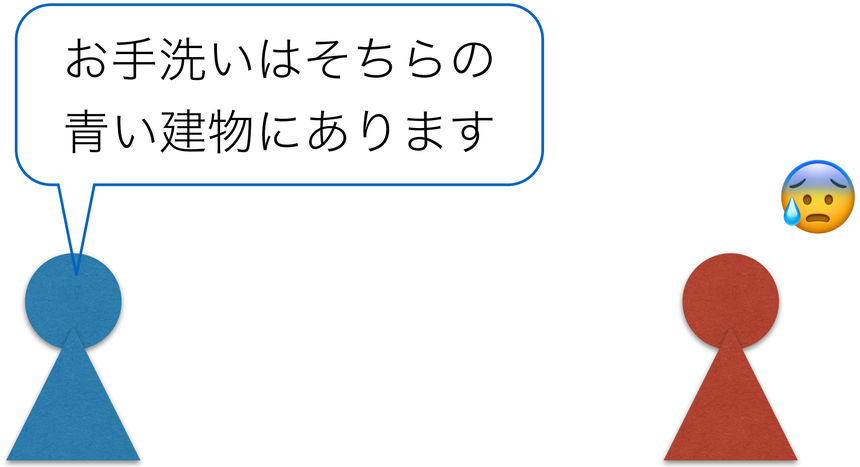
w/ Catherine Cui, Pengfei Liu, Fahim Faisal, Alissa Ostapenko, Yulia Tsvetkov, Antonios Anastasopoulos

and You!

https://github.com/neulab/globalbench

Why do we Build Language Technology?

Why do we Build Language Technology?



Why do we Build Language Technology?

Why do we Build Language Technology?

- We want to make the world a better place!

Why do we Build Language Technology?

- We want to make the world a better place!
- How do we quantify "better"?

Why do we Build Language Technology?

- We want to make the world a better place!
- How do we quantify "better"?
- **Utility (economics):** the total satisfaction received from consuming a good or service.

How Much Utility Do We Derive from Language Technology?

Image Credit: Wikipedia

How Much Utility Do We Derive from Language Technology?

- **American English Speaker:** Use Google assistant, car navigation system, translate text, benefit from good search technology

Image Credit: Wikipedia

How Much Utility Do We Derive from Language Technology?

- **American English Speaker:** Use Google assistant, car navigation system, translate text, benefit from good search technology
- **Japanese Speaker:** Use the above technology, maybe with fewer features, maybe a bit worse

Image Credit: Wikipedia

How Much Utility Do We Derive from Language Technology?

- **American English Speaker:** Use Google assistant, car navigation system, translate text, benefit from good search technology

- **Japanese Speaker:** Use the above technology, maybe with fewer features, maybe a bit worse

- **Marshalese Speaker:** Don't use the above technology, or be forced to use it in a second language

Image Credit: Wikipedia

How Much Utility Do We Derive from Language Technology?

- **American English Speaker:** Use Google assistant, car navigation system, translate text, benefit from good search technology

- **Japanese Speaker:** Use the above technology, maybe with fewer features, maybe a bit worse

- **Marshalese Speaker:** Don't use the above technology, or be forced to use it in a second language

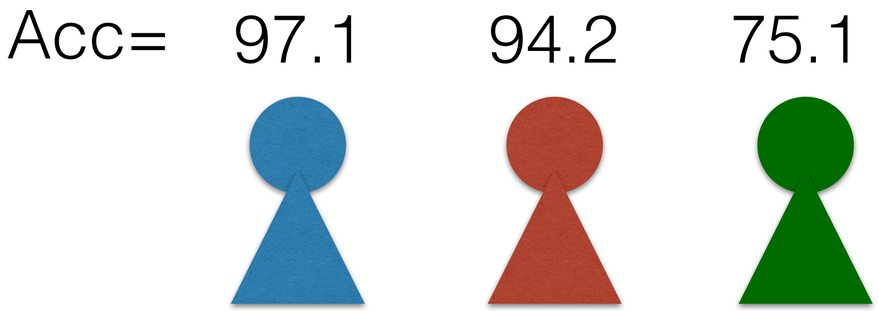
- **English Speaker with a Greek accent:** ~~Use Google assistant,~~ car navigation system, translate text, benefit from good search technology

Image Credit: Wikipedia

A Recipe for Progress?

A Recipe for Progress?

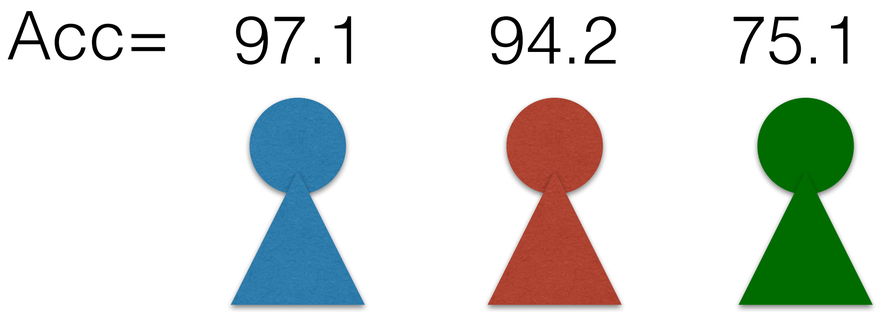
- **Quantify** disparities in language technology performance



A Recipe for Progress?

- **Quantify** disparities in language technology performance

- **Link** performance to derived utility



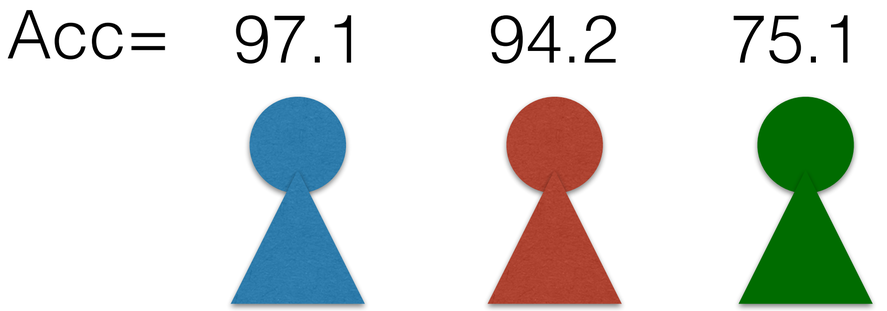
Accuracy=94.2

😀 ?

😰 ?

A Recipe for Progress?

- **Quantify** disparities in language technology performance



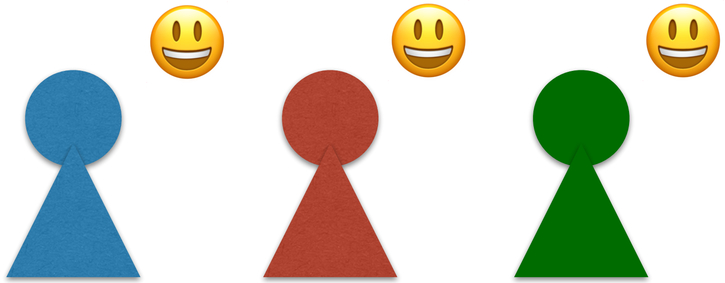
- **Link** performance to derived utility

Accuracy=94.2

😀 ?

😰 ?

- **Mitigate** disparities through research community incentives



Multilingual Leaderboards

Multilingual Leaderboards

...

Multilingual Leaderboards

XTREME

...

Multilingual Leaderboards

XTREME XGLUE ...

Multilingual Leaderboards

XTREME **XGLUE** ...

CoNLL 2018 Shared Task

Multilingual Leaderboards

XTREME XGLUE CoNLL 2018 Shared Task ...

- Very popular! But what does “multilingual” mean?

Multilingual Leaderboards

XTREME **XGLUE** ...

CoNLL 2018 Shared Task

- Very popular! But what does “multilingual” mean?
- **XTREME:** "availability of monolingual data, and typological diversity"

Multilingual Leaderboards

XTREME **XGLUE** ...

CoNLL 2018 Shared Task

- Very popular! But what does “multilingual” mean?
- **XTREME:** "availability of monolingual data, and typological diversity"
- **XGLUE:** ??? (as far as I can tell, not stated)

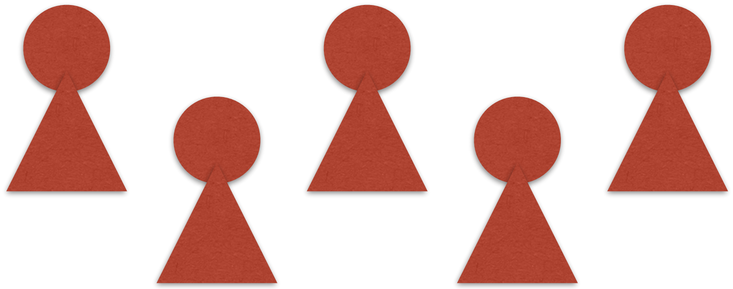
Multilingual Leaderboards

XTREME **XGLUE** ...

CoNLL 2018 Shared Task

- Very popular! But what does “multilingual” mean?
- **XTREME:** "availability of monolingual data, and typological diversity"
- **XGLUE:** ??? (as far as I can tell, not stated)

But what about **people?!**



Global Utility Metrics

(Blasi et al. 2022)

Global Utility Metrics

(Blasi et al. 2022)

- A language technology should be measured by the **utility** it provides to **every person in the world**

Global Utility Metrics

(Blasi et al. 2022)

- A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

Global Utility Metrics

(Blasi et al. 2022)

- A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

- Two problems:

Global Utility Metrics

(Blasi et al. 2022)

- A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

- Two problems:
  - **Problem 1:** how to measure utility of an NLP system?
    → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)

Global Utility Metrics

(Blasi et al. 2022)

- A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

- Two problems:
  - **Problem 1:** how to measure utility of an NLP system?
    → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)
  - **Problem 2:** how to consider different utility provided to every person in the world?
    → Measure over subgroups (here, languages), weighted by demand + coefficient τ.

Global Utility Metrics

(Blasi et al. 2022)

- A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

- Two problems:
  - **Problem 1:** how to measure utility of an NLP system?
    → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)
  - **Problem 2:** how to consider different utility provided to every person in the world?
    → Measure over subgroups (here, languages), weighted by demand + coefficient τ.

$$M_\tau = \sum_{l \in \mathcal{L}} d_l^{(\tau)} \cdot u_l$$

Global Utility Metrics

(Blasi et al. 2022)

- A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

- Two problems:
  - **Problem 1:** how to measure utility of an NLP system?
    → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)
  - **Problem 2:** how to consider different utility provided to every person in the world?
    → Measure over subgroups (here, languages), weighted by demand + coefficient τ.

$$M_\tau = \sum_{l \in \mathcal{L}} d_l^{(\tau)} \cdot u_l$$

"normalized demand"

Global Utility Metrics

(Blasi et al. 2022)

- A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

- Two problems:
  - **Problem 1:** how to measure utility of an NLP system?
    → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)
  - **Problem 2:** how to consider different utility provided to every person in the world?
    → Measure over subgroups (here, languages), weighted by demand + coefficient τ.

$$M_\tau = \sum_{l \in \mathcal{L}} d_l^{(\tau)} \cdot u_l$$

"normalized demand"

"utility"

Global Utility Metrics

(Blasi et al. 2022)

- A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

- Two problems:
  - **Problem 1:** how to measure utility of an NLP system?
    → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)
  - **Problem 2:** how to consider different utility provided to every person in the world?
    → Measure over subgroups (here, languages), weighted by demand + coefficient τ.

$$M_\tau = \sum_{l \in \mathcal{L}} d_l^{(\tau)} \cdot u_l \qquad d_l^{(\tau)} = \frac{n_l^\tau}{\sum_{l' \in \mathcal{L}} n_{l'}^\tau}$$

"normalized demand"

"utility"

Global Utility Metrics

(Blasi et al. 2022)

- A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

- Two problems:
  - **Problem 1:** how to measure utility of an NLP system?
    → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)
  - **Problem 2:** how to consider different utility provided to every person in the world?
    → Measure over subgroups (here, languages), weighted by demand + coefficient τ.

$$M_\tau = \sum_{l \in \mathcal{L}} d_l^{(\tau)} \cdot u_l$$

"normalized demand"

"utility"

$$d_l^{(\tau)} = \frac{n_l^\tau}{\sum_{l' \in \mathcal{L}} n_{l'}^\tau}$$

τ=1 : every person equal ("demographic-average utility")

Global Utility Metrics

(Blasi et al. 2022)

- A language technology should be measured by the **utility** it provides to **every person in the world**

$$M = \sum_i u_i$$

- Two problems:
  - **Problem 1:** how to measure utility of an NLP system?
    → Very hard, use standard accuracy metrics as a proxy now (happy to discuss more!)
  - **Problem 2:** how to consider different utility provided to every person in the world?
    → Measure over subgroups (here, languages), weighted by demand + coefficient τ.

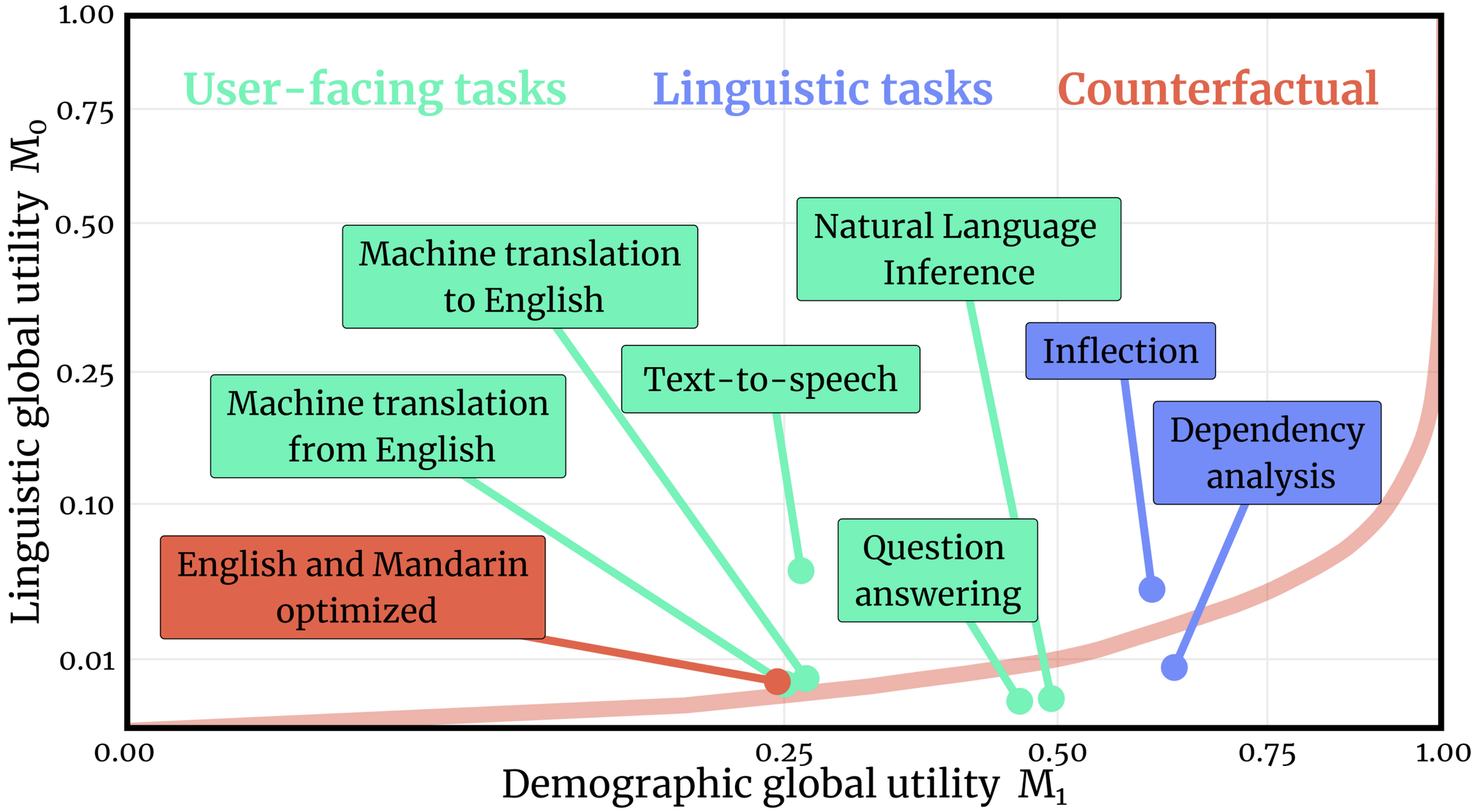
$$M_\tau = \sum_{l \in \mathcal{L}} d_l^{(\tau)} \cdot u_l$$

"normalized demand" "utility"

$$d_l^{(\tau)} = \frac{n_l^\tau}{\sum_{l' \in \mathcal{L}} n_{l'}^\tau}$$

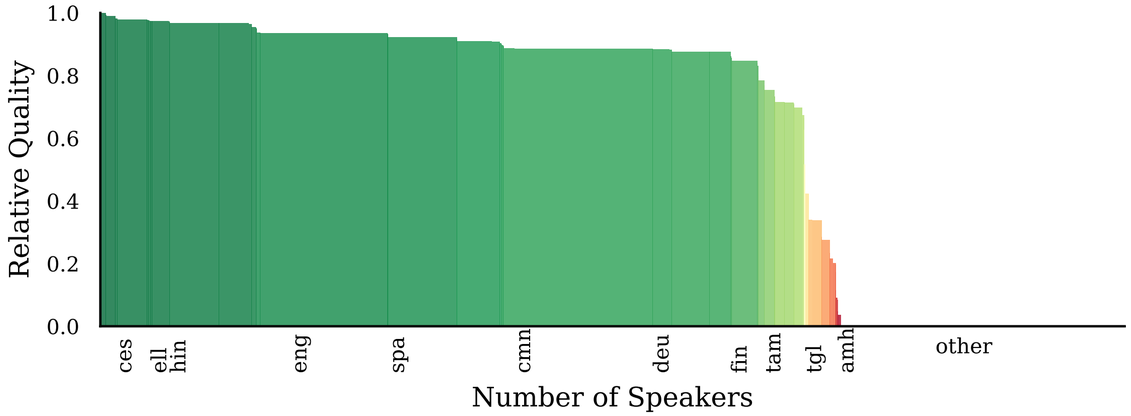
τ=1 : every person equal ("demographic-average utility")

τ=0 : every subgroup equal ("linguistic-average utility")

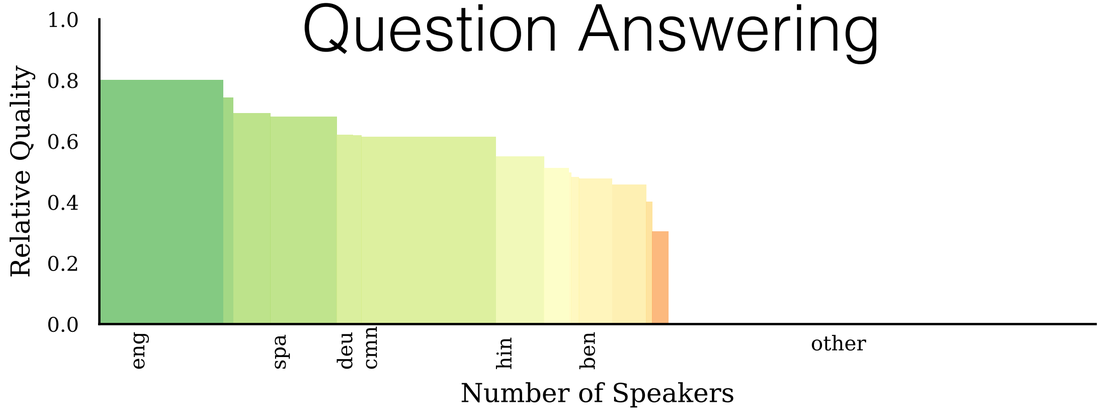

User-facing tasks
Linguistic tasks
Counterfactual
Machine translation to English
Natural Language Inference
Text-to-speech
Inflection
Machine translation from English
Dependency analysis
Question answering
English and Mandarin optimized
1.00
0.75
0.50
0.25
0.10
0.01
Linguistic global utility M_0
0.00
0.25
0.50
0.75
1.00
Demographic global utility M_1

Zooming In

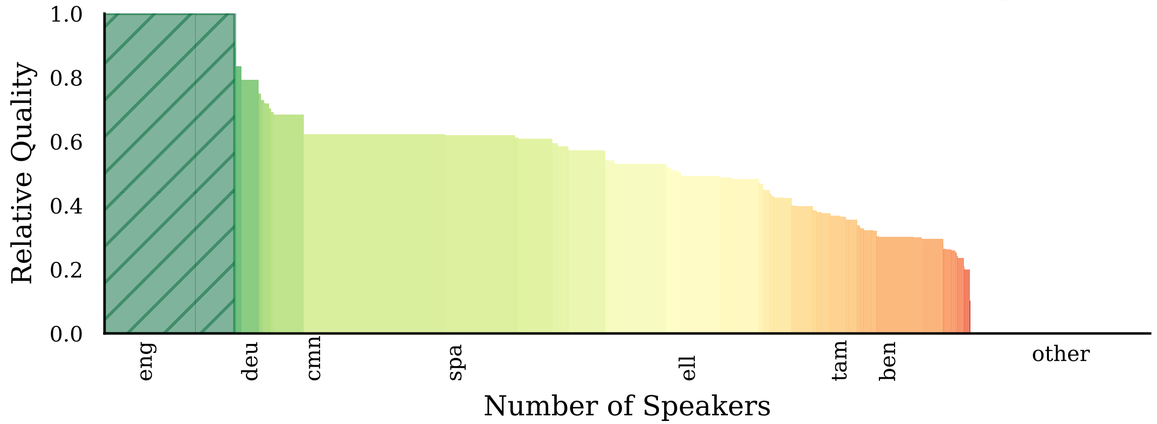
Dependency Parsing



Question Answering

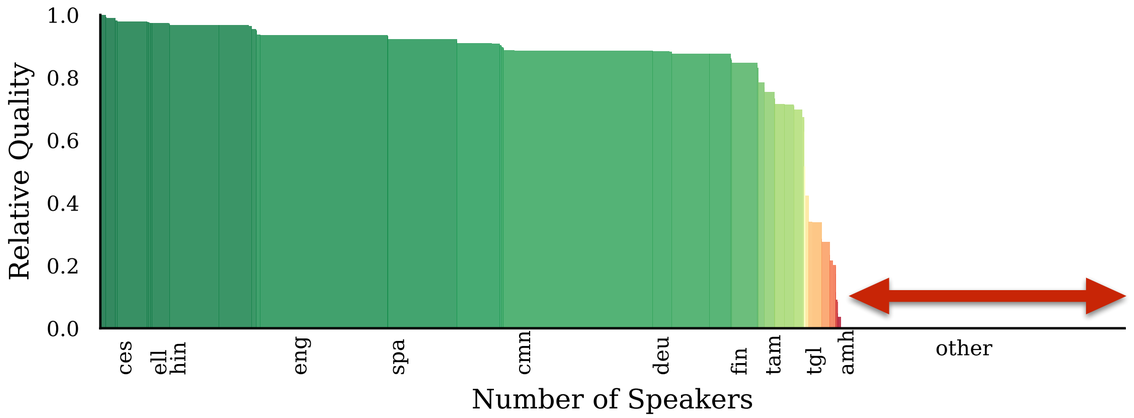


Machine Translation to English

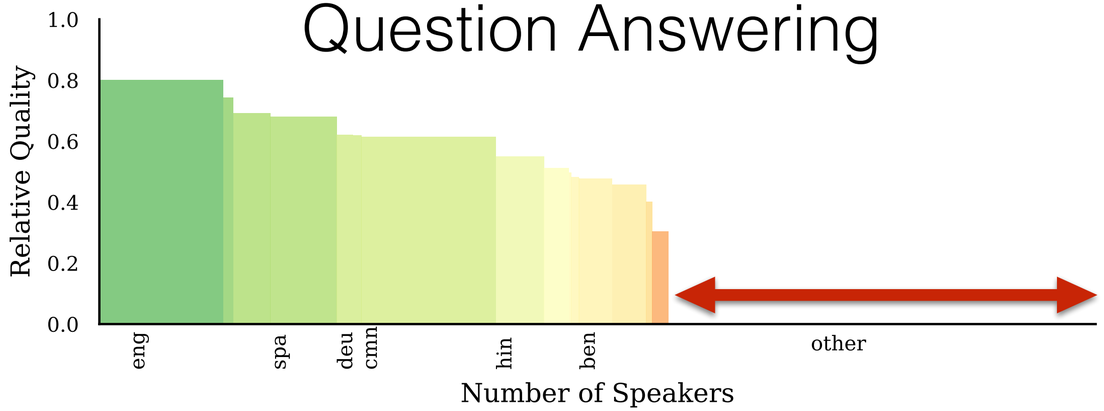


Zooming In

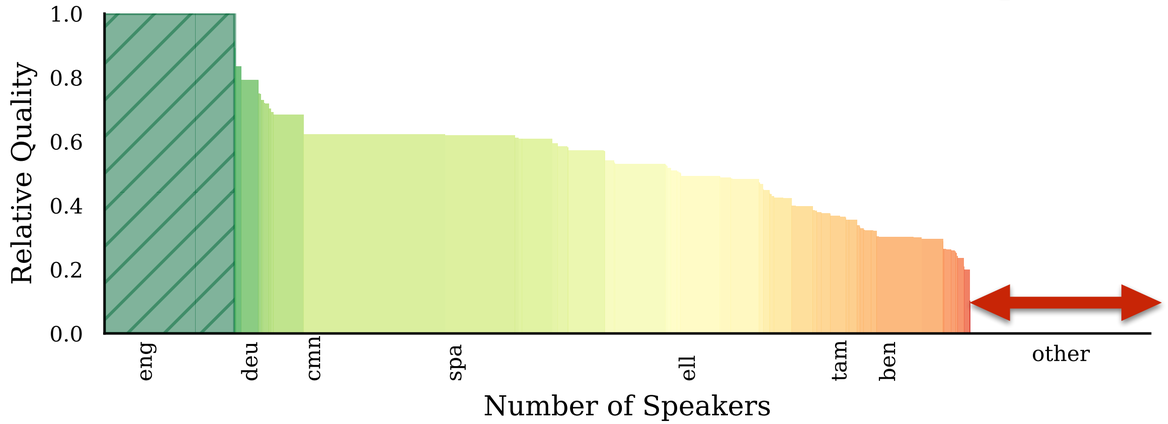
Dependency Parsing



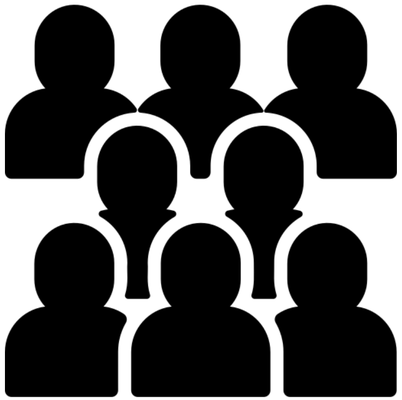
Question Answering



Machine Translation to English

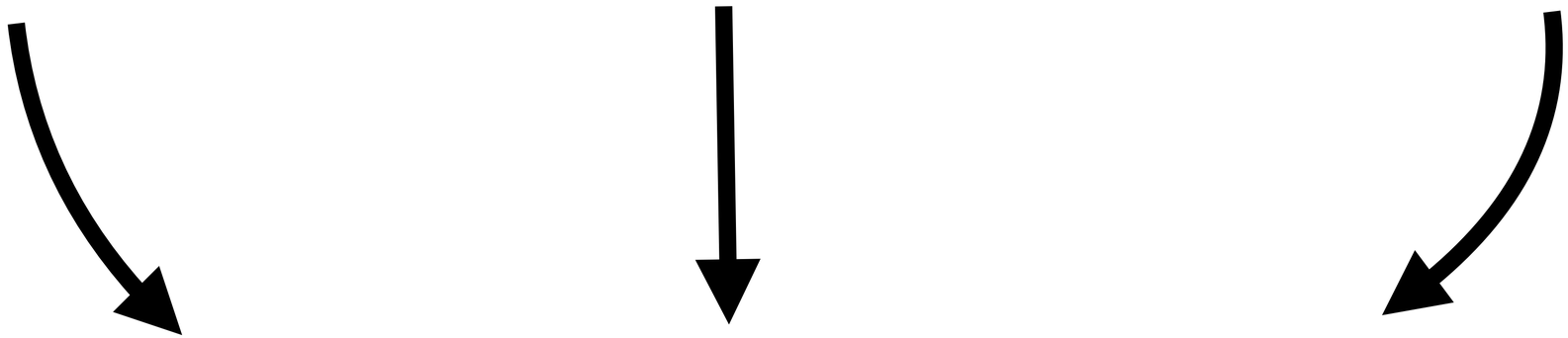


Demographic incentives



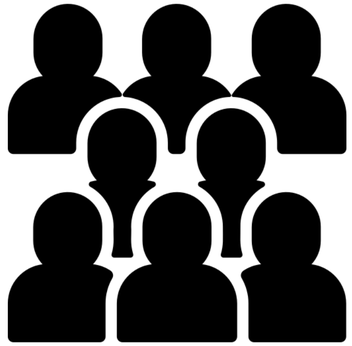
Financial incentives

Academic/professional incentives



Language technologies R&D

Demographic incentives

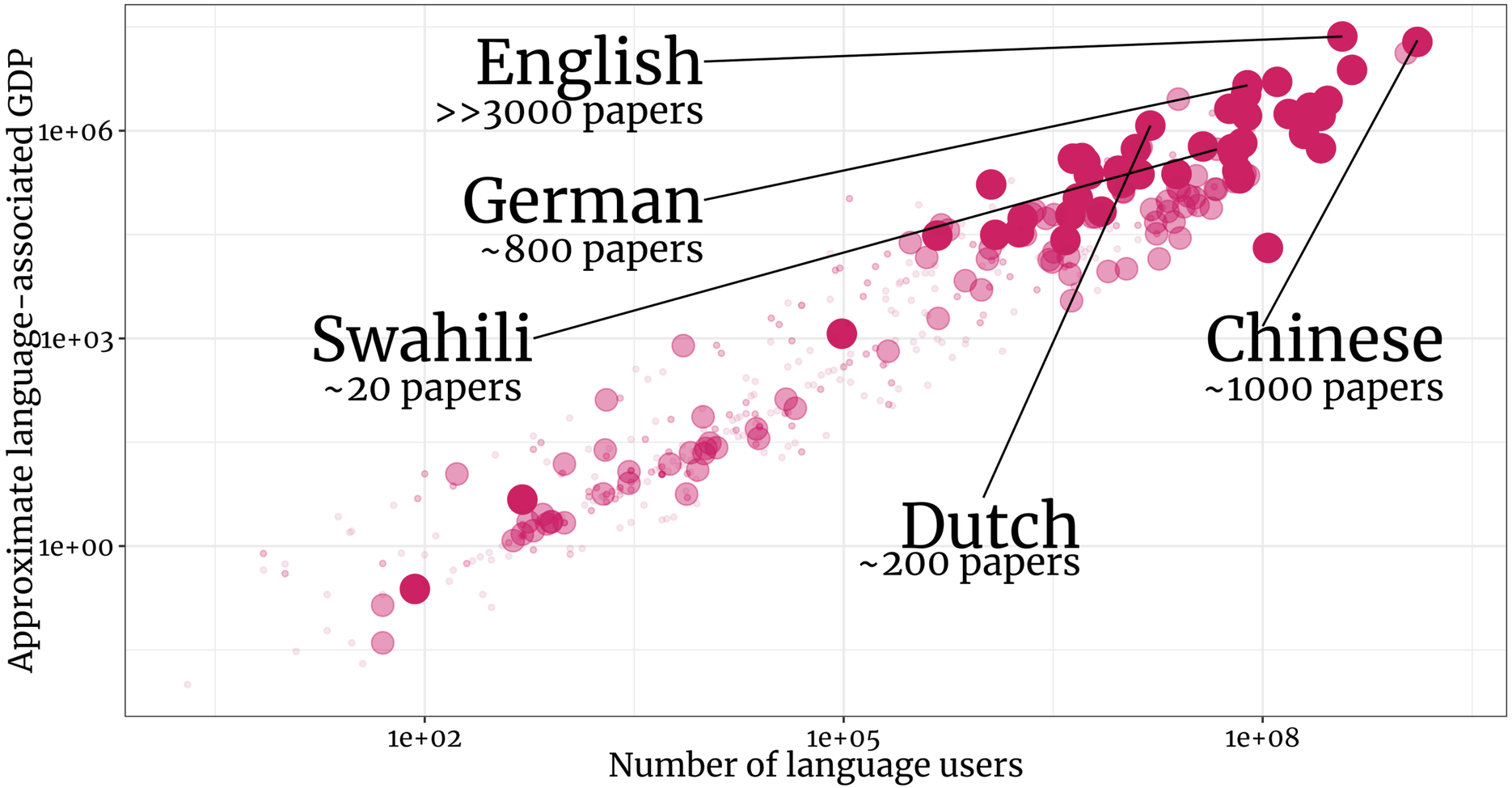


≈ number language users

Financial incentives

≈ approximate GDP associated with language

Which one better predicts the total number of papers on a given language?



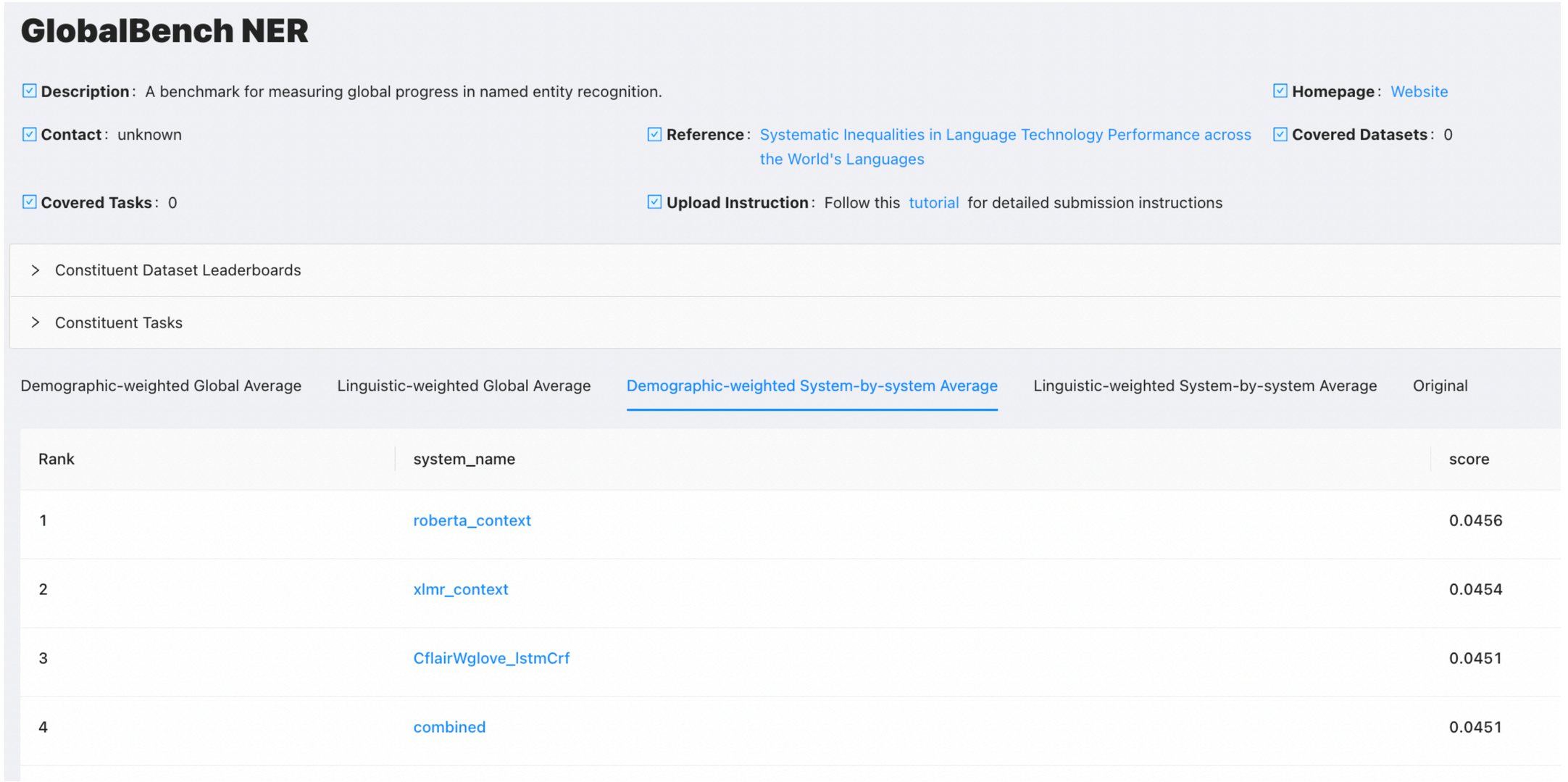
While GDP and number of language users are correlated,

GDP predicts better the total number of papers published on any specific language

How Can We Change our Incentives?

- **GlobalBench:** a benchmark to measure research community progress on equitable language technology

  https://explainaboard.inspiredco.ai/benchmark?id=globalbench



GlobalBench NER

Description: A benchmark for measuring global progress in named entity recognition.

Homepage: Website

Contact: unknown

Reference: Systematic Inequalities in Language Technology Performance across the World's Languages

Covered Datasets: 0

Covered Tasks: 0

Upload Instruction: Follow this tutorial for detailed submission instructions

> Constituent Dataset Leaderboards

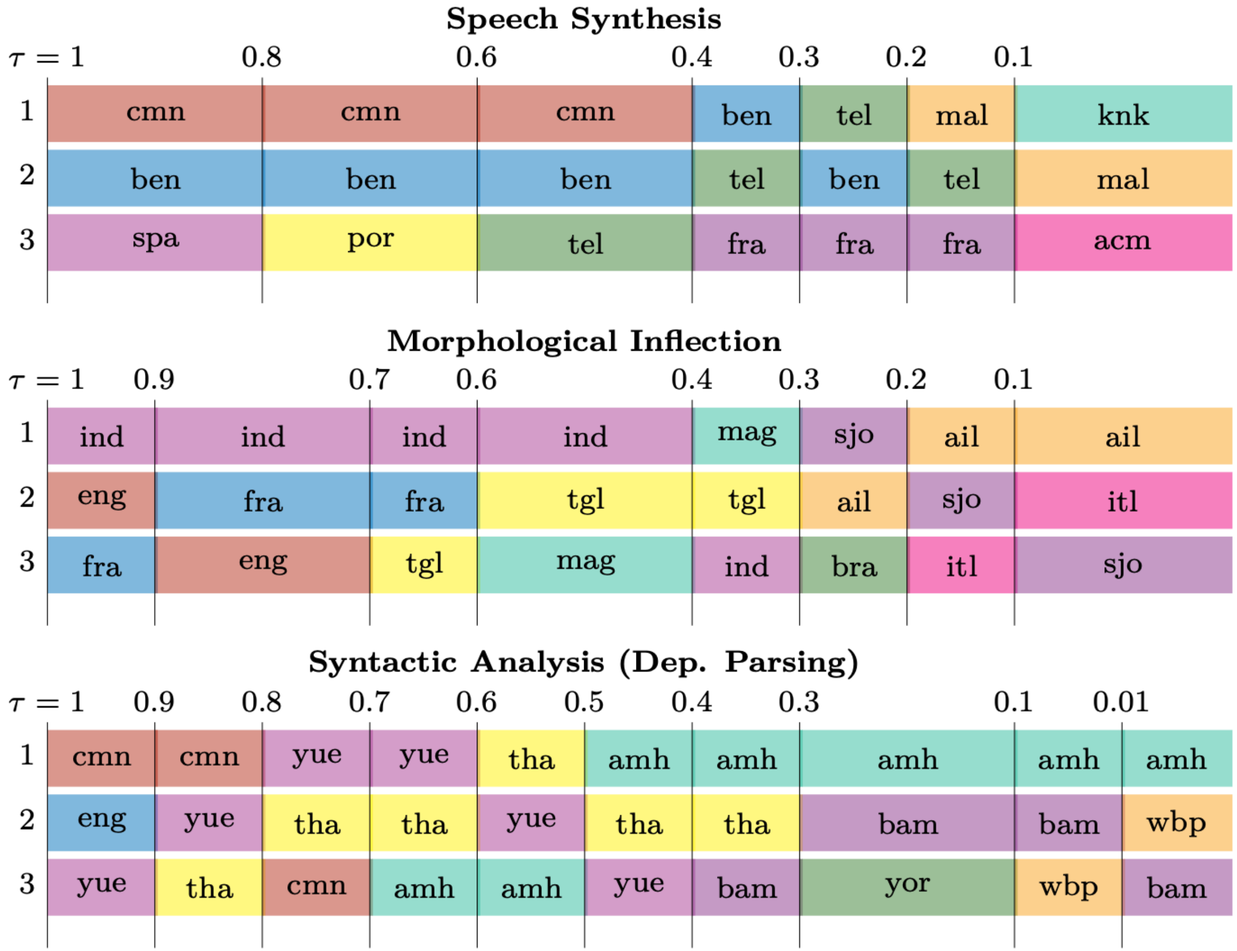
> Constituent Tasks

Demographic-weighted Global Average | Linguistic-weighted Global Average | Demographic-weighted System-by-system Average | Linguistic-weighted System-by-system Average | Original

Rank	system_name	score
1	roberta_context	0.0456
2	xlmr_context	0.0454
3	CflairWglove_lstmCrf	0.0451
4	combined	0.0451

GlobalBench May Help By

- Identifying which languages are *not* covered yet, or under-performing
- Rewarding submissions of *system results* for under-served languages
- Rewarding submissions of *datasets* for under-served languages



We Want You!

- Please submit datasets!
- Please submit systems!
- Please propose tasks!
- Please give us other ideas!

https://www.github.com/neulab/globalbench