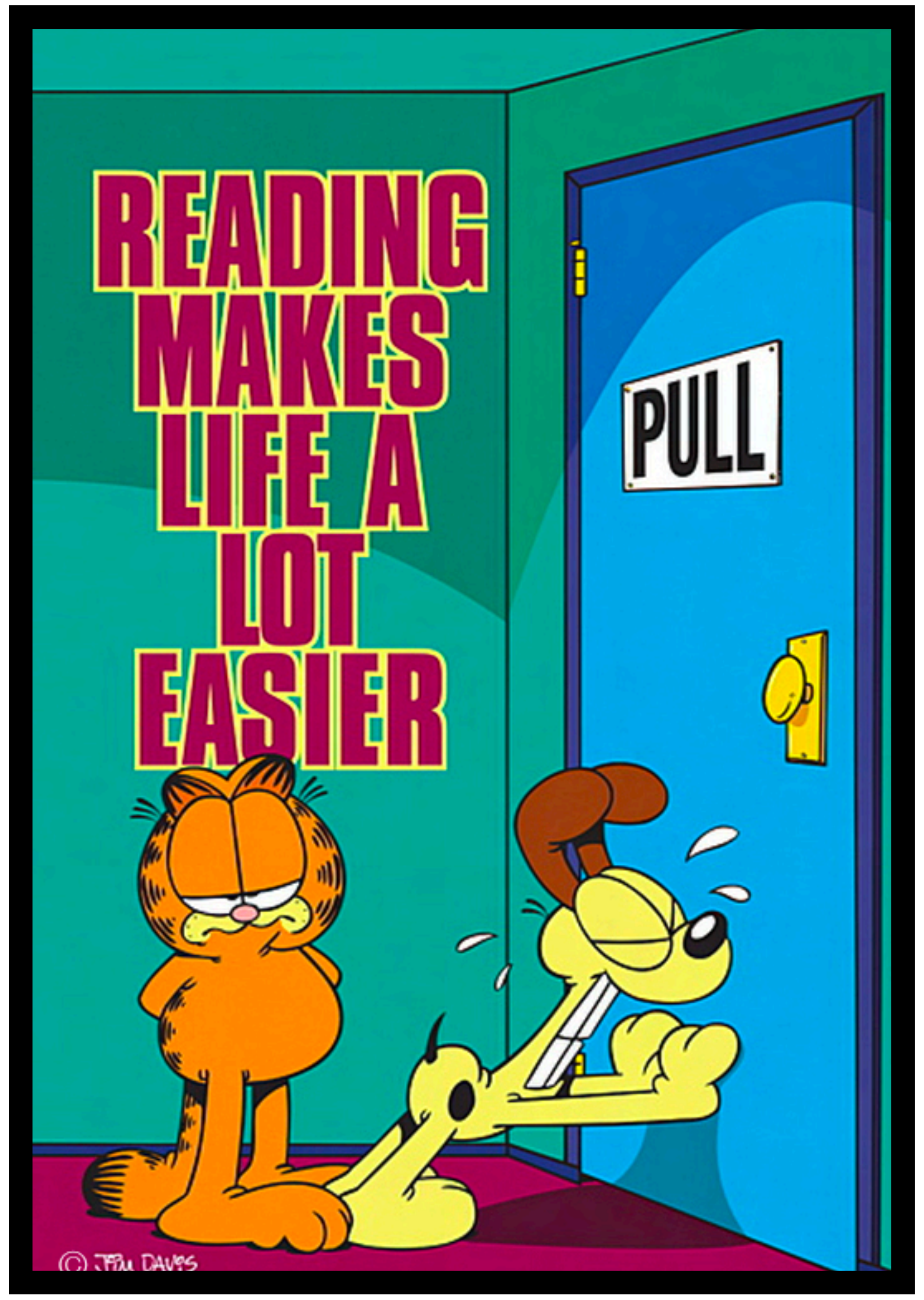


(Image Source: Garfield)

Enhancing Multilingual Capabilities in Large Language Models

Wei Xu (associate professor)
College of Computing
Georgia Institute of Technology
Twitter/X @cocoweixu

NLP X Research Lab

Georgia Tech School of Interactive Computing College of Computing

Generative AI

- generation evaluation
- reading/writing/voice assistant
- human-AI interactive system
- stylistics

Language Models

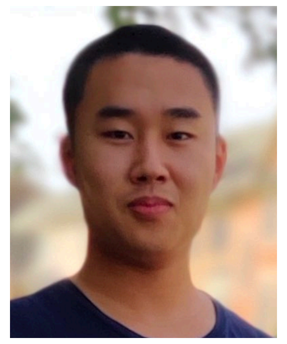
- multi-/cross-lingual capability
- cultural adaptation
- decoding
- privacy, safety

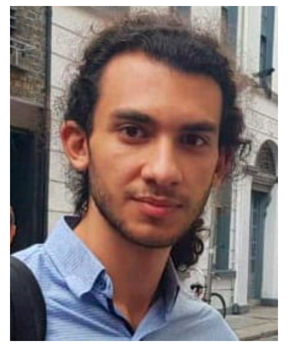
NLP+X Interdisciplinary Research

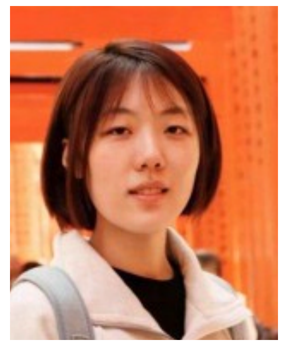
- HCI, human-centered NLP
- Education, Healthcare, Accessibility …

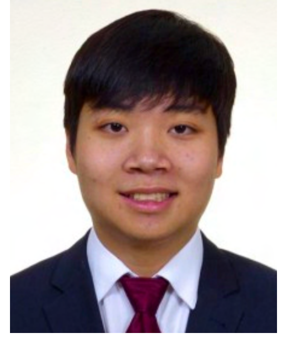
(co-advised with Alan Ritter)


Chao Jiang
PhD student

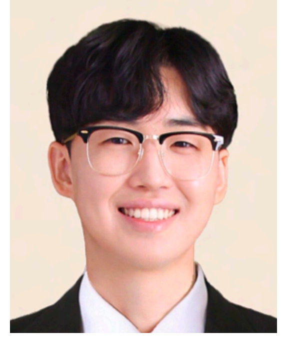

Yao Dou
PhD student


Tarek Naous
PhD student


Geyang Guo
PhD student


Jonathan Zheng
PhD student


Duong Minh Le
PhD student

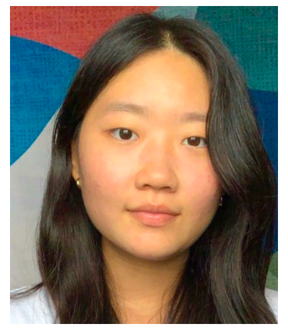

Junmo Kang
PhD student

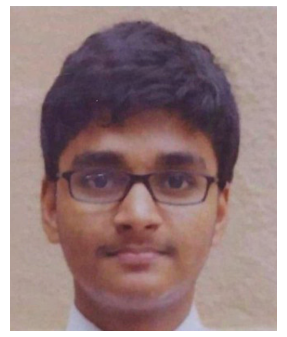

Jeongrok Yu
MS student

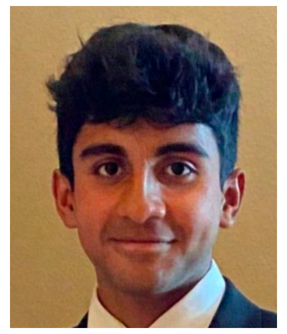

Anton Lavrouk
MS student


Xiaofeng Wu
MS student

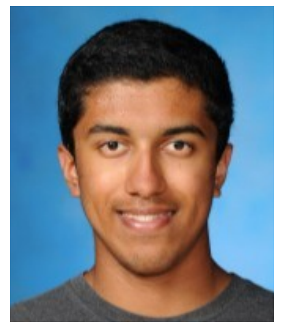

Oleksandr Lavreniuk
Undergrad

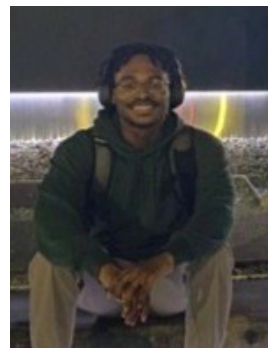

Rachel Choi
Undergrad


Vishnesh Ramanathan
Undergrad

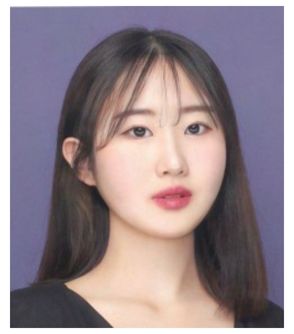

Govind Ramesh
Undergrad


Ian Ligon
Undergrad


Joseph Thomas
Undergrad


Julius Broomfield
Undergrad


Nour Allah El Senary
Undergrad


Siwan Yang
Undergrad


Suraj Mehrotra
Undergrad

Today's Talk —

1 - Cross-lingual Transfer Learning

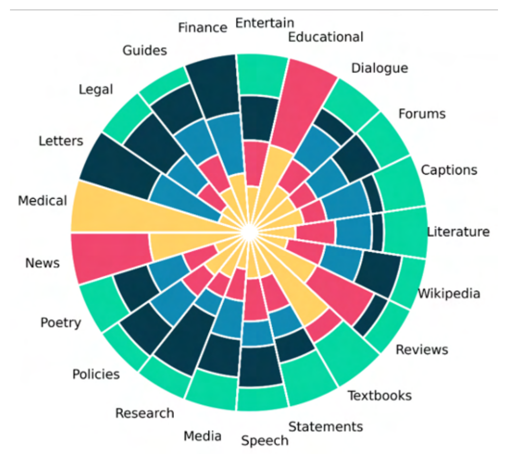
CODEC

(Le et al., ICLR 2024)

Design decoding algorithms to improve performance on non-English languages.

2 - Multilingual Multi-domain Datasets

ReadMe++ & MedReadMe



(Naous et al., EMNLP 2024 & Chao et al., EMNLP 2024)

Support not only more languages but also more text domains/genres.

Today's Talk —

1 - Cross-lingual Transfer Learning

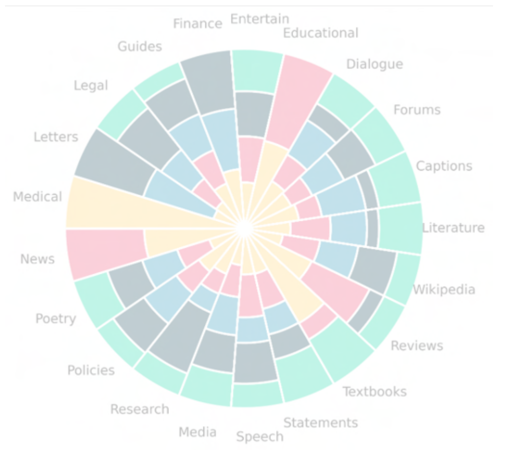
CODEC

(Le et al., ICLR 2024)

Design decoding algorithms to improve performance on non-English languages.

2 - Multilingual Multi-domain Datasets

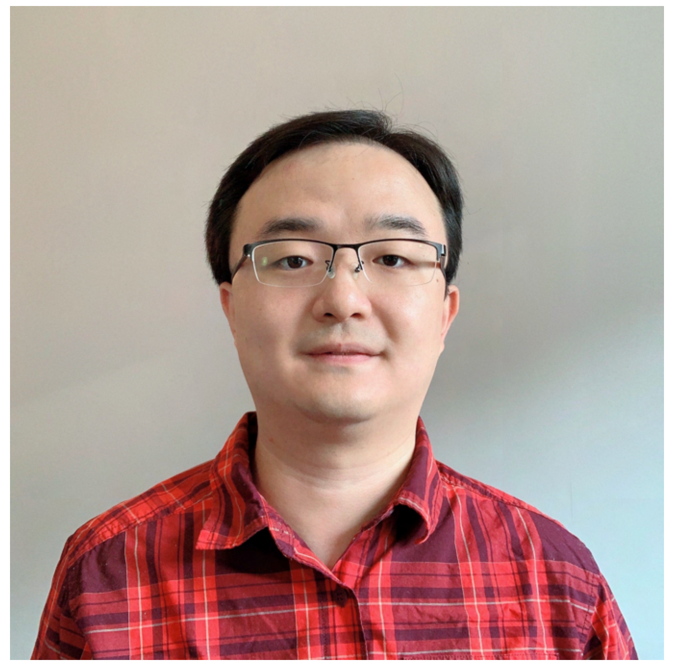
ReadMe++ & MedReadMe



(Naous et al., EMNLP 2024 & Chao et al., EMNLP 2024)

Support not only more languages but also more text domains/genres.

Frustratingly Easy Label Projection for Cross-lingual Transfer (EasyProject)

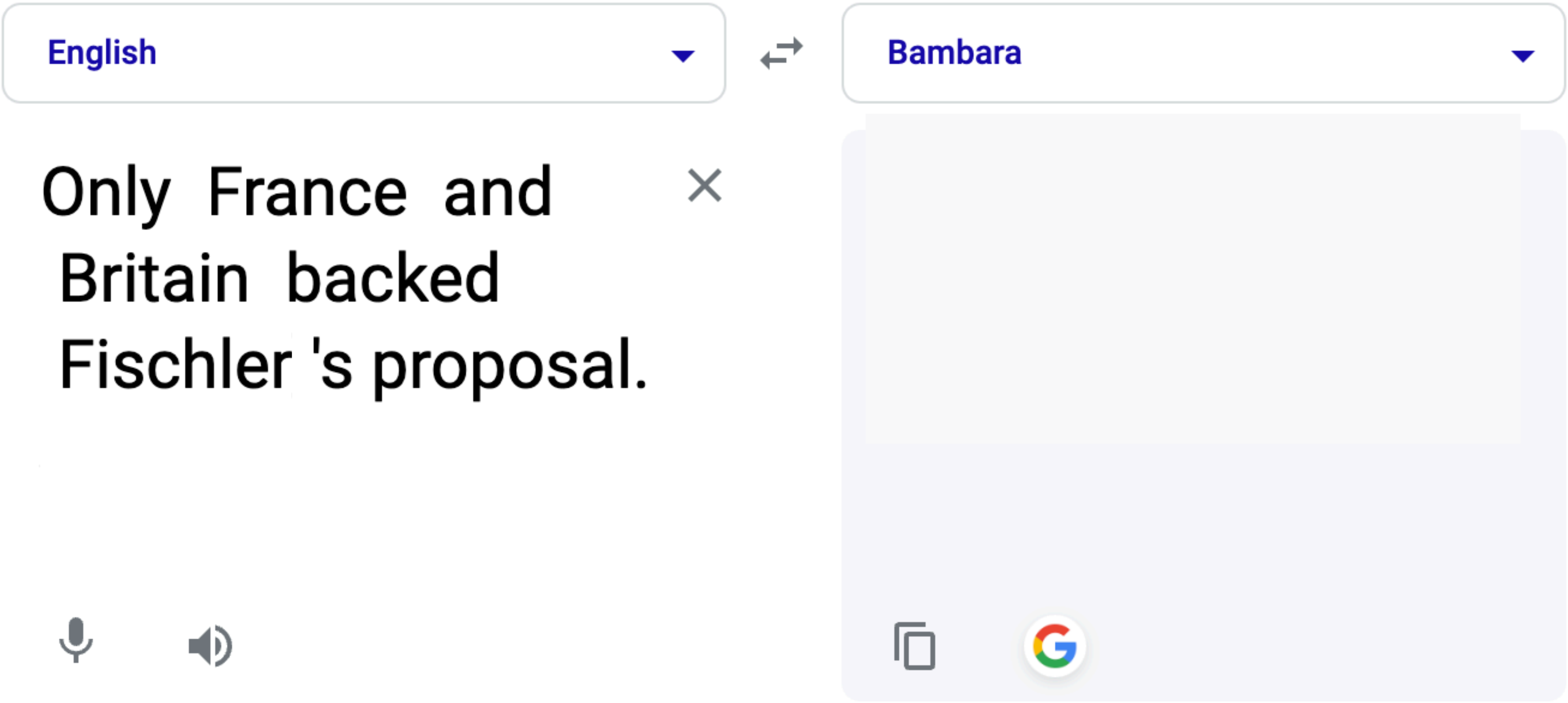



Yang Chen   Chao Jiang   Alan Ritter   Wei Xu

A systematic study of marker-based approach for label projection

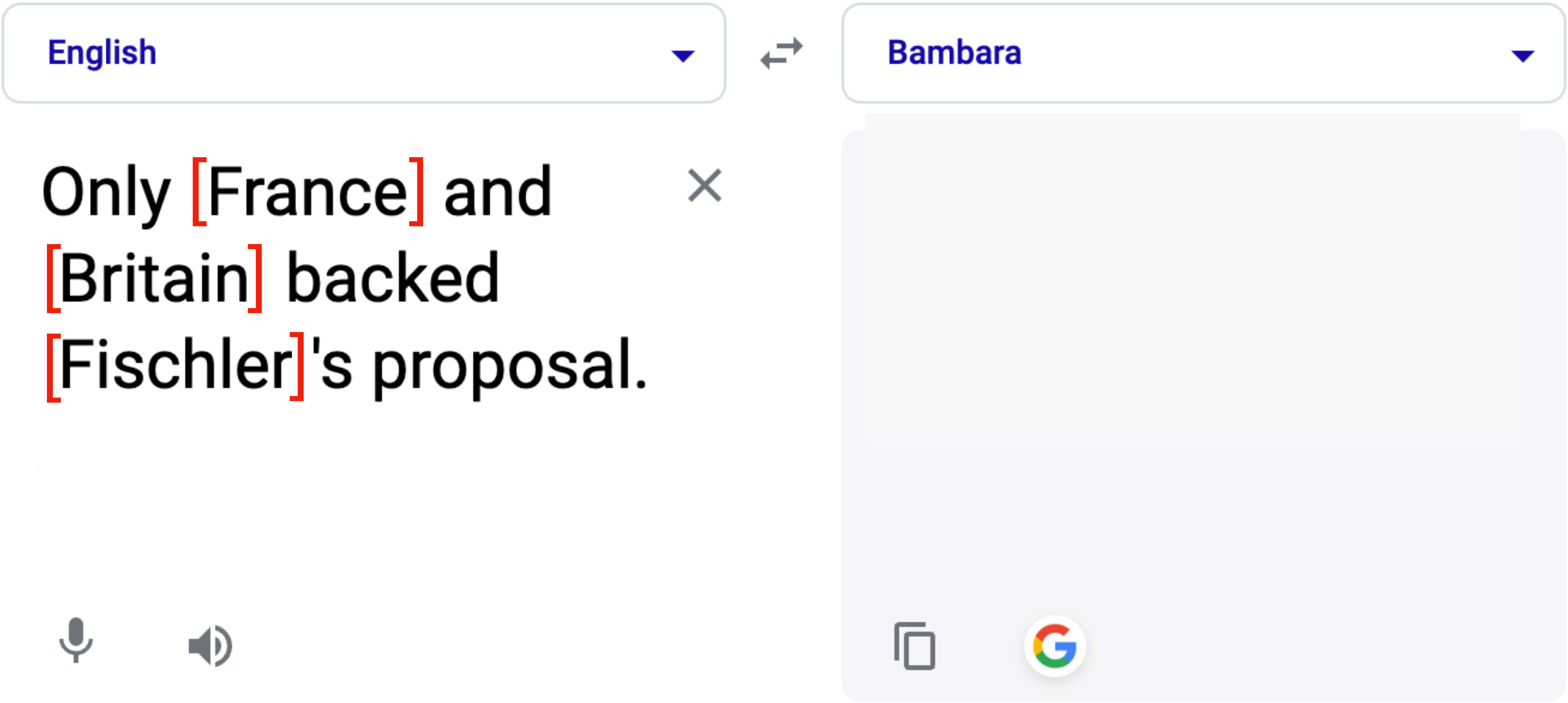
Marker-based Approach

Translating annotated training data from one language to the other



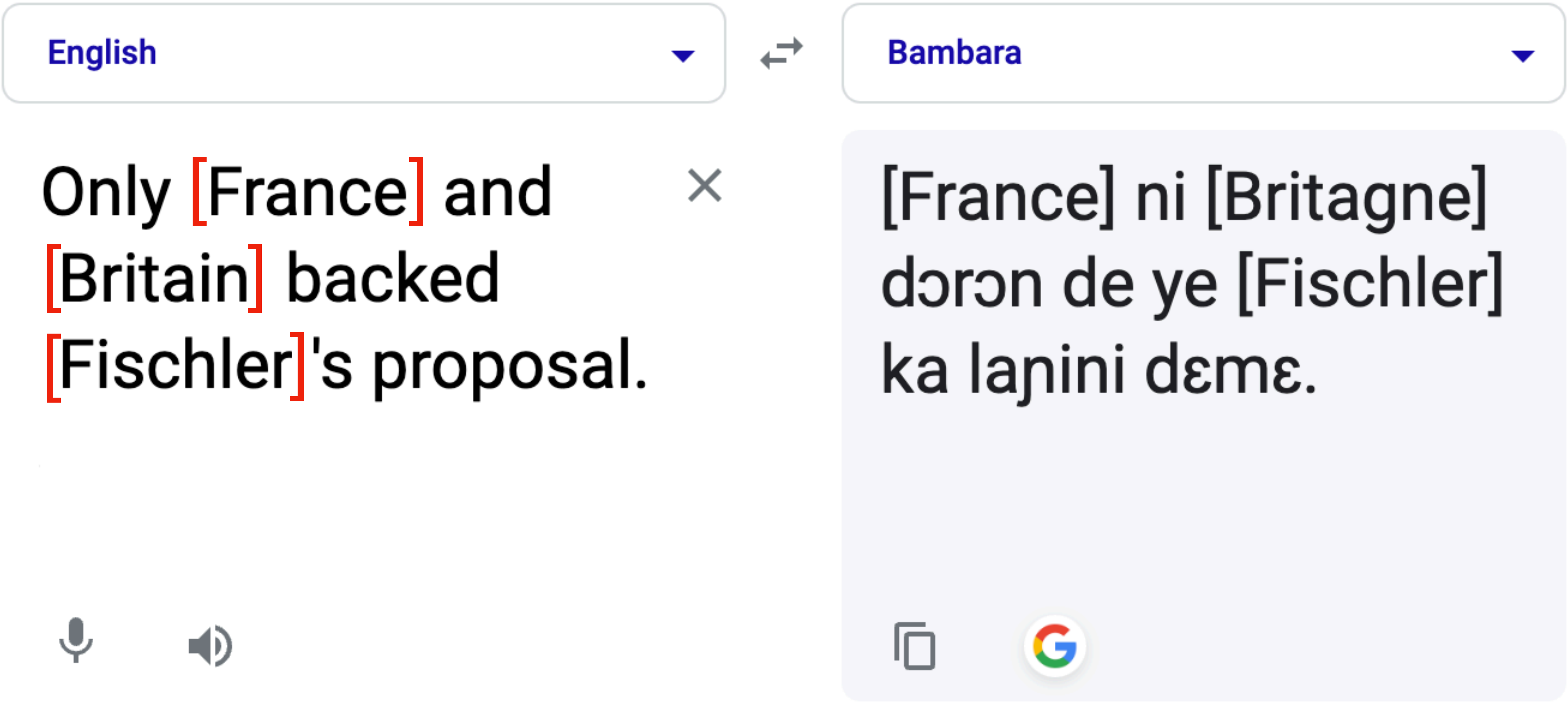
Marker-based Approach

Translating annotated training data from one language to the other by injecting some markers [ ] around the text spans



Marker-based Approach

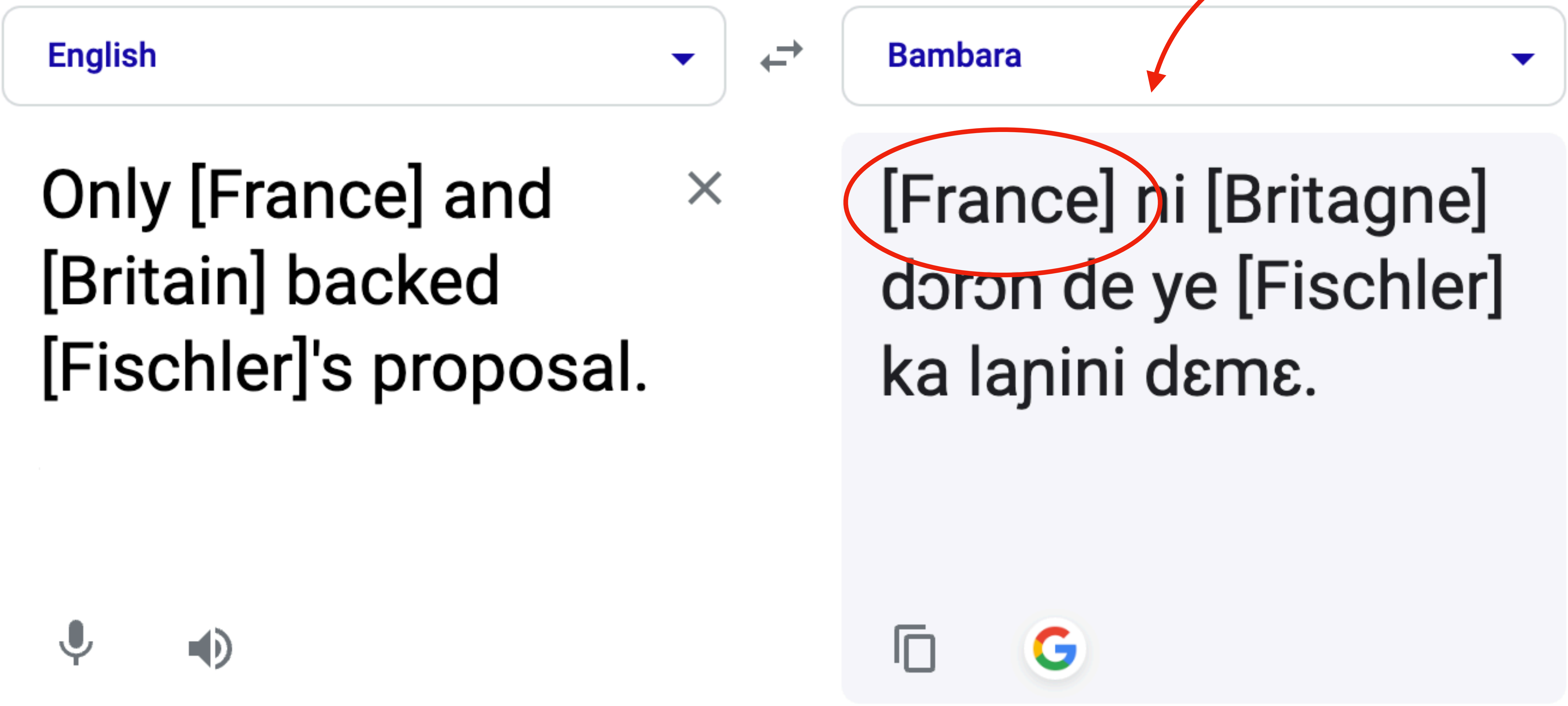
Translating annotated training data from one language to the other by injecting some markers [ ] around the text spans, then sending it directly to a Machine Translation system.



Marker-based Approach

Translating annotated training data from one language to the other by injecting some markers [ ] around the text spans, then sending it directly to a Machine Translation system.

though not without caveat
(will talk more later)



Marker-based Approach

- used by researchers “informally” as a hack
- one of the earliest such accounts is by Lee et al. (2018)
- then, used in MLQA (Lewis et al., 2020), XTREME (Hu et al., 2020) …

- But, only described briefly in each paper
- How well does it work? For different languages, tasks? Better or worse than word alignment?

EasyProject - Easy Marker-based Projection

- Different markers all work to some extents, but vary for languages:

works the best

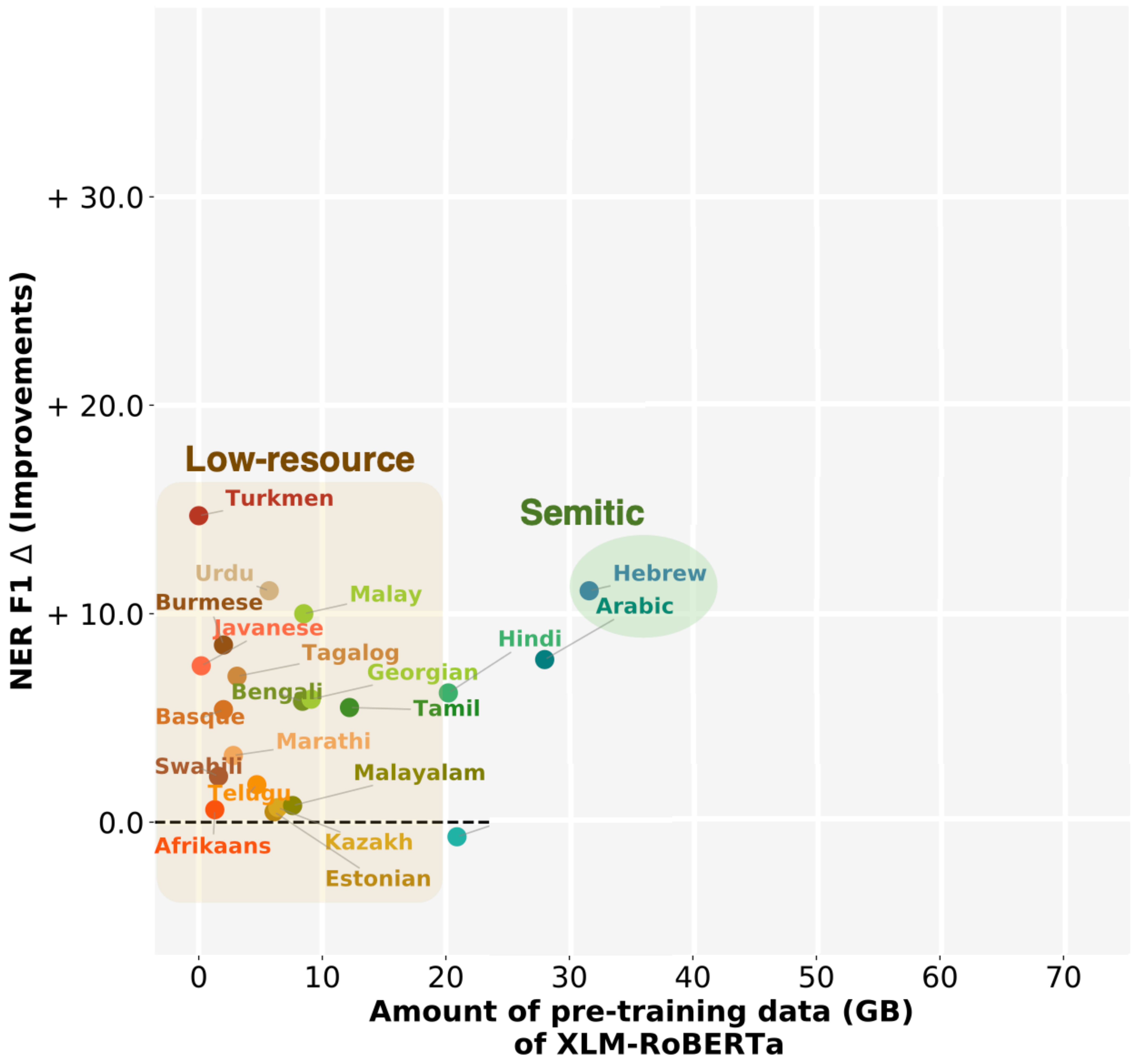
XML tags (e.g., <loc> </loc>) or [ ] “ ” ( ) < > { }

- If >1 spans to be projected in one sentence, do need to map the tags by fuzzy string matching

- Further fine-tuning MT system on synthetic data to make it more robust with punctuations

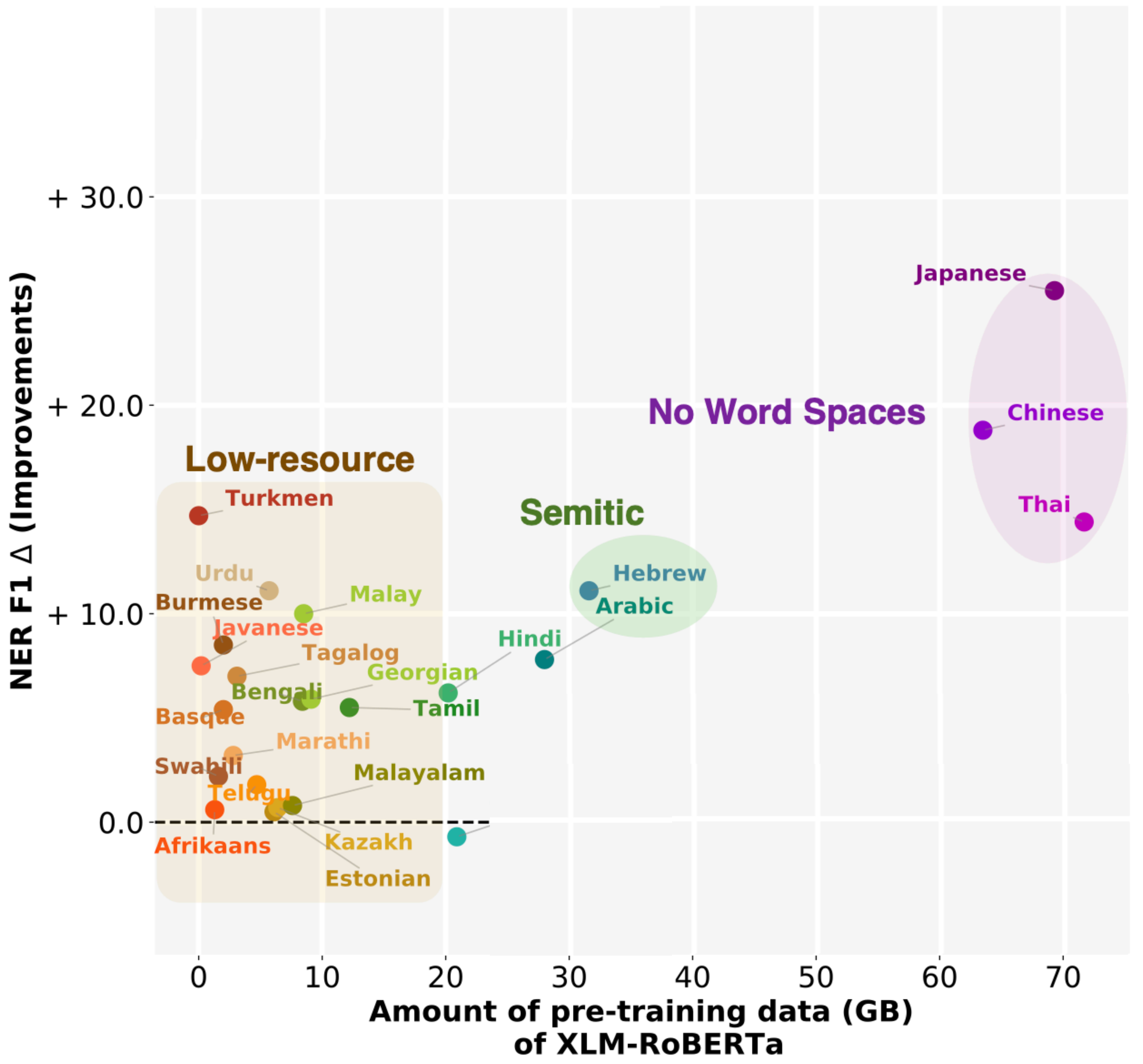
EasyProject - Easy Marker-based Projection

Especially promising for low-resource languages & languages that are written in non-Latin scripts



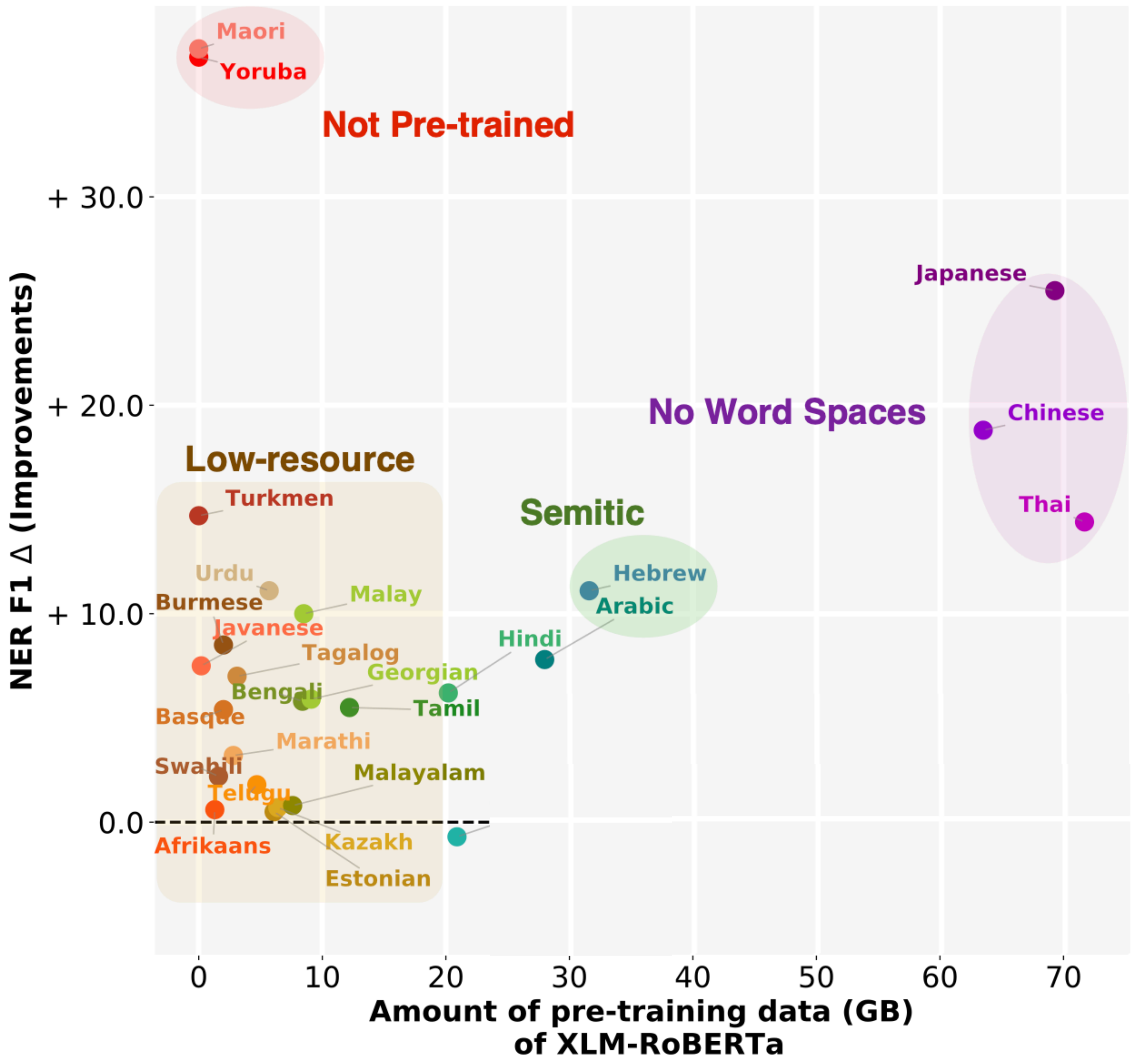
EasyProject - Easy Marker-based Projection

Especially promising for low-resource languages & languages that are written in non-Latin scripts



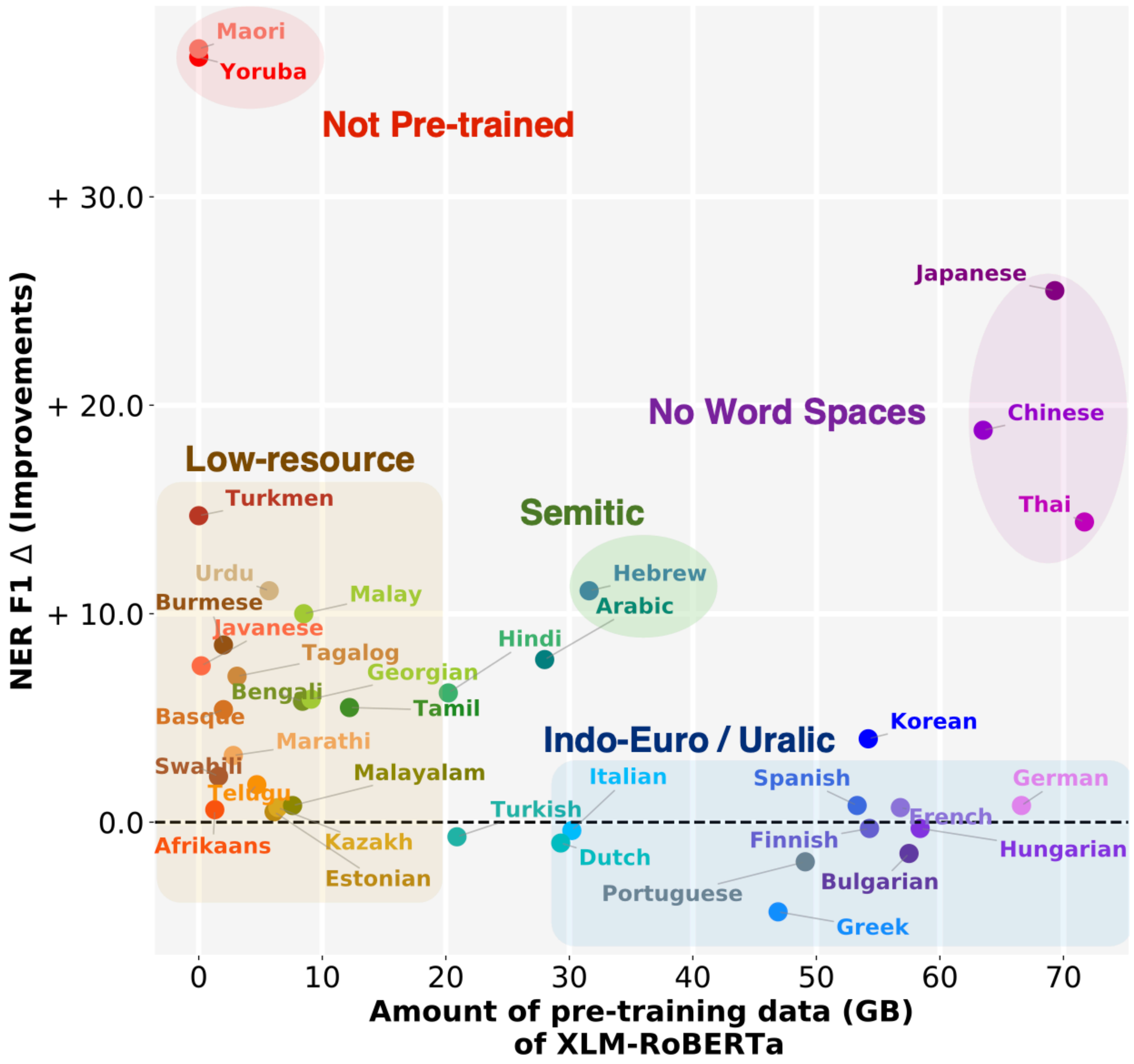
EasyProject - Easy Marker-based Projection

Especially promising for low-resource languages & languages that are written in non-Latin scripts



EasyProject - Easy Marker-based Projection

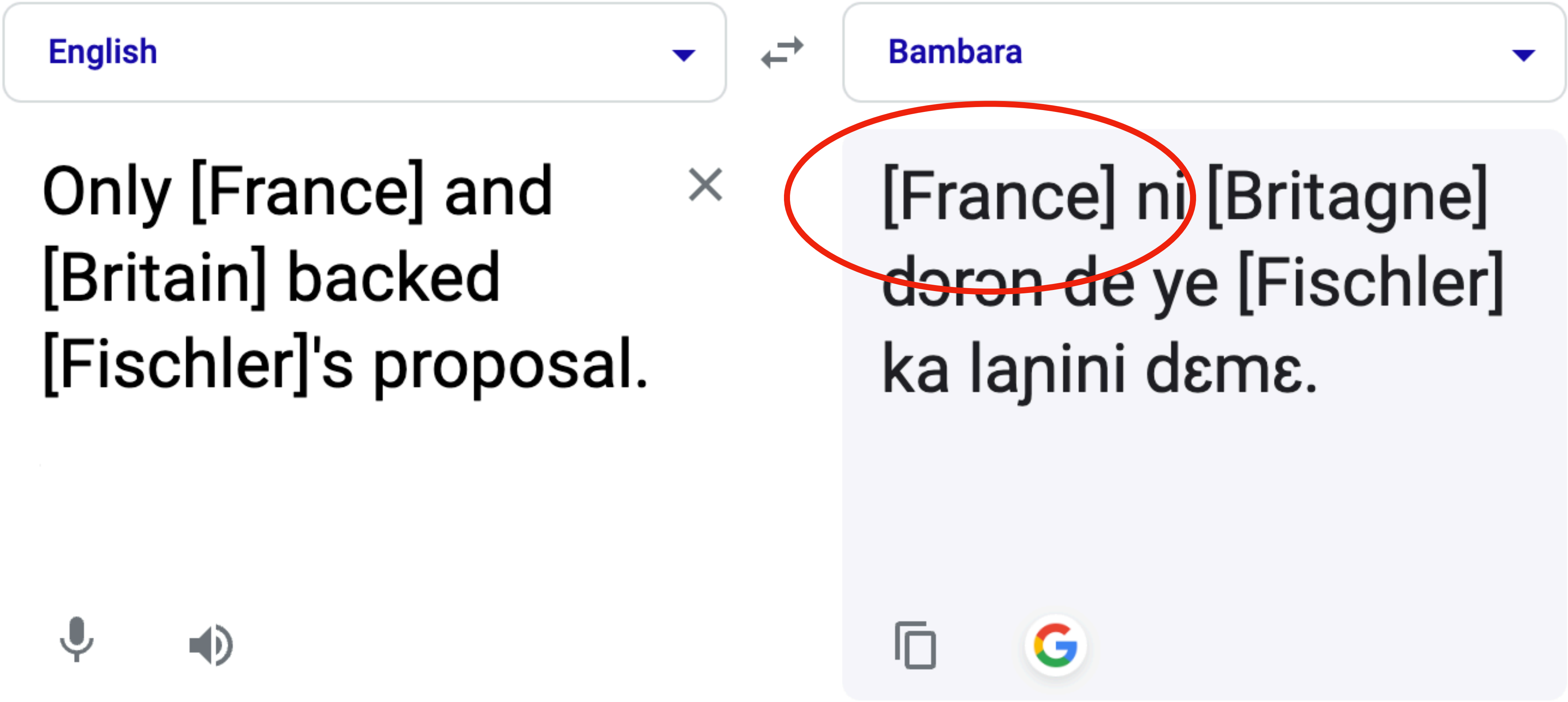
Especially promising for low-resource languages & languages that are written in non-Latin scripts



Zero-shot Cross-lingual Label Projection

Two families of approaches, but each has pros and cons.

marker-based approach



Only need a MT system
&
work surprisingly well !

But, degraded
MT quality
due to injected markers

Zero-shot Cross-lingual Label Projection

Two families of approaches, but each has pros and cons.

marker-based approach

English | Bambara

Only [France] and [Britain] backed [Fischler]'s proposal.

[France] ni [Britagne] dɔrɔn de ye [Fischler] ka laɲini dɛmɛ.

Only need a MT system & work surprisingly well !

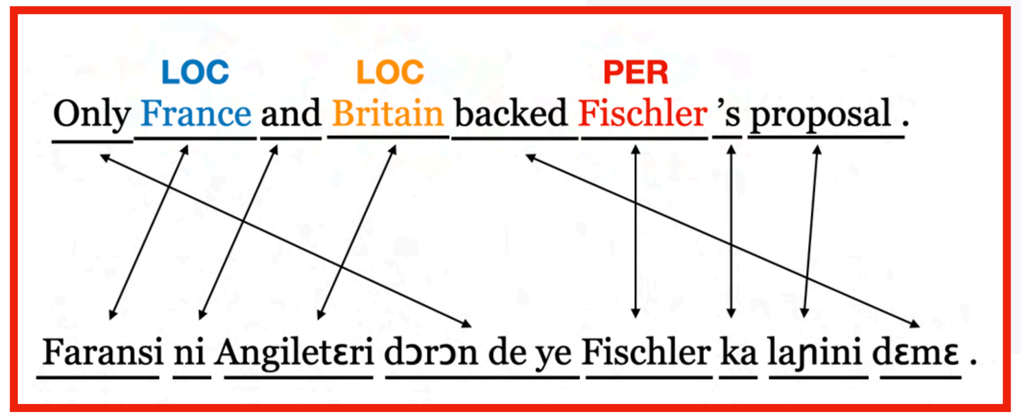
But, degraded MT quality due to injected markers

word alignment-based approach

English | Bambara

Only France and Britain backed Fischler 's proposal .

Faransi ni Angletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ .

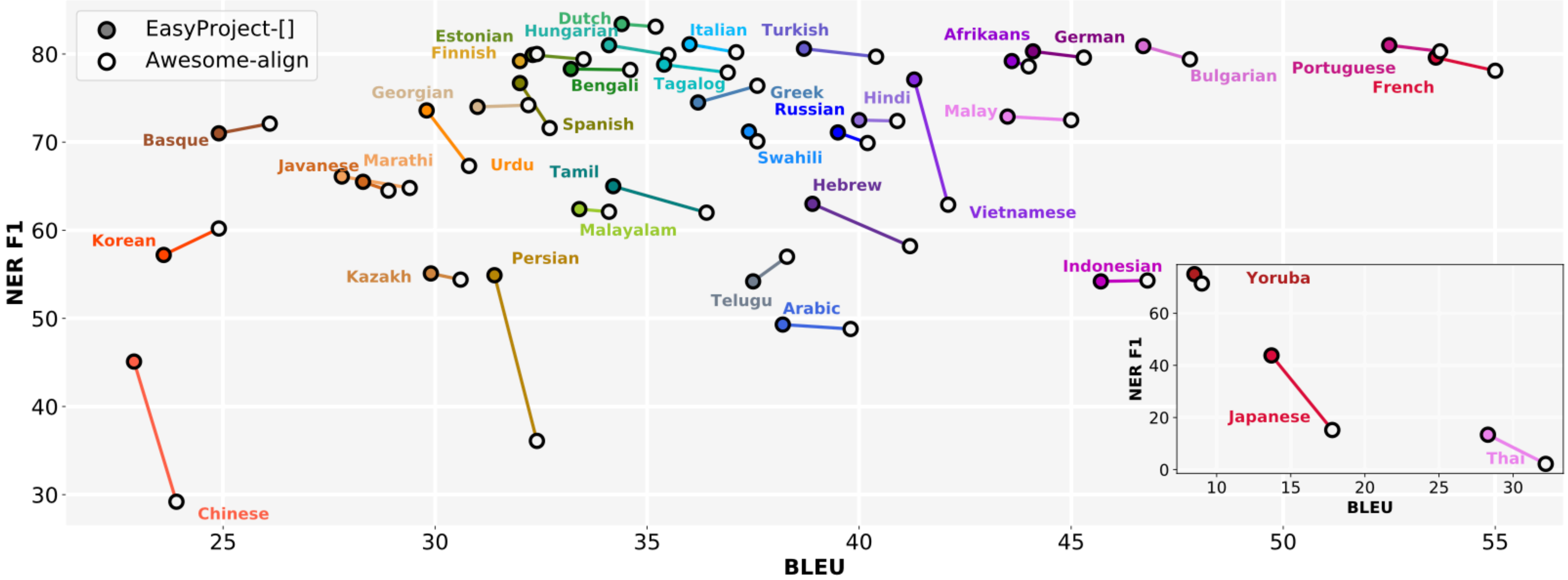


normally better MT quality

Require not only neural MT, but also a separate word alignment model

EasyProject - Easy Marker-based Projection

Despite degraded MT quality, marker-based approach still works surprisingly well for the end task!



Yang Chen, Chao Jiang, Alan Ritter, Wei Xu. “Frustratingly Easy Label Projection for Cross-lingual Transfer” (ACL 2023 Findings)

Can we do marker-based approach without scarifying the translation quality?

Constrained Decoding for Cross-lingual Label Projection (CODEC)

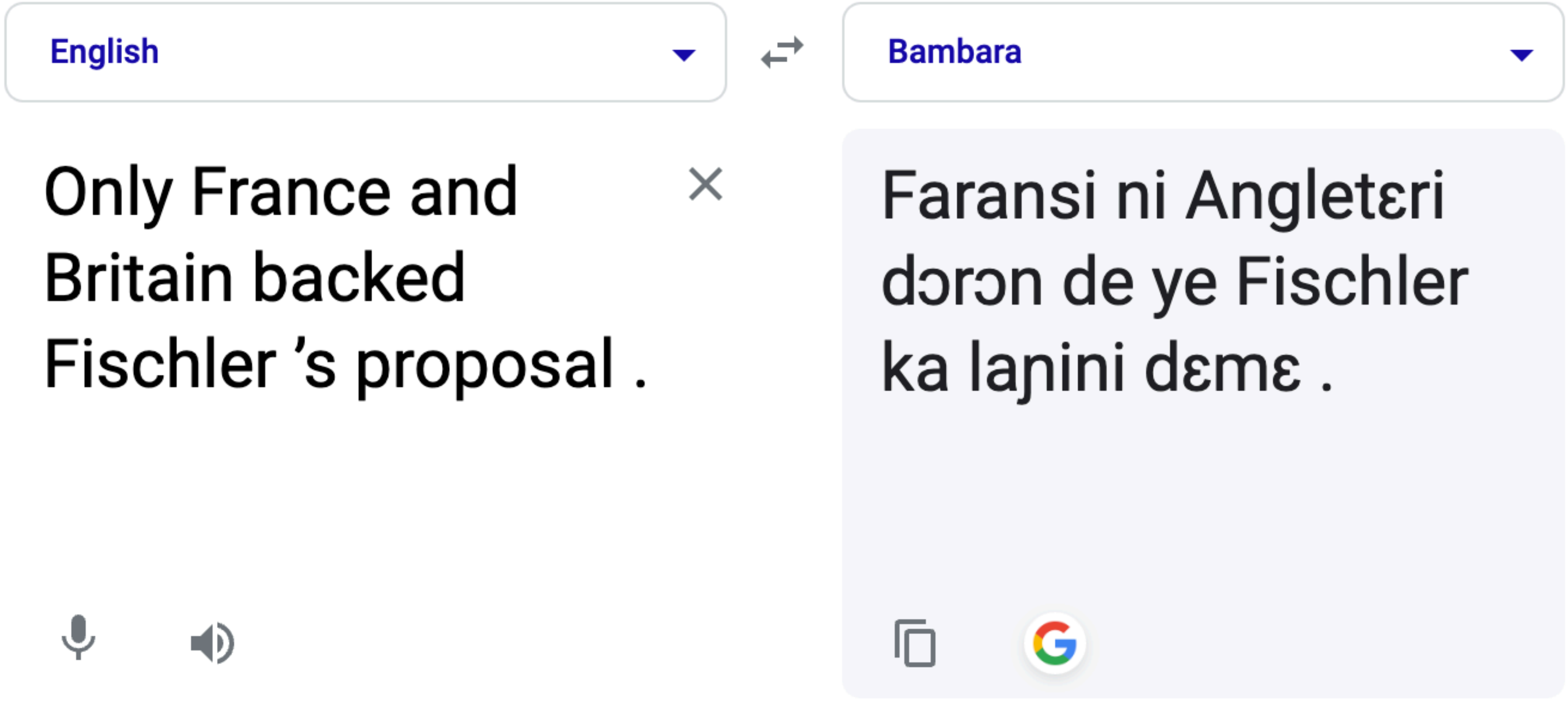



Duong Minh Le Yang Chen Alan Ritter Wei Xu

A better technical solution for marker-based label projection

Key Idea

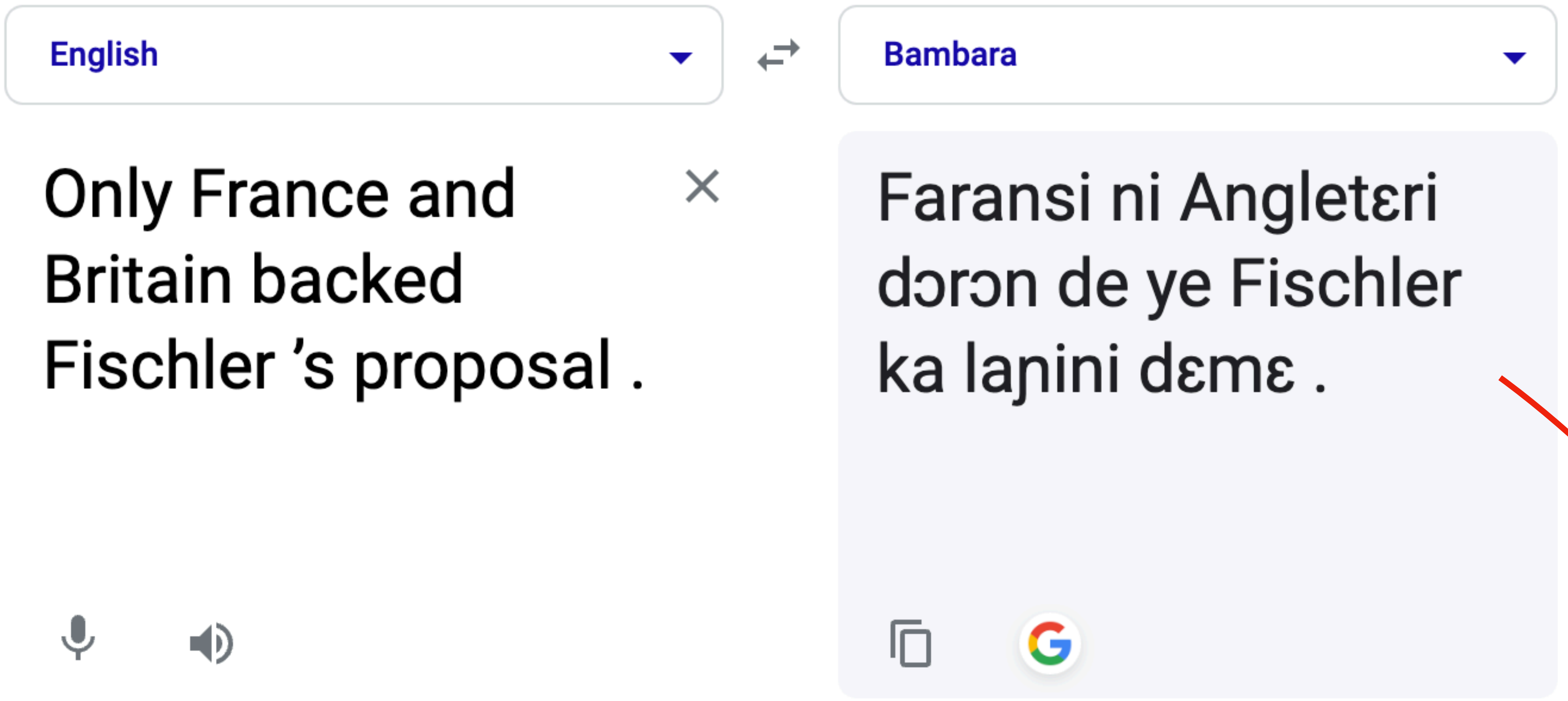
Step 1. Translate the original sentence as usual without markers.



Step 2. Run translation model for a 2nd time to insert markers as a constrained decoding problem.

Key Idea

Step 1. Translate the original sentence as usual without markers.



Impose two constraints:
(1) keeping the same translation
(2) having the correct number of [ ] s

Step 2. Run translation model for a 2nd time to insert markers as a constrained decoding problem.

Input sentece:

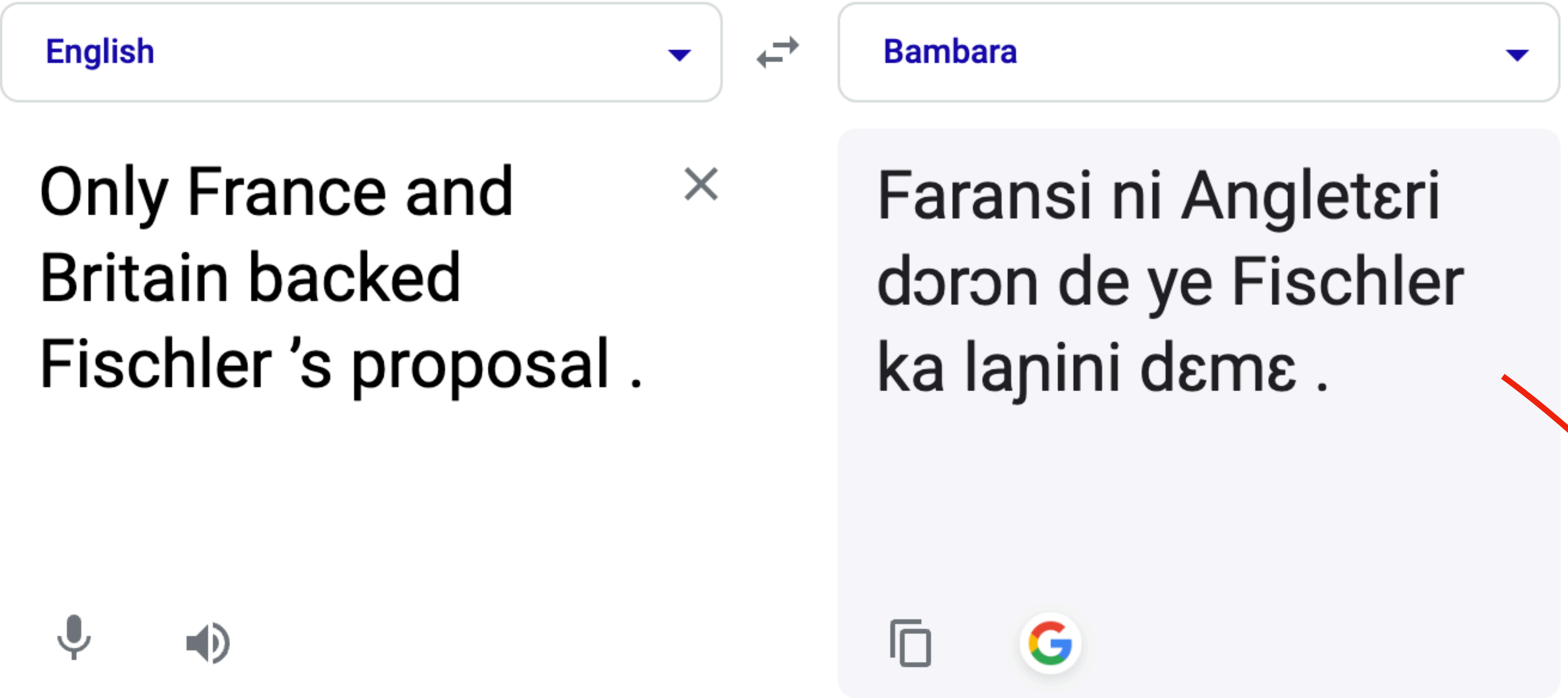
Only [France] and [Britain] backed [Fischler]'s proposal.

→ **LLMs (Translation Models)** →

Translated Output:

Key Idea

Step 1. Translate the original sentence as usual without markers.



Impose two constraints:
(1) keeping the same translation
(2) having the correct number of [ ] s

Step 2. Run translation model for a 2nd time to insert markers as a constrained decoding problem.

Input sentece:

Only [France] and [Britain] backed [Fischler]'s proposal. → **LLMs (Translation Models)** →

Translated Output:

[Faransi] ni [Angiletɛri] dɔrɔn de ye [Fischler] ka laɲini dɛmɛ .

Key Idea — more formally

Step 1. Translate the original sentence as usual without markers.

$$y^{tmpl} = \arg\max_{y} \log P_{\tau}(y|x)$$

Step 2. Run translation model another time to insert m marker pairs [ ] into y^{tmpl} .

$$y^* = \arg\max_{y \in \mathcal{Y}} \log P_{\tau}(y|x^{mark}; y^{tmpl})$$

$\boxed{O(n^{2m})}$

An Efficient Constrained Decoding Algorithm

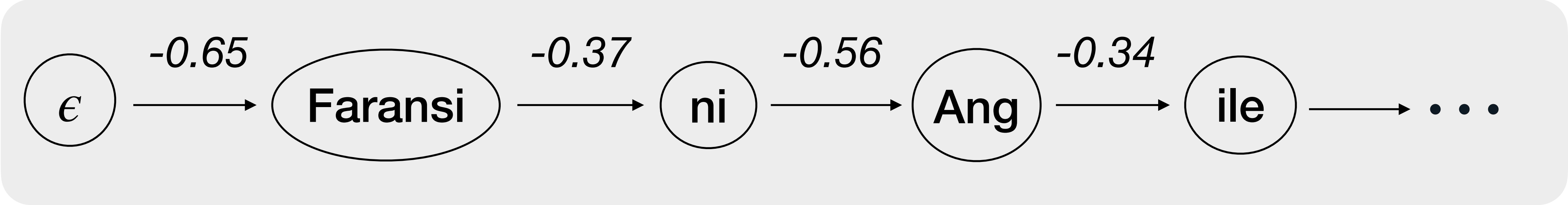
(1) Prune opening marker positions based on the contrastive log-likelihood difference.

An Efficient Constrained Decoding Algorithm

(1) Prune opening marker positions based on the contrastive log-likelihood difference.

Input: $x =$ “Only France and Britain backed Fischler 's proposal .” $x^{mark} =$ “Only France and [Britain] backed Fischler 's proposal .” $y^{tmpl} =$ “Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ .”

$p_1^i = \log P(y_i^{tmpl} | y_{<i}^{tmpl}, x)$ (Conditioned on source text)



$p_2^i = \log P(y_i^{tmpl} | y_{<i}^{tmpl}, x^{mark})$ (Conditioned on source text w/ markers)

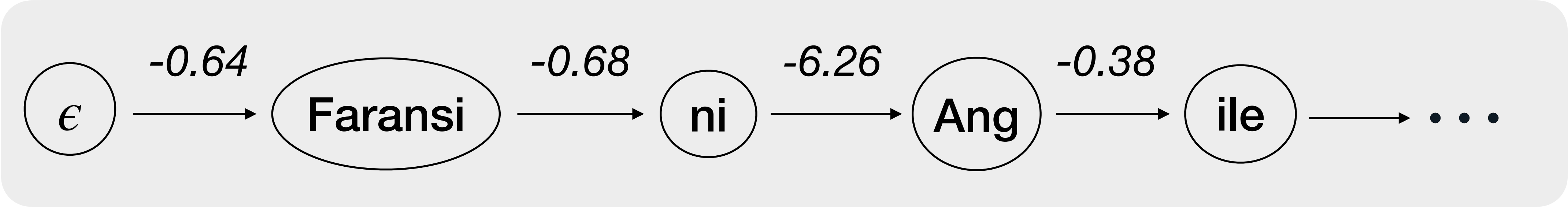
An Efficient Constrained Decoding Algorithm

(1) Prune opening marker positions based on the contrastive log-likelihood difference.

Input: $x =$ "Only France and Britain backed Fischler 's proposal ." $x^{mark} =$ "Only France and [Britain] backed Fischler 's proposal ." $y^{tmpl} =$ "Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ ."

$p_1^i = \log P(y_i^{tmpl} | y_{<i}^{tmpl}, x)$ (Conditioned on source text)

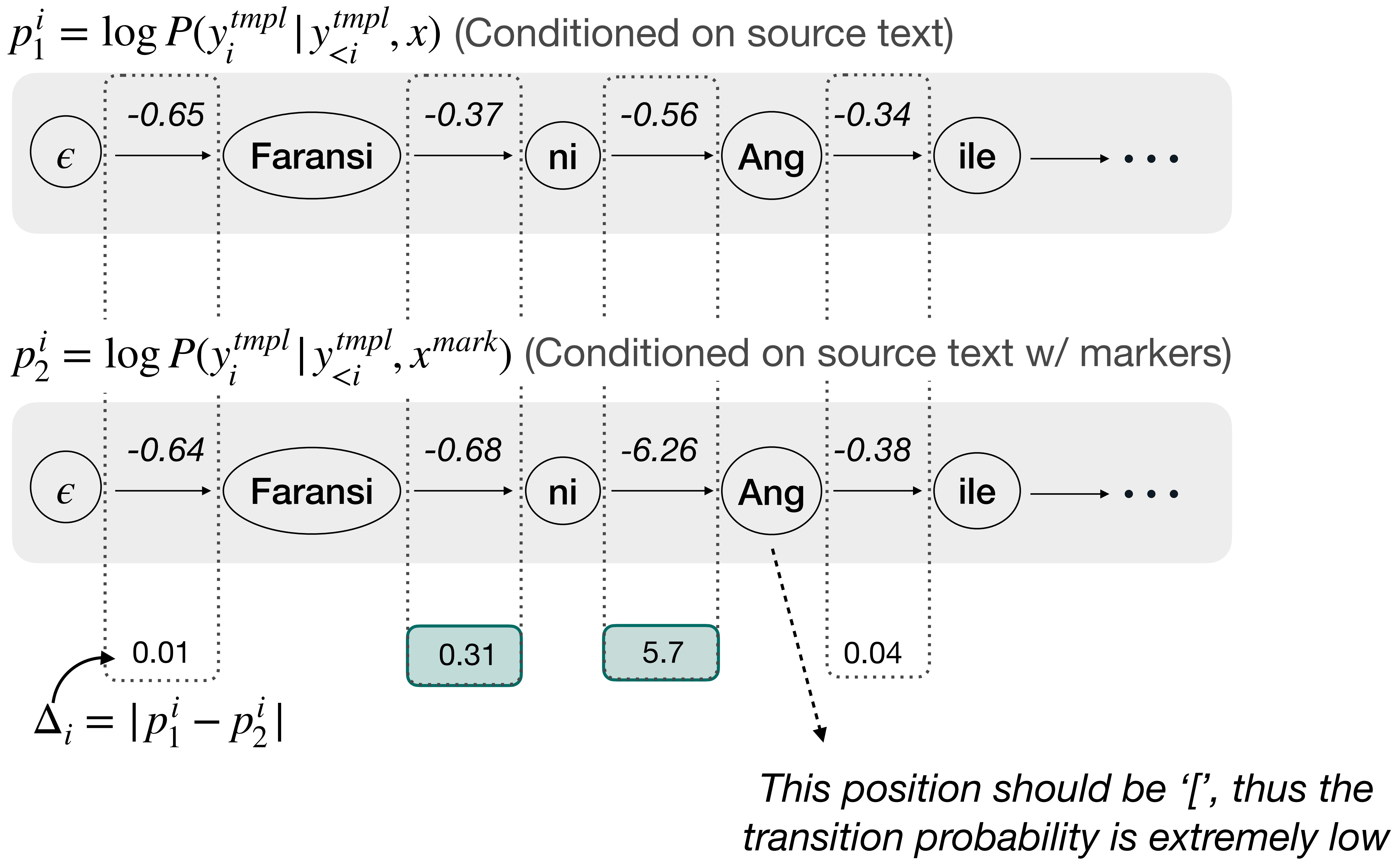
$p_2^i = \log P(y_i^{tmpl} | y_{<i}^{tmpl}, x^{mark})$ (Conditioned on source text w/ markers)



An Efficient Constrained Decoding Algorithm

(1) Prune opening marker positions based on the contrastive log-likelihood difference.

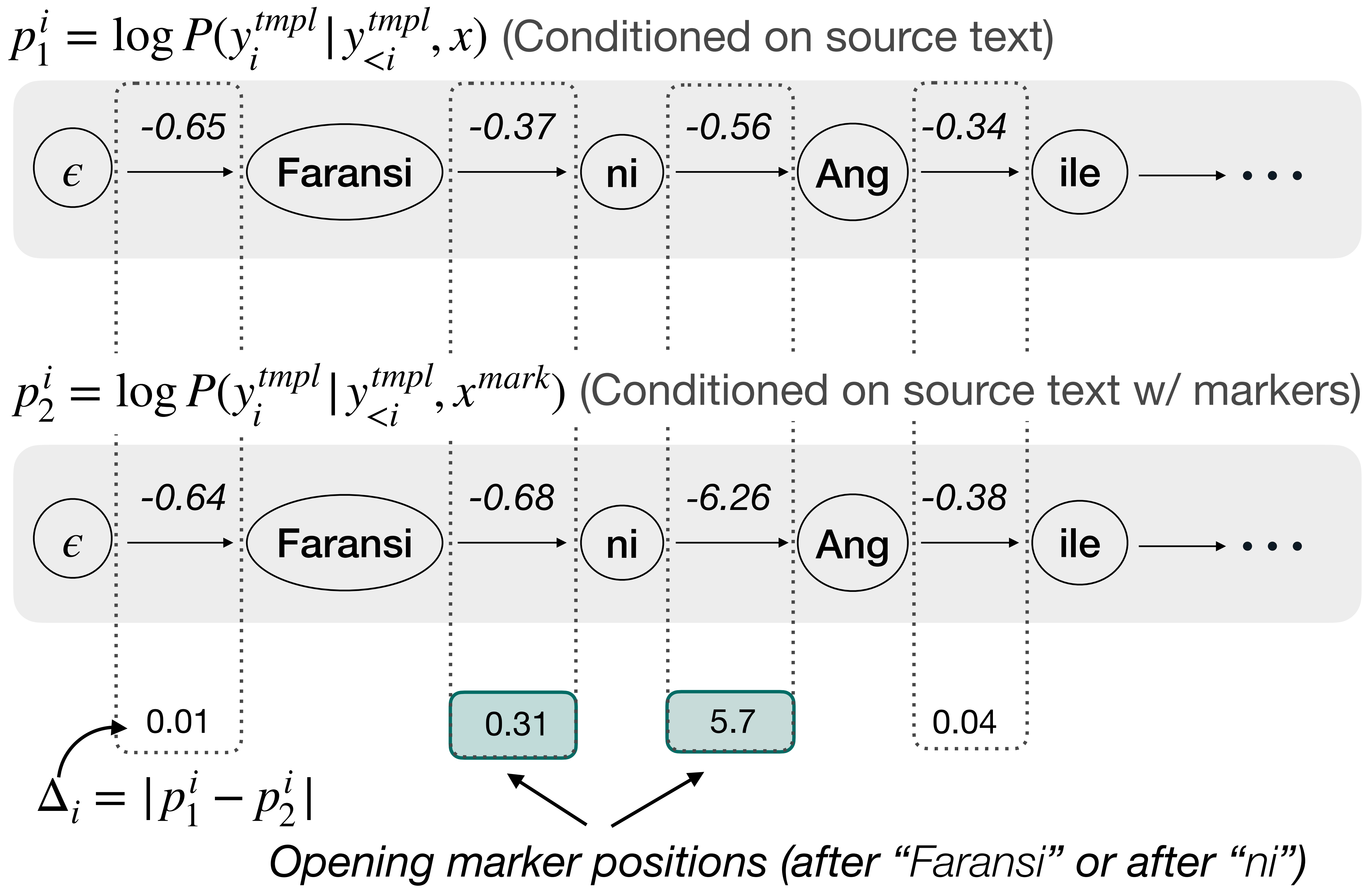
Input: $x =$ "Only France and Britain backed Fischler 's proposal ." $x^{mark} =$ "Only France and [Britain] backed Fischler 's proposal ." $y^{tmpl} =$ "Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ ."



An Efficient Constrained Decoding Algorithm

(1) Prune opening marker positions based on the contrastive log-likelihood difference.

Input: $x =$ "Only France and Britain backed Fischler 's proposal ." $x^{mark} =$ "Only France and [Britain] backed Fischler 's proposal ." $y^{tmpl} =$ "Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ ."



An Efficient Constrained Decoding Algorithm

(2) A branch-and-bound search algorithm with a heuristic lower bound $L_d^k = \log P(y_{1:d}^k | x^{mark})$.

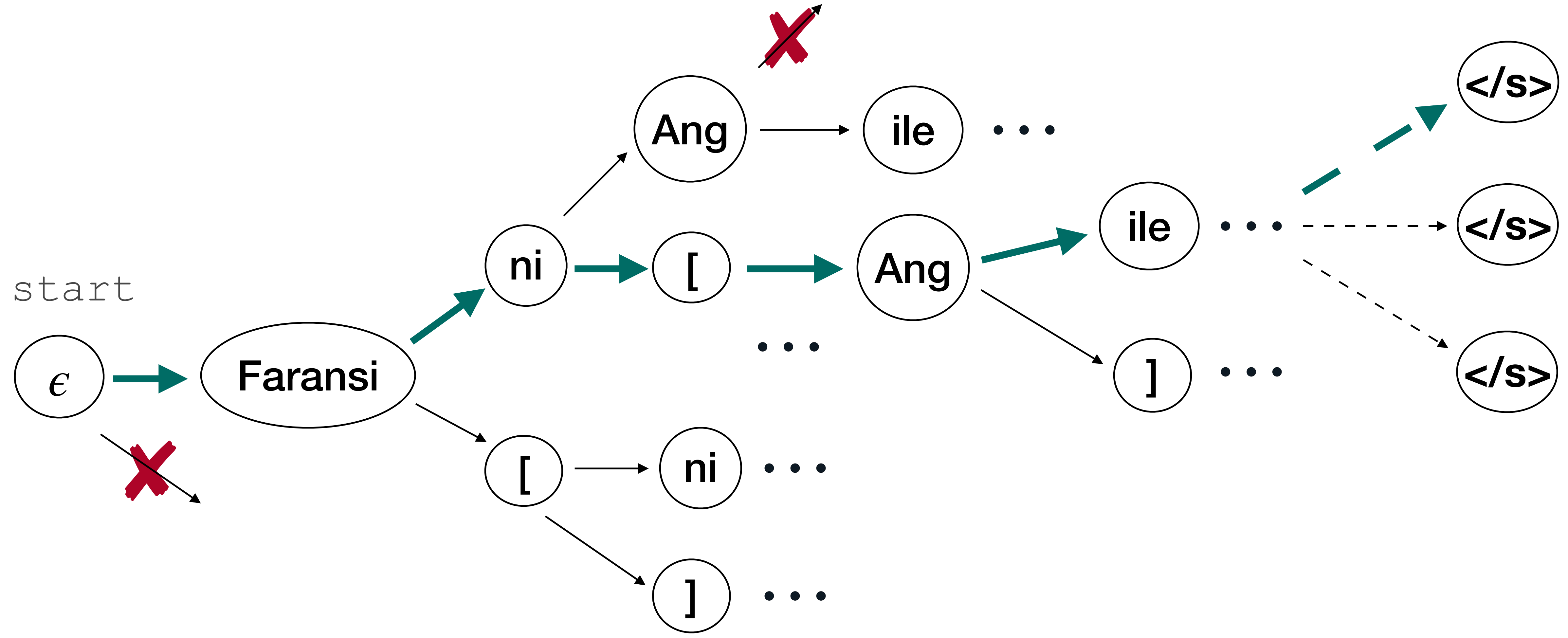
$$d = \min\left(\max\left(j + \delta, q\right), |y^k|\right)$$

An Efficient Constrained Decoding Algorithm

(2) A branch-and-bound search algorithm with a heuristic lower bound $L_d^k = \log P(y_{1:d}^k | x^{mark})$.

$$d = \min\left(\max\left(j + \delta, q\right), |y^k|\right)$$

Input: $x =$ "Only France and Britain backed Fischler 's proposal ." $x^{mark} =$ "Only France and [Britain] backed Fischler 's proposal ." $y^{tmpl} =$ "Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ ."



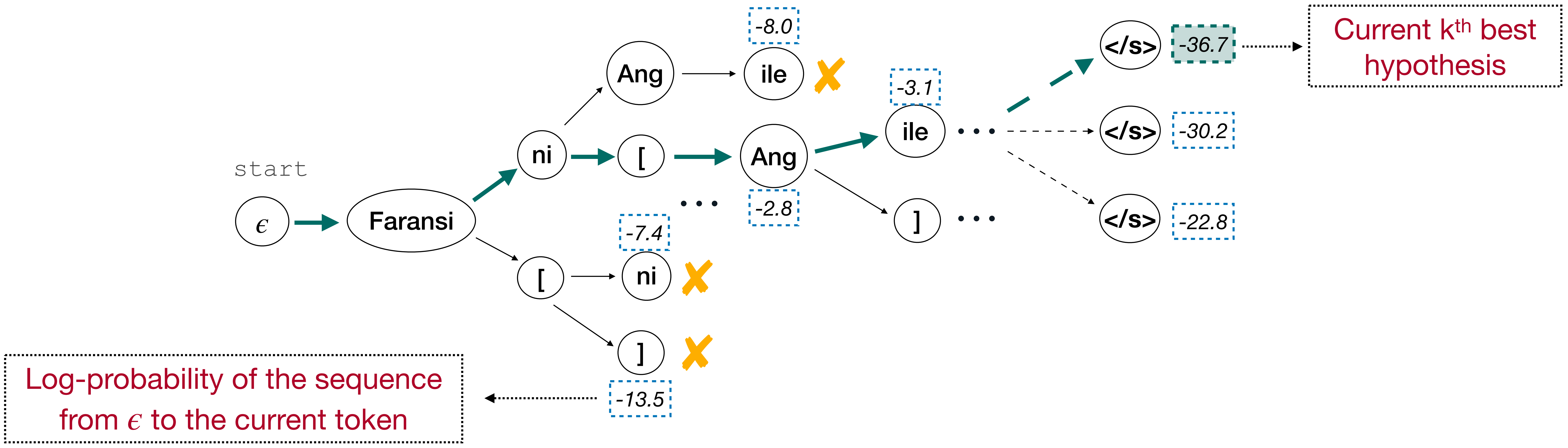
Prune opening-marker positions

An Efficient Constrained Decoding Algorithm

(2) A branch-and-bound search algorithm with a heuristic lower bound $L_d^k = \log P(y_{1:d}^k | x^{mark})$.

$$d = \min\left(\max\left(j + \delta, q\right), |y^k|\right)$$

Input: $x =$ "Only France and Britain backed Fischler 's proposal ." $x^{mark} =$ "Only France and [Britain] backed Fischler 's proposal ." $y^{tmpl} =$ "Faransi ni Angileteri doron de ye Fischler ka laɲini dɛmɛ ."

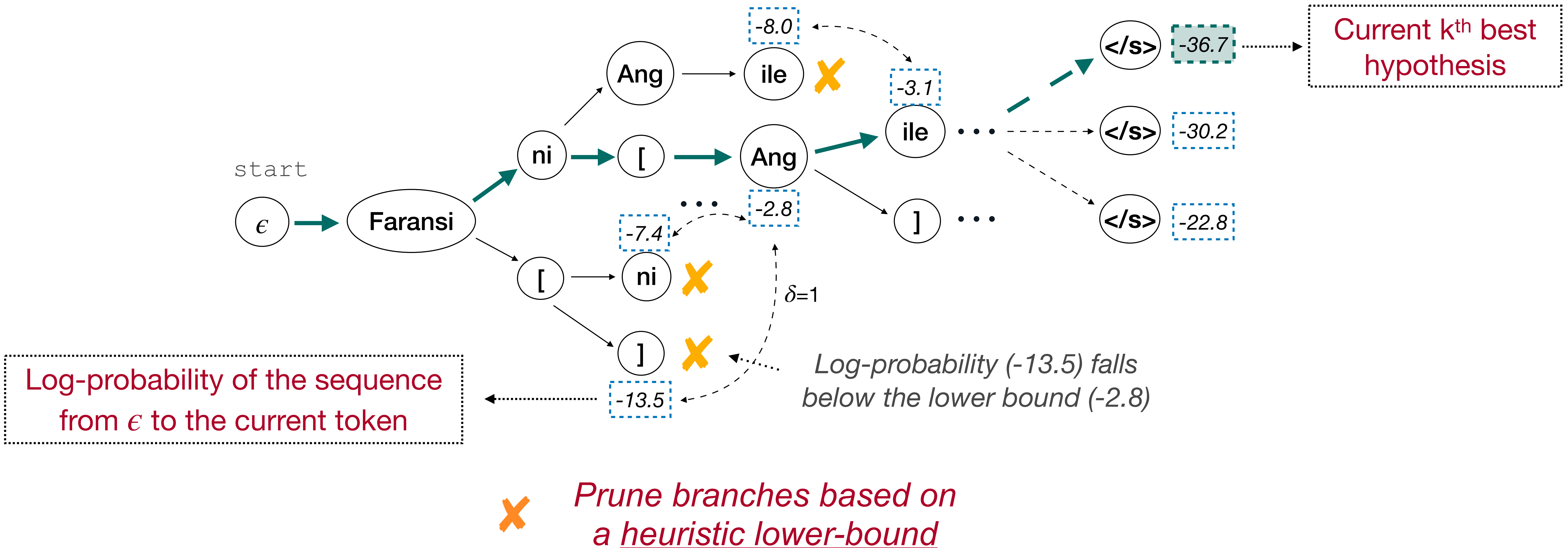


An Efficient Constrained Decoding Algorithm

(2) A branch-and-bound search algorithm with a heuristic lower bound $L_d^k = \log P(y_{1:d}^k | x^{mark})$.

$$d = \min\left(\max\left(j + \delta, q\right), |y^k|\right)$$

Input: $x =$ “Only France and Britain backed Fischler 's proposal .” $x^{mark} =$ ”Only France and [Britain] backed Fischler 's proposal .” $y^{tmpl} =$ ”Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ .”



Prune branches based on a <u>heuristic lower-bound</u>

An Efficient Constrained Decoding Algorithm

Algorithm 1 Constrained_DFS: Searching for top-k best hypotheses

Input x^{mark}: Source sentence with marker, y: translation prefix (default: ϵ), y^{tmpl}: translation template,
L: $[\log P(y_1|x), \log P(y_{1:2}|x), \ldots, \log P(y|x)]$ (default=[0.0]), $\mathcal{M}$: opening marker positions
H: min heap to record the results, k: number of hypotheses, δ: lower bound hyperparameter

$flag \leftarrow$ {check if all markers are generated}
if $y_{|y|} = \texttt{</s>}$ and $flag = \texttt{TRUE}$: **then**
    $H.\mathrm{push}((L_{|y|}, L, y))$ ▷ H sorts by the first element
    if $\mathrm{len}(H) > k$ **then**
        $H.\mathrm{pop}()$
else
    $\mathcal{T} \leftarrow []$
    $w_1 \leftarrow$ {get the next token in y^{tmpl}}
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{(w_1, \log P(w_1|y, x^{mark}))\}$
    $j \leftarrow |y| + 1$ ▷ position of the token to be generated next
    $w_2 \leftarrow$ {get the next marker}
    if $\exists\, w_2$ and not ($w_2 =$ '[' land $j \notin \mathcal{M}$) **then**
        $\mathcal{T} \leftarrow \mathcal{T} \cup \{(w_2, \log P(w_2|y, x^{mark}))\}$
    $\mathcal{T} \leftarrow$ {sort $\mathcal{T}$ by the second element in decreasing order}
    for $(w, p) \in \mathcal{T}$ **do**
        $logp \leftarrow L_{|y|} + p$
        $\gamma \leftarrow$ {compute lower bound following Eq 7}
        if $logp > \gamma$ **then**
            Constrained_DFS($x^{mark}, y \cdot w, y^{tmpl}, L \cup \{logp\}, \mathcal{M}, H, k, \delta$)
return H

Experiment Results

CODEC outperforms GPT-4, EasyProject and Awesome-align for NER and Event Extraction tasks.

- **Label Projection baselines:**
  - Alignment-based (***Awes-align***): Utilize a word-alignment system (*Awesome-align*[1]) to perform label projection
  - Marker-based (***EasyProject***): insert markers into the source sentence then translate

- **Zero-shot Cross-lingual transfer (*FT_{En}*)**

  The multilingual model is fine-tuned only on the English data

[1] *Zi-Yi Dou and Graham Neubig. Word alignment by fine-tuning embeddings on parallel corpora.* In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 2112–2128, Online, April 2021

Experiment Results

More importantly, CODEC shines on low-resource languages, such as MasakhaNER 2.0 dataset.

Lang.	GPT-4†	FT_{En}	Translate-train		
			Awes-align	EasyProject	CODEC (Δ_{FT})
Bambara	46.8	37.1	45.0	45.8	45.8 (+8.7)
Ewe	75.5	75.3	78.3	78.5	**79.1** (+3.8)
Fon	19.4	49.6	59.3	61.4	**65.5** (+15.9)
Hausa	70.7	71.7	72.7	72.2	72.4 (+0.7)
Igbo	51.7	59.3	63.5	65.6	70.9 (+11.6)
Kinyarwanda	59.1	66.4	63.2	71.0	71.2 (+4.8)
Luganda	73.7	75.3	77.7	76.7	77.2 (+1.9)
Luo	**55.2**	35.8	46.5	50.2	49.6 (+13.8)
Mossi	44.2	45.0	52.2	53.1	**55.6** (+10.6)
Chichewa	75.8	**79.5**	75.1	75.3	76.8 (-2.7)
chiShona	66.8	35.2	69.5	55.9	72.4 (+37.2)
Kiswahili	82.6	**87.7**	82.4	83.6	83.1 (-4.6)
Setswana	62.0	64.8	73.8	74.0	74.7 (+9.9)
Akan/Twi	52.9	50.1	62.7	65.3	64.6 (+14.5)
Wolof	62.6	44.2	54.5	58.9	63.1 (+18.9)
isiXhosa	69.5	24.0	61.7	**71.1**	70.4 (+46.4)
Yoruba	**58.2**	36.0	38.1	36.8	41.4 (+5.4)
isiZulu	60.2	43.9	68.9	73.0	**74.8** (+30.9)
AVG	60.4	54.5	63.6	64.9	67.1 (+12.7)

- NER: mDeBERTa-v3
- MT: NLLB

prior marker-based approach cannot do this

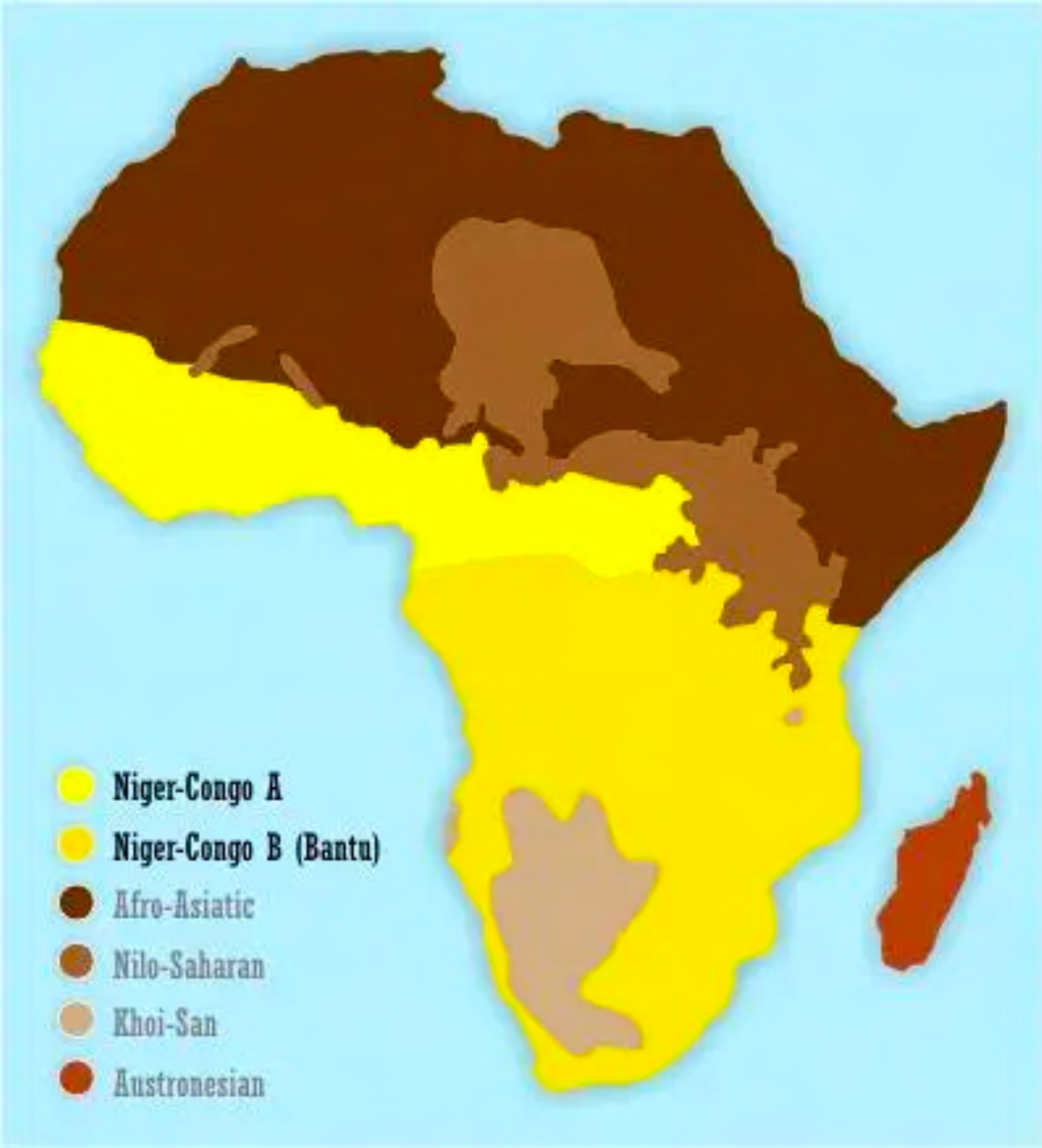
Experiment Results

"Translate-test" - CODEC can also translate test data in source language into a high-resource language to run inference on, then project predicted span labels back to the test data.

Lang.	GPT-4[†]	FT_{En}	Translate-train			Translate-test	
			Awes-align	EasyProject	CODEC (Δ_{FT})	Awes-align	CODEC (Δ_{FT})
Bambara	46.8	37.1	45.0	45.8	45.8 (+8.7)	50.0	**55.6** (+18.5)
Ewe	75.5	75.3	78.3	78.5	**79.1** (+3.8)	72.5	**79.1** (+3.8)
Fon	19.4	49.6	59.3	61.4	**65.5** (+15.9)	62.8	61.4 (+11.8)
Hausa	70.7	71.7	72.7	72.2	72.4 (+0.7)	70.0	**73.7** (+2.0)
Igbo	51.7	59.3	63.5	65.6	70.9 (+11.6)	**77.2**	72.8 (+13.5)
Kinyarwanda	59.1	66.4	63.2	71.0	71.2 (+4.8)	64.9	**78.0** (+11.6)
Luganda	73.7	75.3	77.7	76.7	77.2 (+1.9)	**82.4**	82.3 (+7.0)
Luo	**55.2**	35.8	46.5	50.2	49.6 (+13.8)	52.6	52.9 (+17.1)
Mossi	44.2	45.0	52.2	53.1	**55.6** (+10.6)	48.4	50.4 (+5.4)
Chichewa	75.8	**79.5**	75.1	75.3	76.8 (-2.7)	78.0	76.8 (-2.7)
chiShona	66.8	35.2	69.5	55.9	72.4 (+37.2)	67.0	**78.4** (+43.2)
Kiswahili	82.6	**87.7**	82.4	83.6	83.1 (-4.6)	80.2	81.5 (-6.2)
Setswana	62.0	64.8	73.8	74.0	74.7 (+9.9)	**81.4**	80.3 (+15.5)
Akan/Twi	52.9	50.1	62.7	65.3	64.6 (+14.5)	72.6	**73.5** (+23.4)
Wolof	62.6	44.2	54.5	58.9	63.1 (+18.9)	58.1	**67.2** (+23.0)
isiXhosa	69.5	24.0	61.7	**71.1**	70.4 (+46.4)	52.7	69.2 (+45.2)
Yoruba	**58.2**	36.0	38.1	36.8	41.4 (+5.4)	49.1	58.0 (+22.0)
isiZulu	60.2	43.9	68.9	73.0	**74.8** (+30.9)	64.1	**76.9** (+33.0)
AVG	60.4	54.5	63.6	64.9	67.1 (+12.7)	65.8	**70.4** (+16.0)

Error Analysis

Underline marks the projection errors.



only marks sub-words as an entity

	English Data	Augmented data in low-resource languages		
		EasyProject	Awesome-align	Codec
chiShona	**India**$_{LOC}$ and **Pakistan**$_{LOC}$ have fought … region of **Kashmir**$_{LOC}$ …	**India**$_{LOC}$ ne **Pakistan**$_{LOC}$ … ye **Kashmir**$_{LOC}$ chibviro …	**India**$_{LOC}$ nePakistan … zvinetso **yeKashmir**$_{LOC}$ …	**India**$_{LOC}$ **nePakistan**$_{LOC}$ … zvinetso **yeKashmir**$_{LOC}$ …
isiZulu	State media quoted **China**$_{LOC}$'s top negotiator with **Taipei**$_{LOC}$, **Tang Shubei**$_{PER}$, … from **Taiwan**$_{LOC}$ …	Imithombo … we **China**$_{LOC}$ ne **Taipei**$_{LOC}$, **uTang Shubei**$_{PER}$, … elivela **eTaiwan**$_{LOC}$ …	**Imithombo**$_{LOC}$ … waseChina neTaipei , **uTang Shubei**$_{PER}$, … elivela eTaiwan …	Imithombo … **waseChina**$_{LOC}$ **neTaipei**$_{LOC}$, **uTang Shubei**$_{PER}$, … elivela **eTaiwan**$_{LOC}$ …

having difficulty to project multiple spans

Duong Minh Le, Yang Chen, Alan Ritter, Wei Xu. "Constrained Decoding for Cross-lingual Label Projection" (ICLR 2024)

Today’s Talk —

1 - Cross-lingual Transfer Learning

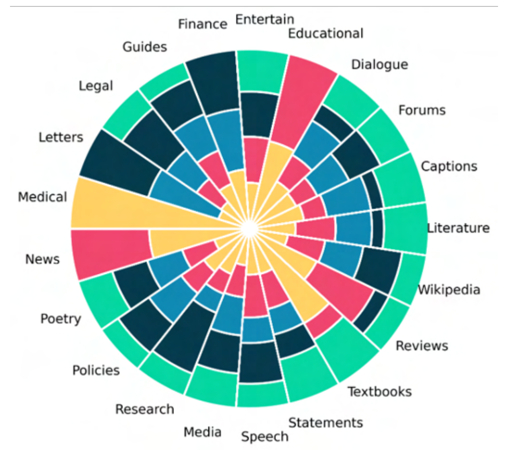
CODEC

(Le et al., ICLR 2024)

Design decoding algorithms to improve performance on non-English languages.

2 - Multilingual Multi-domain Datasets

ReadMe++ & MedReadMe



(Naous et al., EMNLP 2024 & Chao et al., EMNLP 2024)

Support not only more languages but also more text domains/genres.

Today’s Talk —

1 - Cross-lingual Transfer Learning

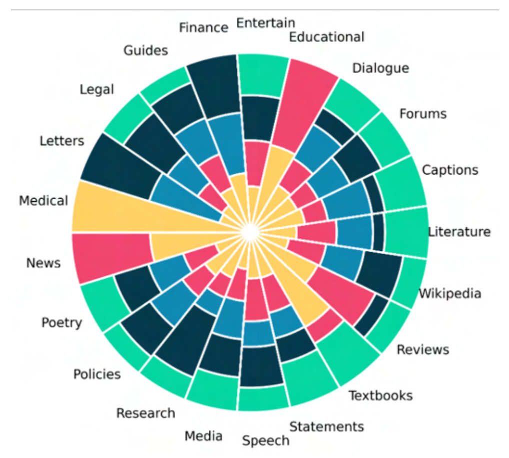
CODEC

(Le et al., ICLR 2024)

Design decoding algorithms to improve performance on non-English languages.

2 - Multilingual Multi-domain Datasets

ReadMe++ & MedReadMe



(Naous et al., EMNLP 2024 & Chao et al., EMNLP 2024)

Support not only more languages but also more text domains/genres.

Text Simplification

Rewrite complex text into simpler language while retain its original meaning.

Science

Preserved on ancient teeth, a fossilized microbial world

By Deborah Netburn, Los Angeles Times
Published: 03/05/2014 Word Count: 682

The layers of calcified plaque entomb the bacteria that also live in our mouths -- turning them into small fossils even when we are alive. And when we die, these dense, calcified micro-fossils remain intact, even as most of the rest of us decomposes.

Text Simplification

Rewrite complex text into simpler language while retain its original meaning.

The layers of calcified plaque entomb the bacteria that also live in our mouths -- turning them into small fossils even when we are alive. And when we die, these dense, calcified micro-fossils remain intact, even as most of the rest of us decomposes.

Text Simplification

Rewrite complex text into simpler language while retain its original meaning.

The layers of calcified plaque entomb the bacteria that also live in our mouths -- turning them into small fossils even when we are alive.

And when we die, these dense, calcified micro-fossils remain intact, even as most of the rest of us decomposes.

Text Simplification

Rewrite complex text into simpler language while retain its original meaning.

The layers of ~~calcified~~ plaque entomb the bacteria that also ~~live~~ in our mouths -- turning them into small fossils ~~even when we are alive.~~

split

The buildup of plaque can trap the bacteria that live in our mouths.

It turns them into tiny fossils.

And when we die, these ~~dense, calcified~~ micro-fossils remain intact, even ~~as most of the rest of us decomposes~~.

paraphrase

Even after death, these micro-fossils don't break down.

Human Text Simplification

Professional editors rewrite news articles into 4 different readability levels for grade 3-12 students.

Wei Xu, Chris Callison-Burch, Courtney Napoles. "Problems in Current Text Simplification Research: New Data Can Help" (TACL 2015)
Yang Zhong, Chao Jiang, Wei Xu, Jessy Li. "Discourse Level Factors for Sentence Deletion in Text Simplification" (AAAI 2020)

Why Text Simplification?

It can help a lot of people!

- Children (Leonardo et al., 2018) ← research on education using Newsela data
- Second language learners (Housel et al., 2020)
- Deaf and hard-of-hearing students (Alonzo et al., 2020) ← using our EMNLP 2018 work on lexical simplification
- People with dyslexia (Rello at al., 2013)
- People with autism spectrum disorder (González-Navarro et al., 2014)

- and many others … e.g., to read legal & medical documents (Trienes et al. 2024; Joseph et al. 2024), etc.

Mounica Maddela, Wei Xu. "A Neural Readability Ranking Model and A Word-Complexity Lexicon for Lexical Simplification" (EMNLP 2018)

Other Text Generation Tasks

- **Multilingual split and rephrase** (Daniel Kim*, Mounica Maddela*, Reno Kriz, Wei Xu, Chris Callison-Burch — EMNLP 2021)

  An additional advantage is that a shorter ramp can be used, thereby reducing weight and improving the rear view of the driver.

  Another advantage is that a shorter ramp can be used. || This saves weight and improves the look of the rear of the vehicle.

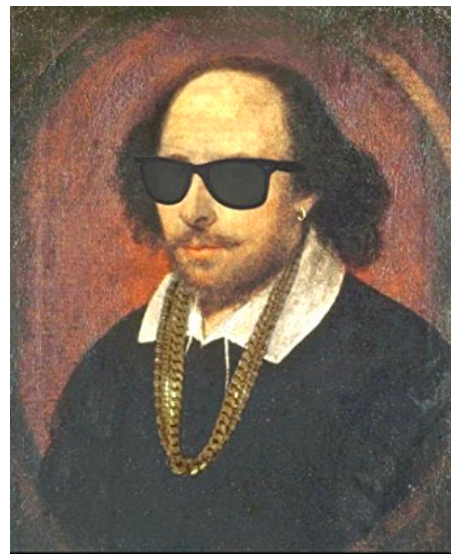
- **Neutralizing biased languages** (Zhong Yang, Jingfeng Yang, Diyi Yang, Wei Xu — EMNLP 2021 Findings)

  A Golden duck may refer to: A cricket 'golden' duck in which a batsman is out for nought on the first ball he faces.
  A cricket 'golden' duck in which a batter is out for nought on the first ball they face.

- **Large-scale paraphrase identification and generation** (Yao You, Chao Jiang, Wei Xu - EMNLP 2022)

- **Style transfer** (Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, Colin Cherry - COLING 2012)

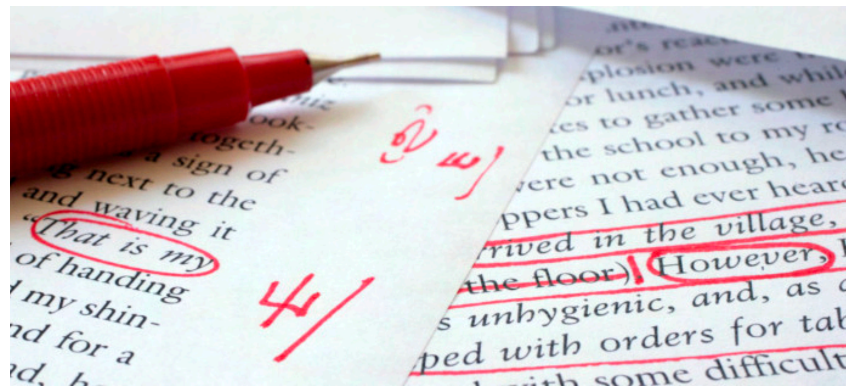
  If you will not be turned, you will be destroyed! — Star Wars

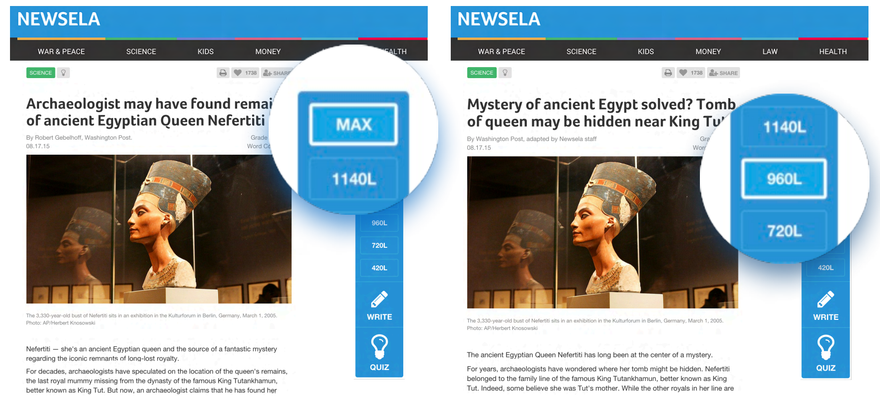
  If you will not be turn'd, you will be undone!

Automatic Text Simplification

It is a great benchmark for natural language generation (NLG) models.

Need both **diversity** and **controllability** from the model to meet users' varied reading needs.





complicated rewriting

good training data

~reliable evaluation

(covers other text-to-text tasks: splitting, compression, paraphrase generation, style transfer, etc.)

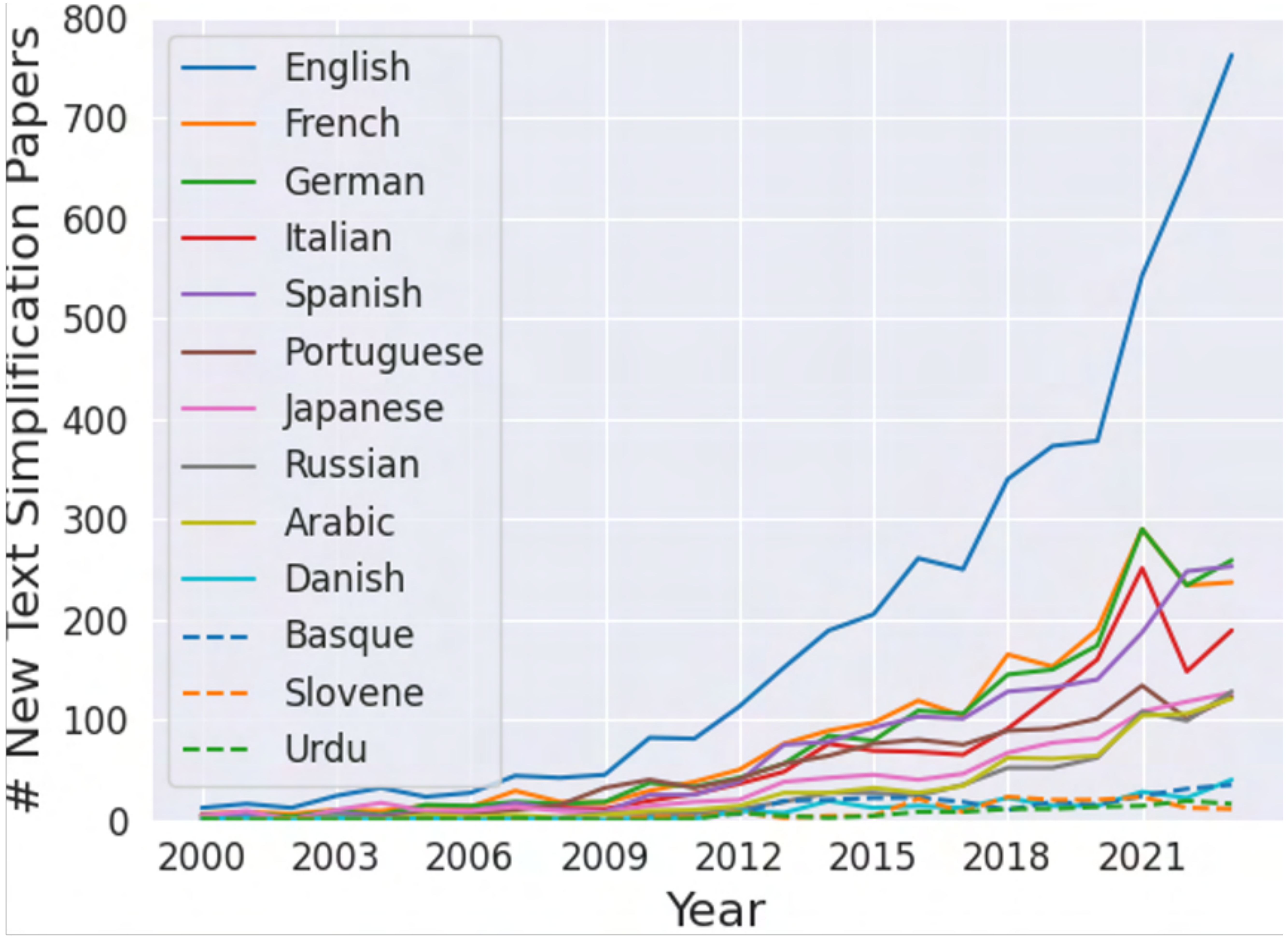
Revisiting Non-English Text Simplification: a Unified Multilingual Benchmark

Michael J. Ryan Tarek Naous. Wei Xu

🏆 **Best Paper Award Honorable Mention - ACL 2023**

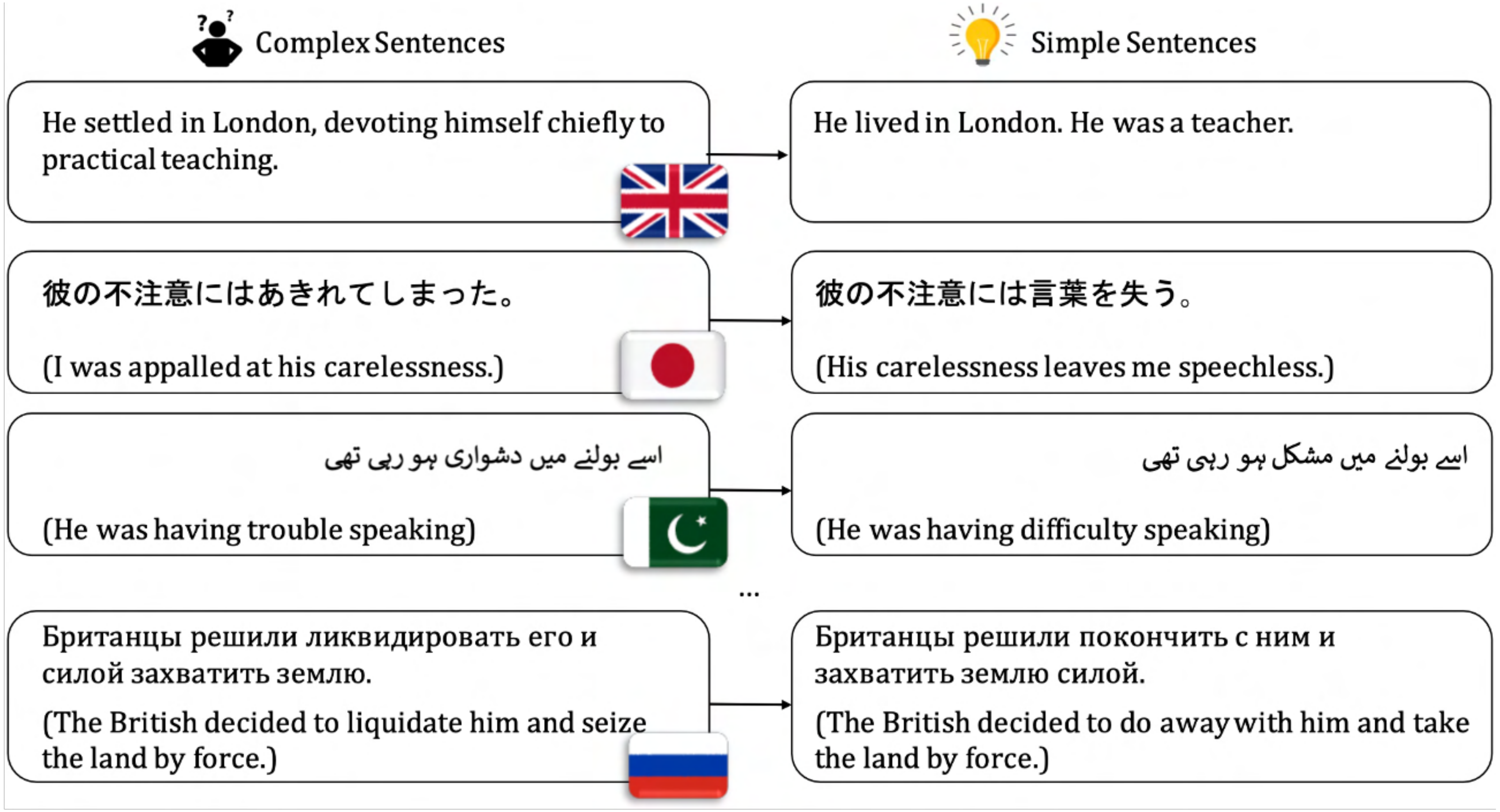
Growth of Text Simplification Research

- **In 2023 alone:**
  - 763 new papers on English text simplification
  - 237 new papers on French
  - <20 papers related to Urdu or Slovene simplification



(Count based on Google Scholar)

We introduce MultiSim of parallel texts



12 languages and growing (now 15)

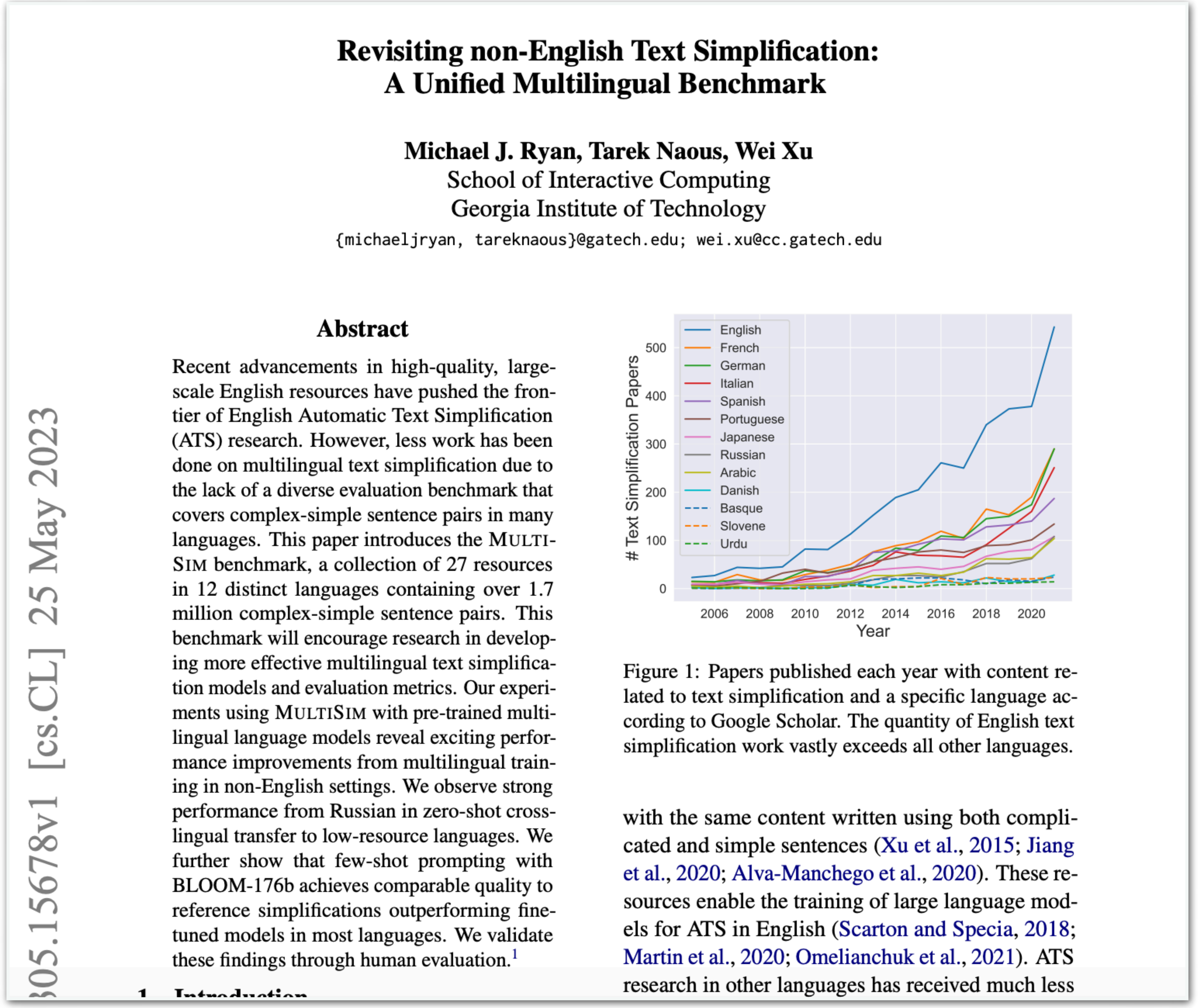
Corpus	Source(s)	Simplification Author	Collection Strategy	Alignment Level	Sentence Aligned	Complex Sentences	Simple Sentences	Access
Arabic Corpora								
Saaq al-Bambuu (Khallaf and Sharoff, 2022)	Literature	writer	Target Audience Resource	sentence	auto	2,980	2,980	private
Basque Corpora								
CBST (Gonzalez-Dios et al., 2018)	Science Communications	translator, teacher	Annotator	document	manual	458	591	on request
Brazilian Portuguese Corpora								
PorSimples (Aluísio and Gasperin, 2010)	News, Science Communications	linguist	Annotator	document	manual	7,902	10,174	on request
Danish Corpora								
DSim (Klerke and Søgaard, 2012)	News	journalists	Target Audience Resource	sentence	auto	47,887	60,528	on request
English Corpora†								
ASSET (Alva-Manchego et al., 2020)	W	crowdsource	Annotator	sentence	manual	2,359	23,590	open source
Newsela EN (Xu et al., 2015)	News	experts	Target Audience Resource	document	auto	393,798	402,222	on request
Wiki-Auto (Jiang et al., 2020)	W	crowdsource	Automatic	document	auto	10,144,476	1,241,671	open source
French Corpora								
Alector (Gala et al., 2020)	Websites, Science Communications	experts	Annotator	document	NA	1,230	1,192	open source
CLEAR (Grabar and Cardon, 2018)	W, Medical Documents	crowdsource, experts	Automatic	sentence	auto	4,596	4,596	open source
WikiLarge FR (Cardon and Grabar, 2020)	W	crowdsource	Translation	sentence	auto	307,067	308,409	open source
German Corpora								
GEOLinoTest (Mallinson et al., 2020)	News	linguist	Annotator	sentence	manual	1,198	1,198	open source
German News (Säuberli et al., 2020)	News	news agency	Target Audience Resource	document	auto	15,239	14,344	on request
Klexikon (Aumiller and Gertz, 2022)	W	crowdsource	Automatic	document	NA	771,059	96,870	open source
Simple Patho (Trienes et al., 2023)	Medical Documents	medical students	Annotator	paragraph	manual	22,191	26,551	private
Simple German (Battisti et al., 2020)	Websites	government	Target Audience Resource	document	auto	12,806	8,400	on request*
TextComplexityDE (Naderi et al., 2019)	W	native speaker	Annotator	document	manual	250	250	open source
Italian Corpora								
AdminIT (Miliani et al., 2022)	Government	researchers	Annotator	sentence	manual	777	763	open source
SIMPITIKI Wiki (Tonelli et al., 2016)	W	crowdsource	Automatic	sentence	manual	575	575	open source
PaCCSS-IT (Brunato et al., 2016)	Websites	crowdsource	Automatic	sentence	auto	63,006	63,006	open source
Teacher (Brunato et al., 2015)	Websites	teachers	Annotator	document	manual	204	195	open source
Terence (Brunato et al., 2015)	Literature	experts	Annotator	document	manual	1,035	1,060	open source
Japanese Corpora								
EasyJapanese (Maruyama and Yamamoto, 2018)	News, Websites	students	Annotator	sentence	manual	50,000	50,000	open source
EasyJapaneseExtended (Katsuta and Yamamoto, 2018)	News, Websites	crowdsource	Annotator	sentence	manual	34,400	35,000	open source
Japanese News (Goto et al., 2015)	News	journalists, teachers	Target Audience Resource	document	auto	13,356	13,356	private
Russian Corpora								
RuAdapt Encyclopedia (Dmitrieva et al., 2021)	Encyclopedic	researchers	Annotator	document	auto	9,729	10,230	open source
RuAdapt Fairytale (Dmitrieva et al., 2021)	Literature	researchers	Annotator	document	auto	310	404	open source
RuAdapt Lit (Dmitrieva and Tiedemann, 2021)	Literature	writers	Annotator	document	auto	24,152	28,259	on request
RSSE (Sakhovskiy et al., 2021)	W	crowdsource	Annotator	sentence	manual	2,000	6,804	open source
RuWikiLarge (Sakhovskiy et al., 2021)	W	crowdsource	Translation	sentence	auto	278,499	289,788	on request
Slovene Corpora								
SloTS (Gorenc and Robnik-Šikonja, 2022)	Literature	experts	Target Audience Resource	sentence	manual	1,181	1,287	open source
Spanish Corpora								
FIRST (Orasan et al., 2013)	Literature, News, Medical Documents	experts	Annotator	document	manual	320	332	private
Newsela ES (Xu et al., 2015)	News	experts	Target Audience Resource	document	auto	46,256	45,519	on request
Simplext (Saggion et al., 2015)	News	researchers	Annotator	document	manual	1,108	1,742	on request
Urdu Corpora								
SimplifyUREval (Qasmi et al., 2020)	Literature, News	expert	Annotator	sentence	manual	500	736	open source

Table 1: Important properties of text simplification parallel corpora. †Common English corpora included for comparison. Many other English corpora omitted. *Only scripts to replicate the corpus are available upon request. Simple German results differ from original paper because of changes to availability of online articles. *Sources*: Literature, Science Communications, News, Wikipedia, Websites, Medical Documents, Government, Encyclopedic. *Collection Strategies*: Automatic, Translation, Annotator, Target Audience Resource.

Open Source

MultiSim data and code (loaders) are available - https://github.com/XenonMolecule/MultiSim

Paper on arXiv



Revisiting non-English Text Simplification:
A Unified Multilingual Benchmark

Michael J. Ryan, Tarek Naous, Wei Xu
School of Interactive Computing
Georgia Institute of Technology
{michaeljryan, tareknaous}@gatech.edu; wei.xu@cc.gatech.edu

arXiv:2305.15678v1 [cs.CL] 25 May 2023

Abstract

Recent advancements in high-quality, large-scale English resources have pushed the frontier of English Automatic Text Simplification (ATS) research. However, less work has been done on multilingual text simplification due to the lack of a diverse evaluation benchmark that covers complex-simple sentence pairs in many languages. This paper introduces the MULTISIM benchmark, a collection of 27 resources in 12 distinct languages containing over 1.7 million complex-simple sentence pairs. This benchmark will encourage research in developing more effective multilingual text simplification models and evaluation metrics. Our experiments using MULTISIM with pre-trained multilingual language models reveal exciting performance improvements from multilingual training in non-English settings. We observe strong performance from Russian in zero-shot cross-lingual transfer to low-resource languages. We further show that few-shot prompting with BLOOM-176b achieves comparable quality to reference simplifications outperforming fine-tuned models in most languages. We validate these findings through human evaluation.[1]

1 Introduction

Figure 1: Papers published each year with content related to text simplification and a specific language according to Google Scholar. The quantity of English text simplification work vastly exceeds all other languages.

with the same content written using both complicated and simple sentences (Xu et al., 2015; Jiang et al., 2020; Alva-Manchego et al., 2020). These resources enable the training of large language models for ATS in English (Scarton and Specia, 2018; Martin et al., 2020; Omelianchuk et al., 2021). ATS research in other languages has received much less

Data on Huggingface

Michael J. Ryan, Tarek Naous, Wei Xu. "Revisiting Non-English Text Simplification: a Unified Multilingual Benchmark" (ACL 2023)

Benchmarking Multilingual LMs for Multi-domain Readability Assessment (ReadMe++)


Tarek Naous

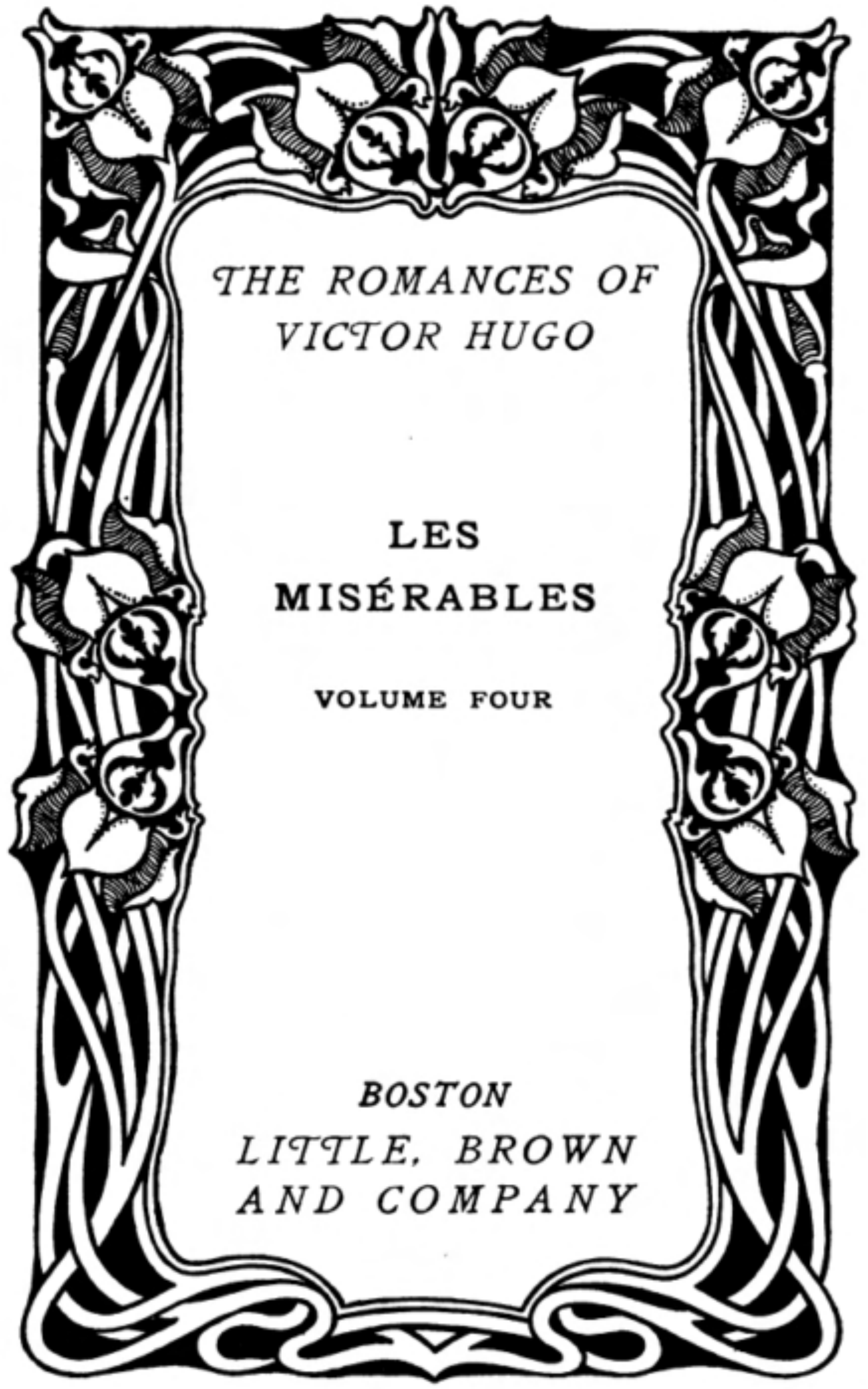

Michael J. Ryan


Anton Lavrouk


Mohit Chandra


Wei Xu

Different Readability Levels


THE ROMANCES OF
VICTOR HUGO

LES
MISÉRABLES

VOLUME FOUR

BOSTON
LITTLE, BROWN
AND COMPANY

“In the uncoerced slowness of its gait, suppleness and agility were discernible.”

“In its voluntary slow movement, its flexibility and agility were noticeable.”

“In its voluntary slow movement, you could still see how flexible and quick it is.”

Prior Work on Readability Measurements

Human-annotated Resources (Arase et al. 2022, Brunato et al. 2018, and more)

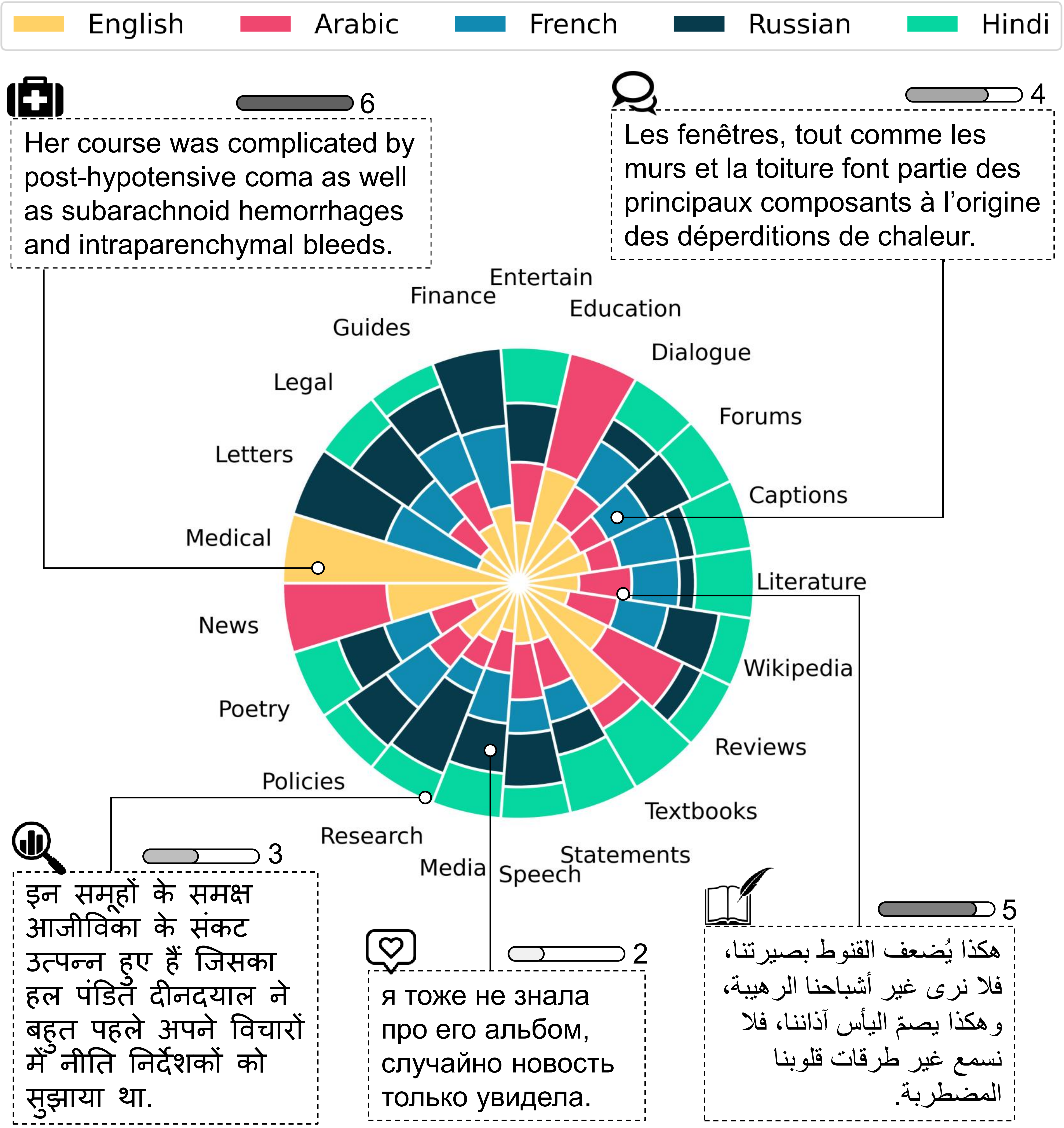
- CEFR: Common European Framework of Reference for Languages
- Mostly using either Wikipedia or news data

Level	Description	Rating
A1	Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.	1
A2	Can understand short, simple texts on familiar matters of a concrete type.	2
B1	Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.	3
B2	Can read with a large degree of independence, adapting style and speed of reading to different texts and purpose.	4
C1	Can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of speciality, provided he/she can reread difficult sections.	5
C2	Can understand and interpret critically virtually all forms of the written language including abstract, structurally complex, or highly colloquial literary and non-literary writings.	6

Yuki Arase, Satoru Uchida, Tomoyuki Kajiwara. "CEFR-Based Sentence Difficulty Annotation and Assessment" (EMNLP 2024)

Our Work - Readme++

- **More diverse languages**
  - 5 different languages
  - written in 4 different scripts
  - 9,465 human-annotated sentences
- **And, more diverse domains**
  - 21 top-level domains
  - 112 data sources
  - all with open license



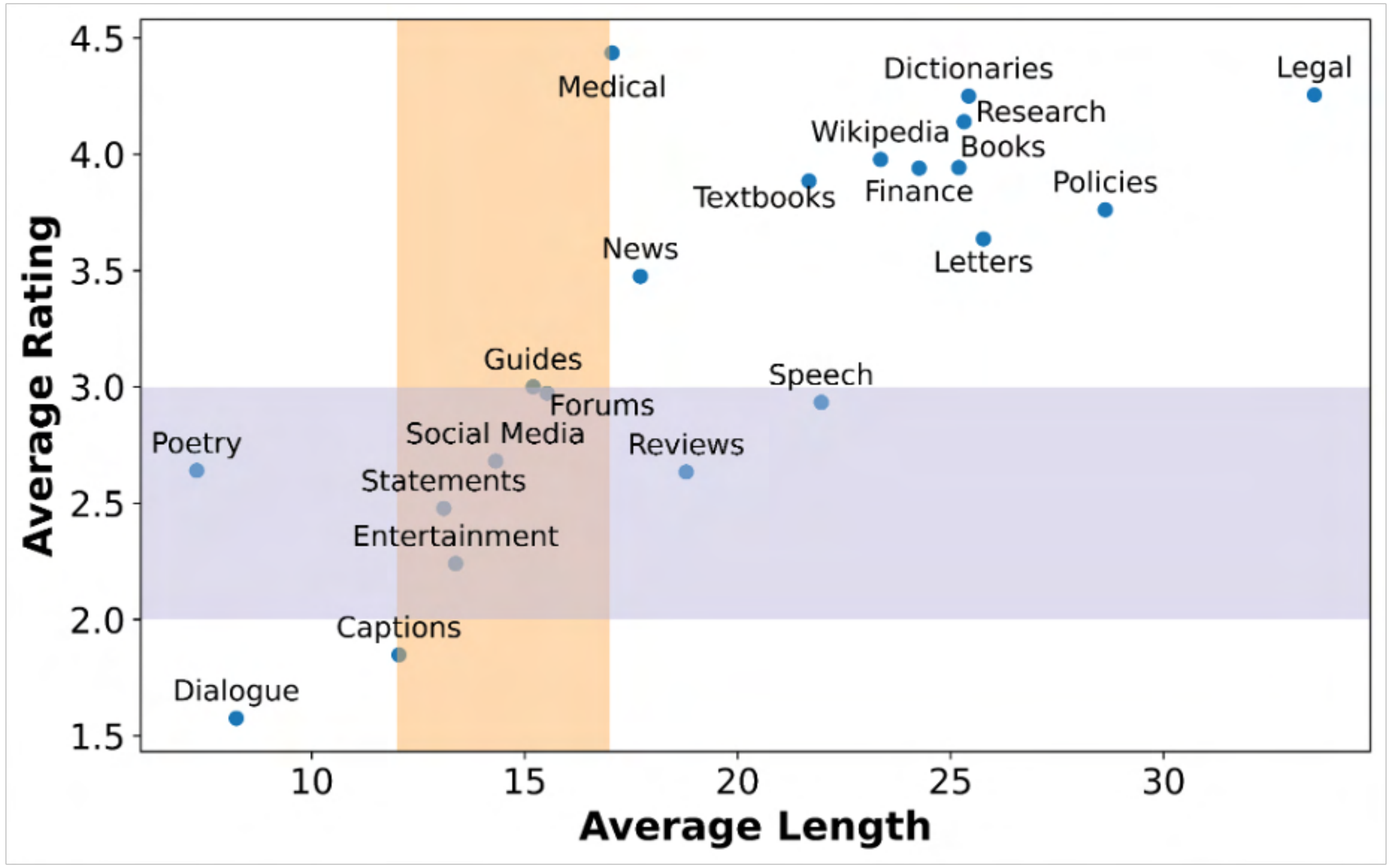
Our Work - Readme++

- A (partial) list of representative sources we sampled data from:

Domain (Abrv)	#	Examples of Data Sources — Full list for all languages in Appendix A		
		Arabic (ar)	English (en)	Hindi (hi)
Captions (Cap)	9	**Images** (ElJundi et al., 2020)	**Videos** (Wang et al., 2019)	**Movies** (Lison and Tiedemann, 2016)
Dialogue (Dia)	7	**Open-domain** (Naous et al., 2020)	**Negotiation** (He et al., 2018)	**Task-oriented** (Malviya et al., 2021)
Dictionaries (Dic)	2	**Dictionaries** (almaany.com)	**Dictionaries** (dictionary.com)	—
Entertainment (Ent)	4	**Jokes** (almrsal.com)	**Jokes** (Weller and Seppi, 2019)	**Jokes** (123hindijokes.com)
Finance (Fin)	3	—	**Finance** (Malo et al., 2014)	—
Forums (For)	7	**QA Websites** (hi.quora.com)	**StackOverflow** (Tabassum et al., 2020)	**Reddit** (reddit.com)
Guides (Gui)	6	**Online Tutorials** (ar.wikihow.com)	**Code Documentation** (mathworks.com)	**Cooking Recipes** (narendramodi.in)
Legal (Leg)	9	**UN Parliament** (Ziemski et al., 2016)	**Constitutions** (constitutioncenter.org)	**Judicial Rulings** (Kapoor et al., 2022)
Letters (Let)	3	—	**Letters** (oflosttime.com)	—
Literature (Lit)	3	**Novels** (hindawi.org/books/)	**History** (gutenberg.org)	**Biographies** (Public Domain Books)
Medical Text (Med)	1	—	**Clinical Reports** (Uzuner et al., 2011)	—
News Articles (New)	2	**Sports** (Alfonse and Gawich, 2022)	**Economy** (Misra, 2022)	—
Poetry (Poe)	5	**Poetry** (aldiwan.net)	**Poetry** (poetryfoundation.org)	**Poetry** (hindionlinejankari.com)
Policies (Pol)	7	**Olympic Rules** (specialolympics.org)	**Contracts** (honeybook.com)	**Code of Conduct** (lonza.com)
Research (Res)	15	**Politics** (jcopolicy.uobaghdad.edu.iq)	**Science & Engineering** (arxiv.org)	**Economics** (journal.ijarms.org)
Social Media (Soc)	3	**Twitter** (Zheng et al., 2022)	**Twitter** (Zheng et al., 2022)	**Twitter** (Zheng et al., 2022)
Speech (Spe)	4	**Public Speech** (state.gov/translations)	**Public Speech** (whitehouse.gov)	**Ted Talks** (ted.com/talks)
Statements (Sta)	6	**Quotes** (arabic-quotes.com)	**Rumours** (Zheng et al., 2022)	**Quotes** (wahh.in)
Textbooks (Tex)	3	**Business** (hindawi.org/books/)	**Agriculture** (open.umn.edu)	**Psychology** (ncert.nic.in)
User Reviews (Rev)	12	**Products** (ElSahar and El-Beltagy, 2015)	**Books** (goodreads.com)	**Movies** (hindi.webdunia.com)
Wikipedia (Wik)	1	**Wikipedia** (wikipedia.com)	**Wikipedia** (wikipedia.com)	**Wikipedia** (wikipedia.com)
Total	112			

What difference does this make?

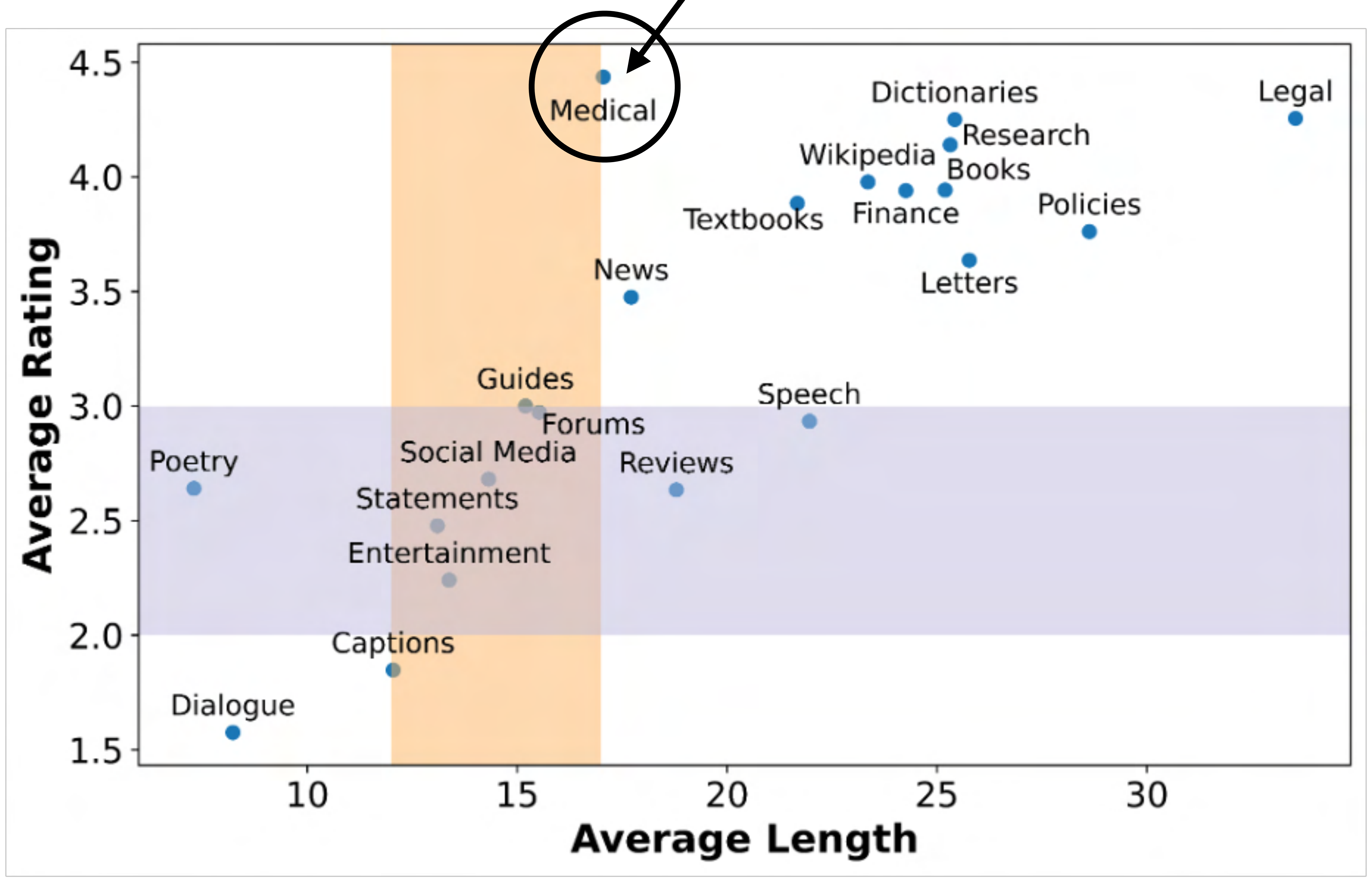
A wider range of topics and lengths of sentences that impact the readability are accounted for.



What difference does this make?

A wider range of topics and lengths of sentences that impact the readability are accounted for.

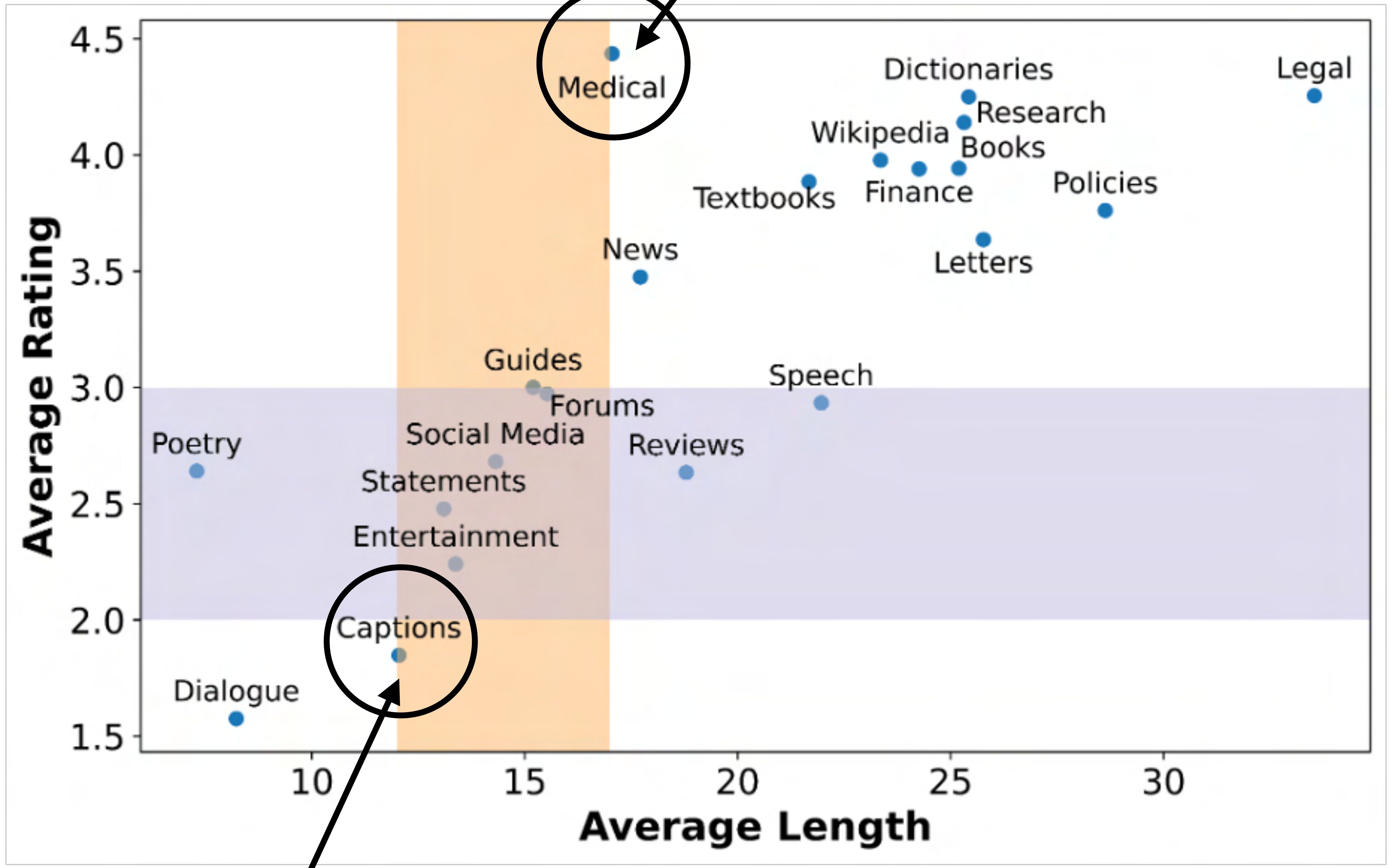
"With history, will go for cardiac catheterization evaluation."



What difference does this make?

A wider range of topics and lengths of sentences that impact the readability are accounted for.

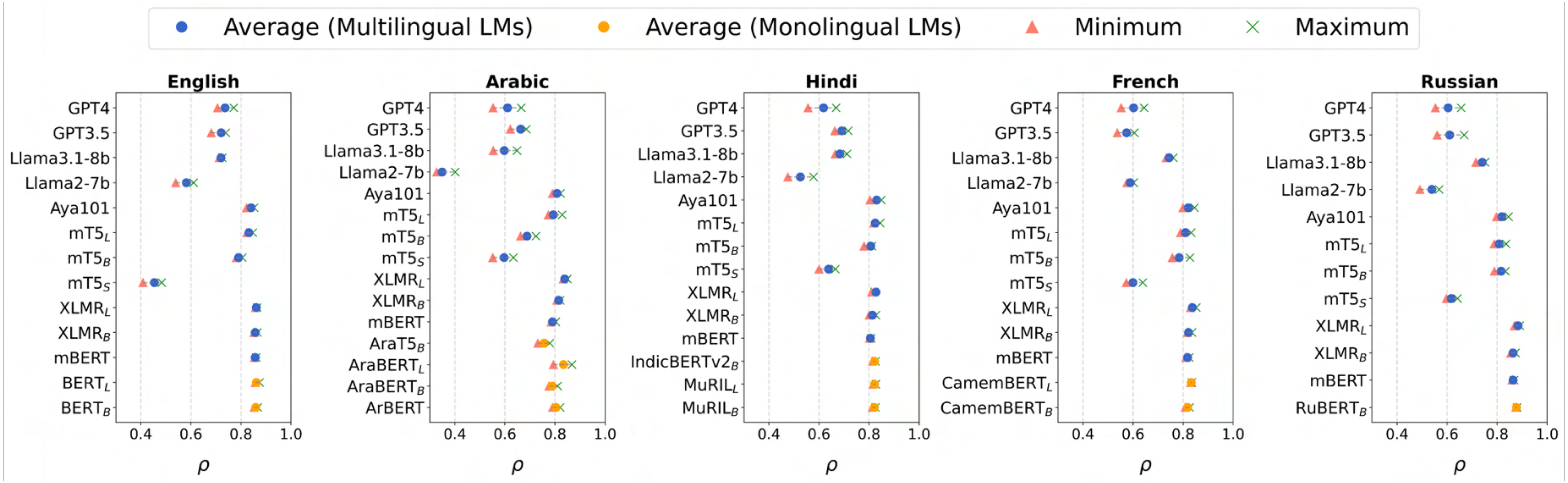
"With history, will go for cardiac catheterization evaluation."



"A young boy is indoors showing his family his dance moves."

Benchmarking multilingual LLMs

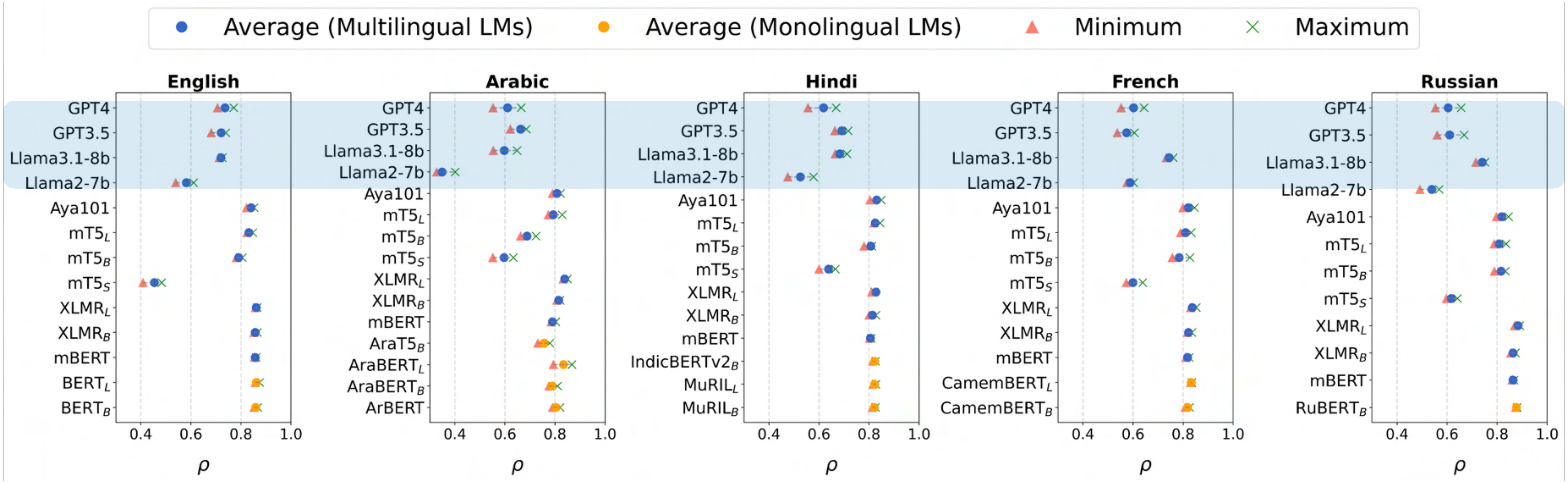
Fine-tuning LLMs perform better than 5-shot prompting of GPT-4 / Llama-3.1 (6-way classification)



i.e., human annotated data is very useful, not only for evaluation but also for fine-tuning.

Benchmarking multilingual LLMs

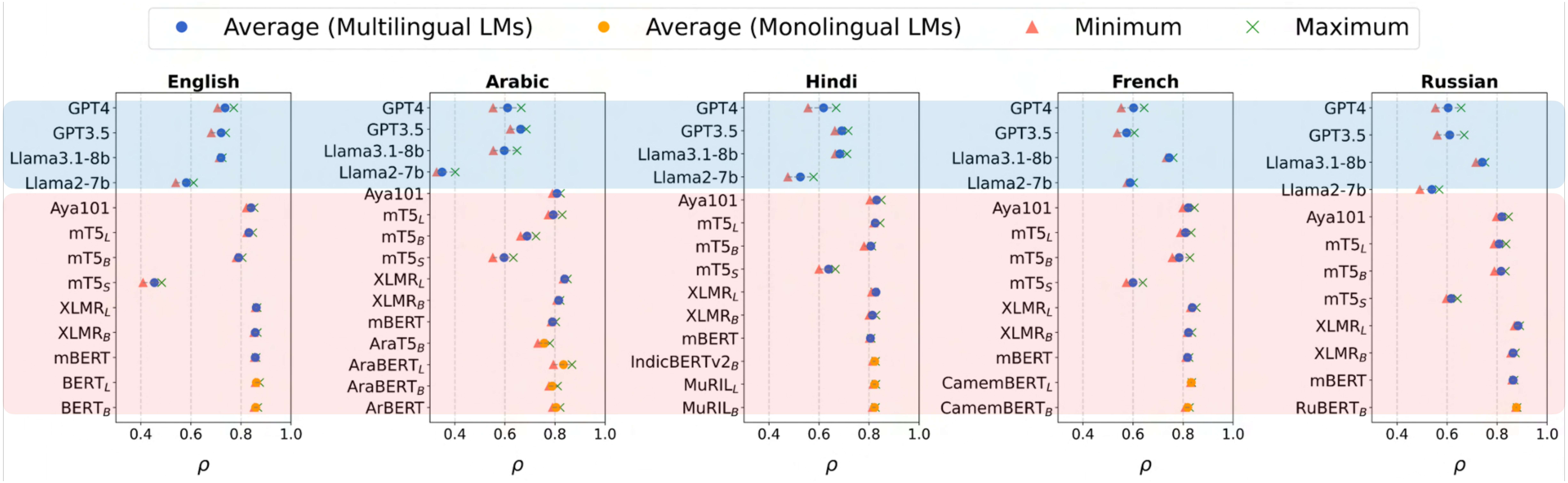
Fine-tuning LLMs perform better than 5-shot prompting of GPT-4 / Llama-3.1 (6-way classification)



i.e., human annotated data is very useful, not only for evaluation but also for fine-tuning.

Benchmarking multilingual LLMs

Fine-tuning LLMs perform better than 5-shot prompting of GPT-4 / Llama-3.1 (6-way classification)

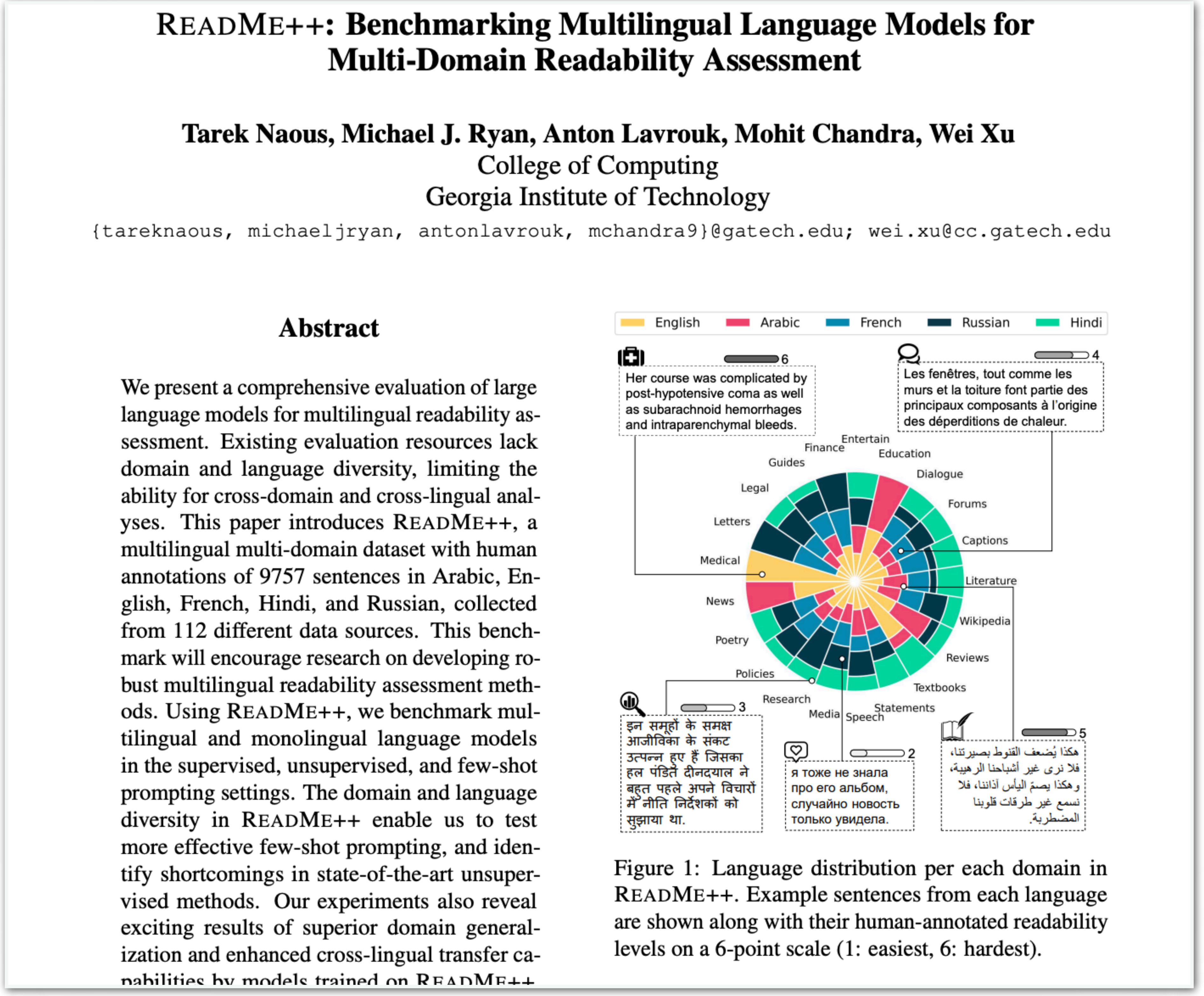


i.e., human annotated data is very useful, not only for evaluation but also for fine-tuning.

Open Source

ReadMe++ data and models are available - https://github.com/tareknaous/readme

Paper on arXiv



READMe++: Benchmarking Multilingual Language Models for Multi-Domain Readability Assessment

Tarek Naous, Michael J. Ryan, Anton Lavrouk, Mohit Chandra, Wei Xu
College of Computing
Georgia Institute of Technology
{tareknaous, michaeljryan, antonlavrouk, mchandra9}@gatech.edu; wei.xu@cc.gatech.edu

Abstract

We present a comprehensive evaluation of large language models for multilingual readability assessment. Existing evaluation resources lack domain and language diversity, limiting the ability for cross-domain and cross-lingual analyses. This paper introduces READMe++, a multilingual multi-domain dataset with human annotations of 9757 sentences in Arabic, English, French, Hindi, and Russian, collected from 112 different data sources. This benchmark will encourage research on developing robust multilingual readability assessment methods. Using READMe++, we benchmark multilingual and monolingual language models in the supervised, unsupervised, and few-shot prompting settings. The domain and language diversity in READMe++ enable us to test more effective few-shot prompting, and identify shortcomings in state-of-the-art unsupervised methods. Our experiments also reveal exciting results of superior domain generalization and enhanced cross-lingual transfer capabilities by models trained on READMe++

Figure 1: Language distribution per each domain in READMe++. Example sentences from each language are shown along with their human-annotated readability levels on a 6-point scale (1: easiest, 6: hardest).

Models on Huggingface

Installation

```
pip install readmepp
```

Usage

First import the class `ReadMe` and create a BERT predictor instance of it.
The parameter `lang` is to specify language (we support "en", "ar", "fr", "ru", and "hi").

```
from readmepp import ReadMe

predictor = ReadMe(lang='en')
```

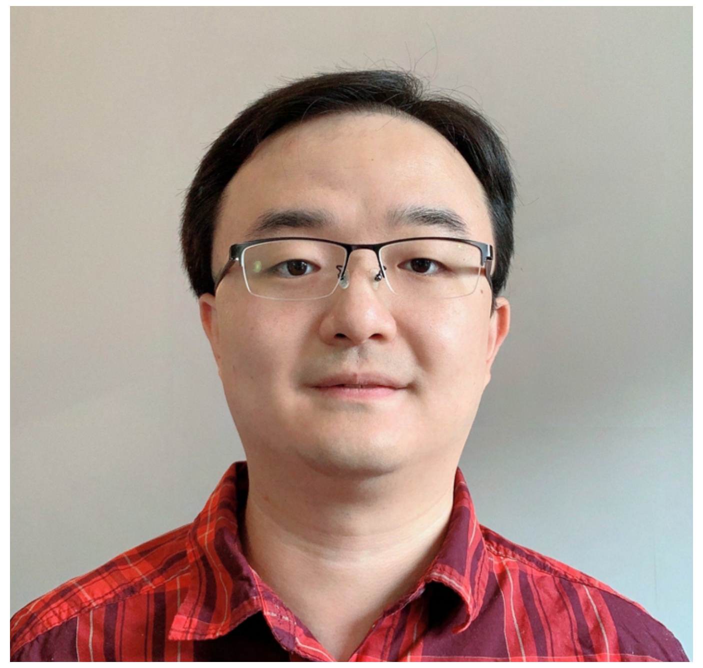
To assess the readability of a sentence, use the `predict` function of the model:

```
sentence = 'Eukaryotes differ from prokaryotes in multiple ways, with unique biochemical pathway

prediction = predictor.predict(sentence)

print(f"Predicted Readability Level: {prediction}")
```

English: https://huggingface.co/tareknaous/readabert-en
Arabic: https://huggingface.co/tareknaous/readabert-ar
Hindi: https://huggingface.co/tareknaous/readabert-hi
French: https://huggingface.co/tareknaous/readabert-fr
Russian: https://huggingface.co/tareknaous/readabert-ru

Tarek Naous, Michael J. Ryan, Anton Lavrouk, Mohit Chandra, Wei Xu. "ReadMe++: Benchmarking Multilingual LMs for Multi-domain Readability Assessment" (EMNLP 2024)

MedReadMe: A Systematic Study for Fine-grained Sentence Readability in Medical Domain



Chao Jiang

Wei Xu

“An oro-antral communication (OAC) is an unnatural opening between the oral cavity and maxillary sinus. When it fails to close spontaneously, it remains patent and is epithelialized to develop into an oro-antral fistula. These complications occur most commonly during extraction of upper molar and premolar teeth (48%).”

Review language : English | Website language : English | Sign In

Cochrane Library — Trusted evidence. Informed decisions. Better health.

Title Abstract Keyword | Browse | Advanced search

Cochrane Reviews | Searching for trials | Clinical Answers | About | Help | About Cochrane

Cochrane Database of Systematic Reviews | Review - Intervention | New search

Interventions for treating oro-antral communications and fistulae due to dental procedures

Salian Kiran Kumar Krishanappa, Prashanti Eachempati, Sumanth Kumbargere Nagraj, Naresh Yedthare Shetty, Soe Moe, Himanshi Aggarwal, Rebecca J Mathew   Authors' declarations of interest

Version published: 16 August 2018   Version history

https://doi.org/10.1002/14651858.CD011784.pub3

Download PDF | Cite this Review | Print | Comment | Share | Follow

Cited in 1 guideline

Collapse all   Expand all

Abstract

Available in English | Español | فارسی | Português | ภาษาไทย | 简体中文

Background

An oro-antral communication is an unnatural opening between the oral cavity and maxillary sinus. When it fails to close spontaneously, it remains patent and is epithelialized to develop into an oro-antral fistula. Various surgical and non-surgical techniques have been used for treating the condition. Surgical procedures include flaps, grafts and other techniques like re-implantation of third molars. Non-surgical techniques include allogenic materials and xenografts. This is an update of a review first published in May 2016.

Contents

Abstract
PICOs
Plain language summary
Authors' conclusions
Summary of findings
Background
Objectives
Methods
Results
Discussion
Figures and tables
References

an snippet discussing oral and dental health from Cochrane

"An *oro-antral communication (OAC)* is an unnatural opening between the *oral cavity* and *maxillary sinus*. When it fails to close *spontaneously*, it remains patent and is *epithelialized* to develop into an *oro-antral fistula*. These *complications* occur most commonly during extraction of *upper molar and premolar teeth* (48%)."

an snippet discussing oral and dental health from Cochrane

Cochrane Database of Systematic Reviews | Review - Intervention

New search

Interventions for treating oro-antral communications and fistulae due to dental procedures

Salian Kiran Kumar Krishanappa, Prashanti Eachempati, Sumanth Kumbargere Nagraj, Naresh Yedthare Shetty, Soe Moe, Himanshi Aggarwal, Rebecca J Mathew Authors' declarations of interest

Version published: 16 August 2018 Version history

https://doi.org/10.1002/14651858.CD011784.pub3

Collapse all Expand all

Abstract

Available in English | Español | فارسی | Português | ภาษาไทย | 简体中文

Background

An oro-antral communication is an unnatural opening between the oral cavity and maxillary sinus. When it fails to close spontaneously, it remains patent and is epithelialized to develop into an oro-antral fistula. Various surgical and non-surgical techniques have been used for treating the condition. Surgical procedures include flaps, grafts and other techniques like re-implantation of third molars. Non-surgical techniques include allogenic materials and xenografts. This is an update of a review first published in May 2016.

Contents

“*An oro-antral communication (OAC) is an unnatural opening between the oral cavity and maxillary sinus. When it fails to close spontaneously, it remains patent and is epithelialized to develop into an oro-antral fistula. These complications occur most commonly during extraction of upper molar and premolar teeth (48%).*”

Review language : English | Website language : English | Sign In

Cochrane Library

Trusted evidence.
Informed decisions.
Better health.

Title Abstract Keyword

Browse | Advanced search

Cochrane Reviews | Searching for trials | Clinical Answers | About | Help | About Cochrane

Cochrane Database of Systematic Reviews | Review - Intervention | New search

Interventions for treating oro-antral communications and fistulae due to dental procedures

Salian Kiran Kumar Krishanappa, Prashanti Eachempati, Sumanth Kumbargere Nagraj, Naresh Yedthare Shetty, Soe Moe, Himanshi Aggarwal, Rebecca J Mathew   Authors' declarations of interest

Version published: 16 August 2018   Version history

https://doi.org/10.1002/14651858.CD011784.pub3

Collapse all   Expand all

Download PDF

Cite this Review

Print | Comment | Share | Follow

31 | Cited in 1 guideline

Abstract

Available in English | Español | فارسی | Português | ภาษาไทย | 简体中文

Background

An oro-antral communication is an unnatural opening between the oral cavity and maxillary sinus. When it fails to close spontaneously, it remains patent and is epithelialized to develop into an oro-antral fistula. Various surgical and non-surgical techniques have been used for treating the condition. Surgical procedures include flaps, grafts and other techniques like re-implantation of third molars. Non-surgical techniques include allogenic materials and xenografts. This is an update of a review first published in May 2016.

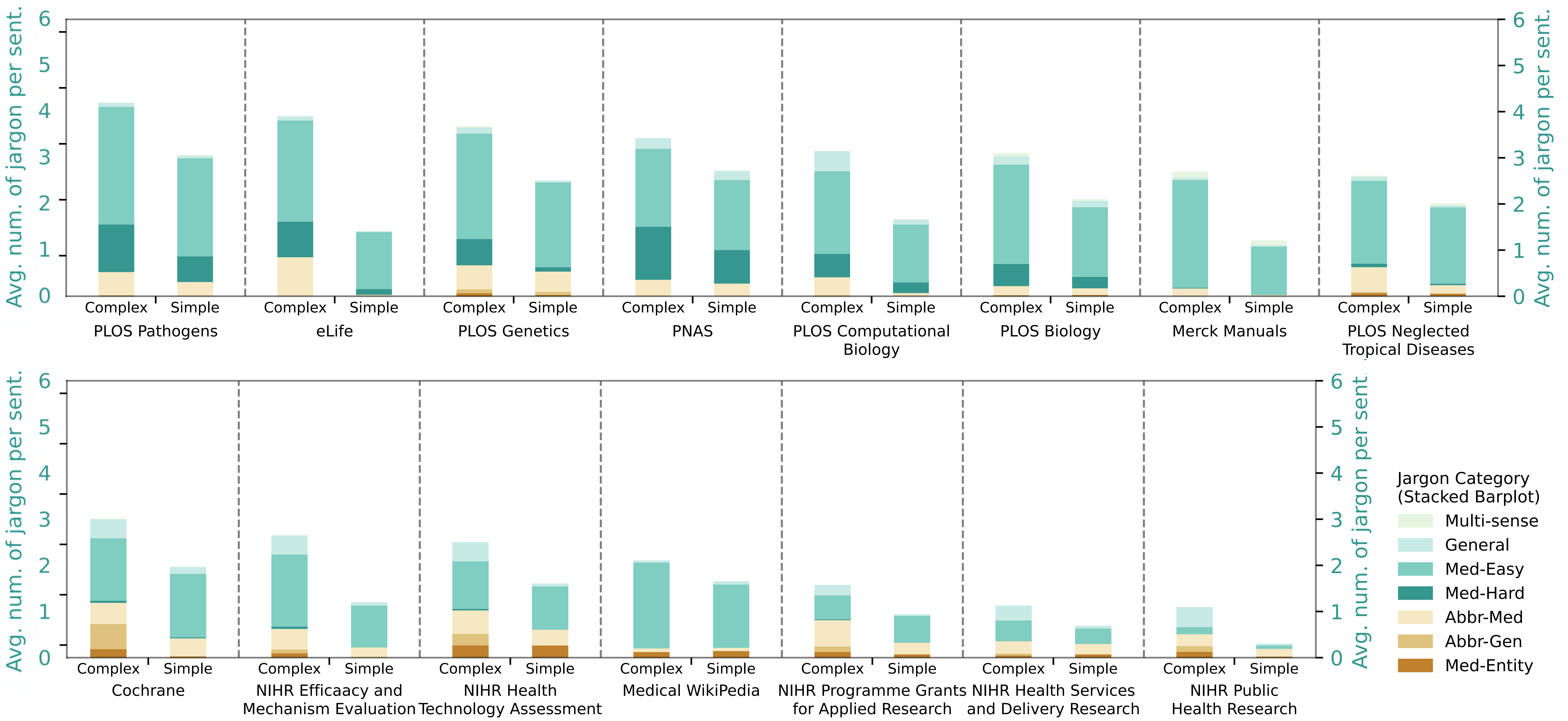
Contents

Abstract
PICOs
Plain language summary
Authors' conclusions
Summary of findings
Background
Objectives
Methods
Results
Discussion
Figures and tables
References

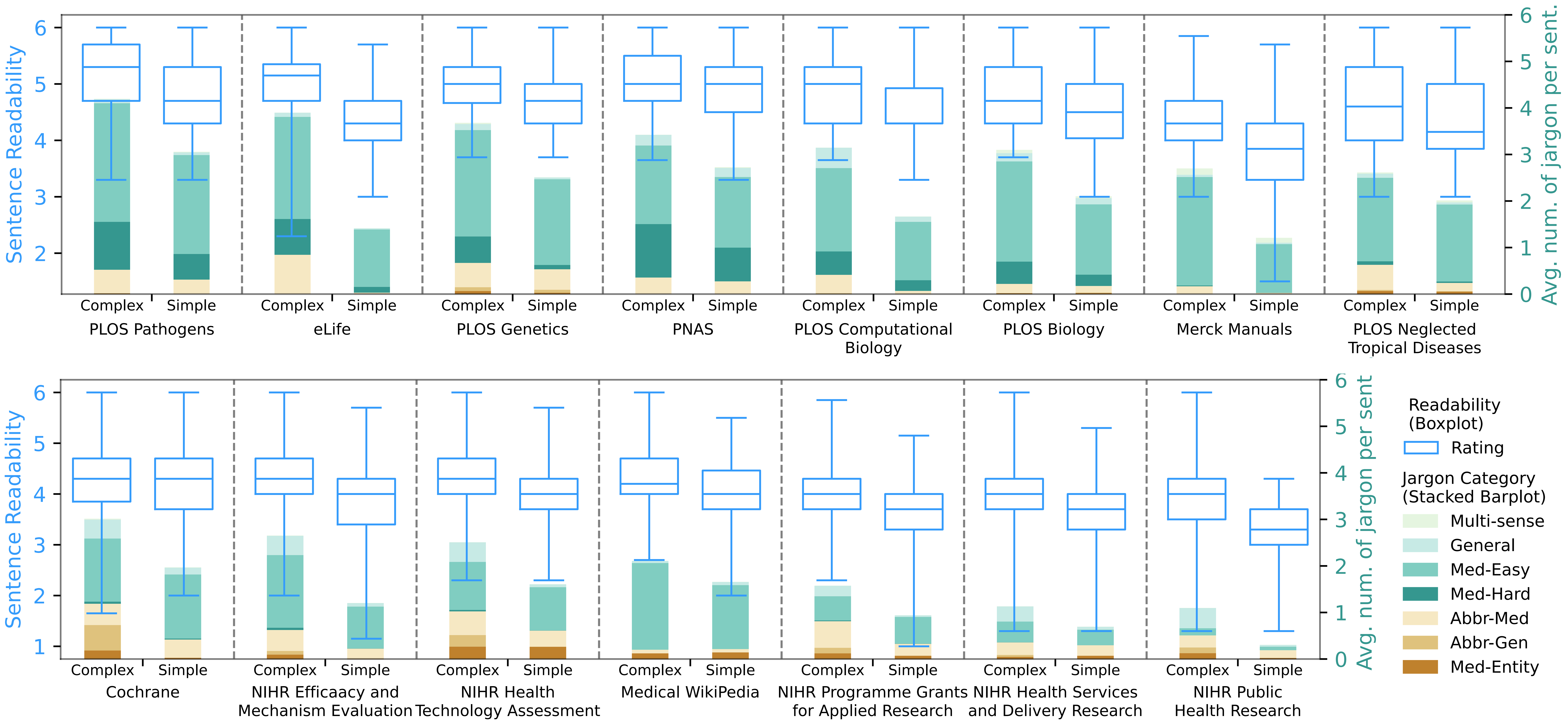
not all jargon and complex terms are equally difficult

Medical - Google Hard	Abbreviations
Medical - Google Easy	General Complex Words

Different Biomedical Data Sources also Vary



Different Biomedical Data Sources also Vary



Rank-and-Rate Annotation Framework

Rank and Rate Sentences on Readability

Signed in as [redacted]

Sign out

Batch ID: [redacted]

Submit and Continue

3

3
3-
3+

Jean Valjean remained silent, motionless, with his back towards the door, seated on the chair from which he had not stirred, and holding his breath in the dark.

These bead-like structures are called nucleosomes, and interactions between histones in different nucleosomes can link one nucleosome to another, to package the DNA into a very condensed form.

+ Context

In a sketch or outline drawing, lines drawn often follow the contour of the subject, creating depth by looking like shadows cast from a light in the artist's position.

+ Context

The long-term functional outcomes of early administration of RDI of amino acids and the use of SMOFlipid, including neurodevelopment, body composition and metabolic health, should be evaluated.

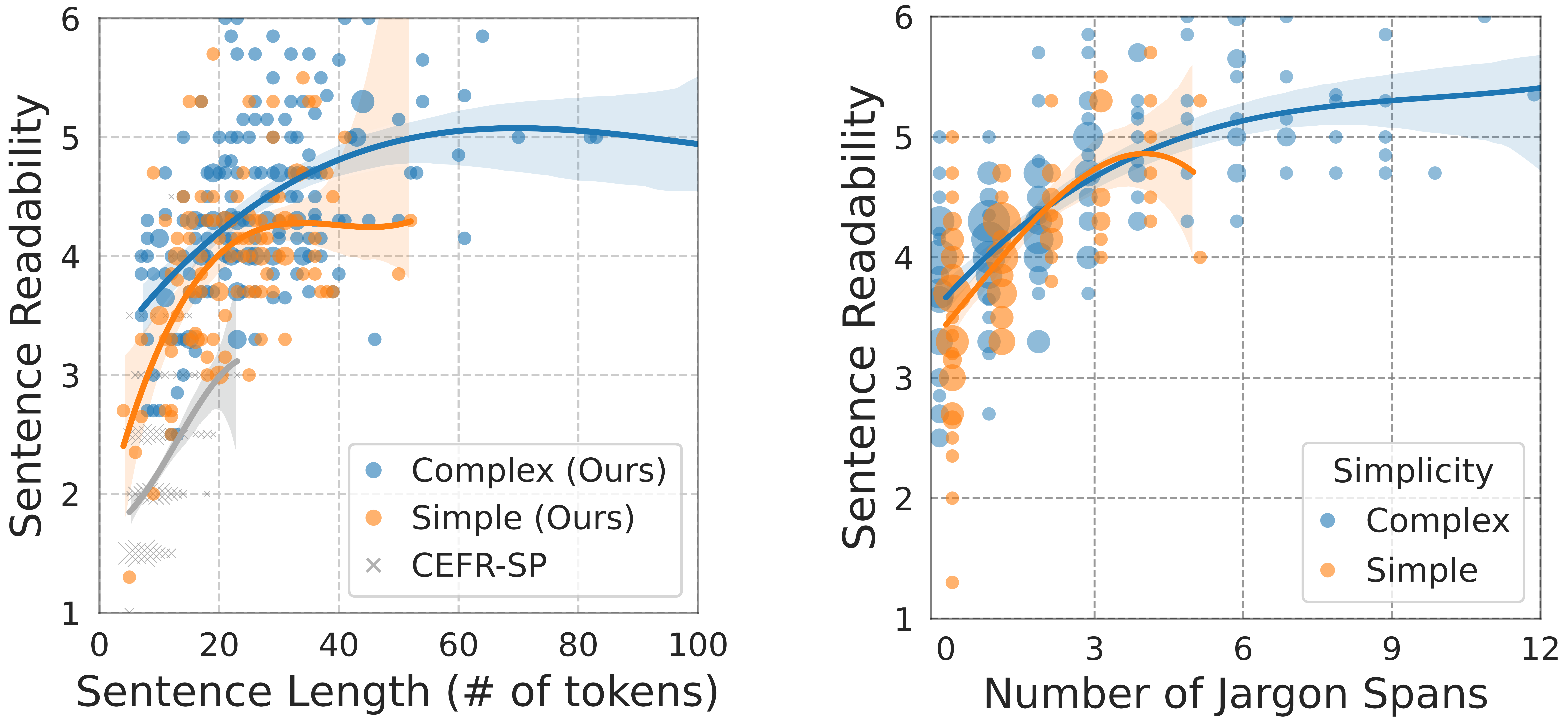
+ Context

All these initiatives take hold as they do, from lead pipes being removed from schools and homes, to new factories being built in communities with a resurgence of American manufacturing.

+ Context

The illumination of the subject is also a key element in creating an artistic piece, and the interplay of light and shadow is a valuable method in the artist's toolbox.

Jargon Greatly Affects Readability



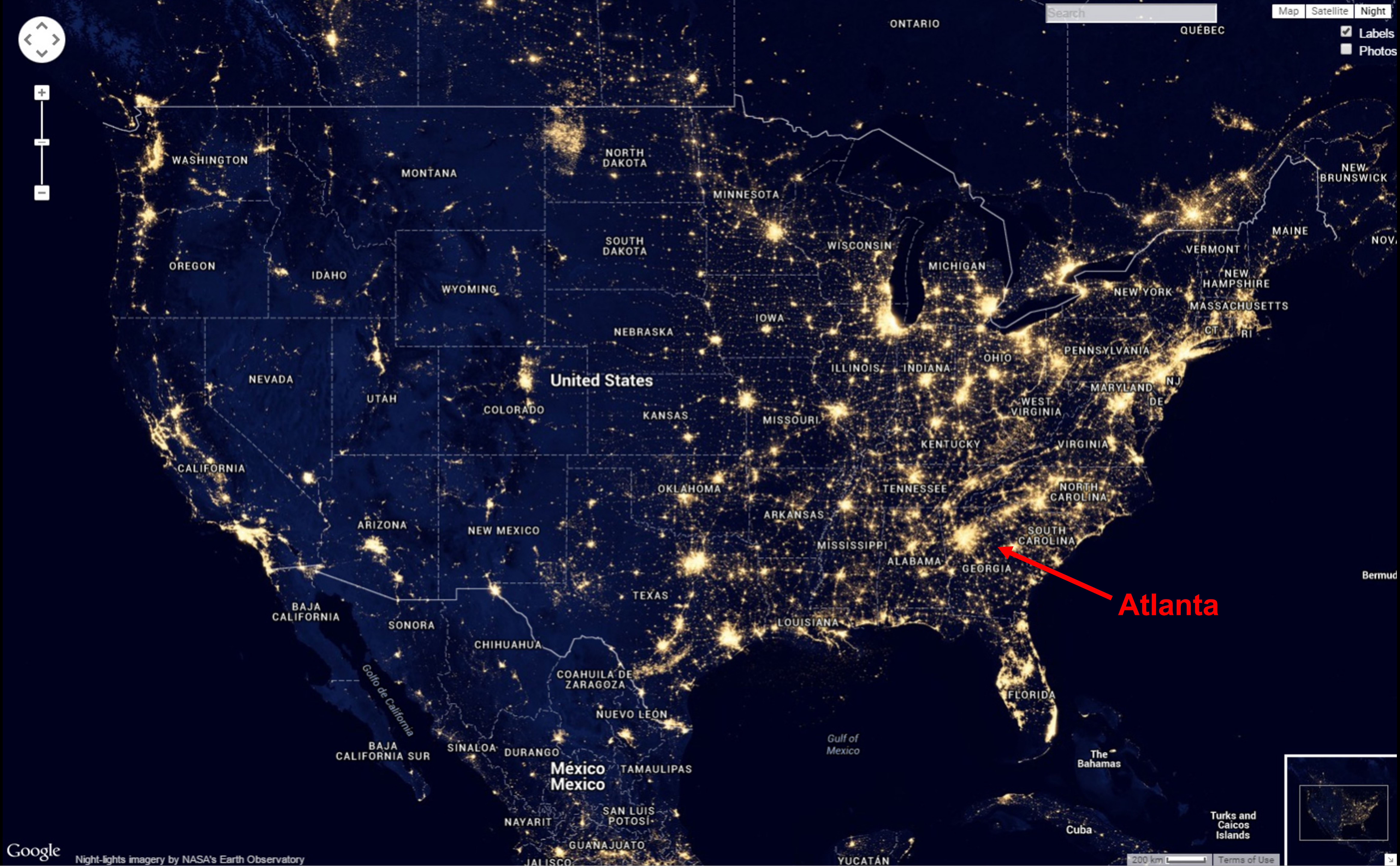
Medical Sentence Readability Measurements

ReadMe++Jar = RoBERTa-large (fine-tuned on ReadMe++) + $\alpha \times$ *# Jargon*

Sources	5-shots		Trained on Each Corpus				The Trained + an Jargon Term			
	GPT-4 (Achiam et al.)	Llama 2-7b (Touvron et al.)	ReadMe++ (Naous et al.)	CEFR-SP (Arase et al.)	CompDS (Brunato et al.)	MEDREADME (Ours)	ReadMe++$_{\text{Jar}}$ (Ours)	CEFR-SP$_{\text{Jar}}$ (Ours)	CompDS$_{\text{Jar}}$ (Ours)	MEDREADME$_{\text{Jar}}$ (Ours)
Cochrane	0.908	0.549	0.858	0.899	0.870	0.947	0.842	0.850	0.785	0.882
PNAS	0.780	0.574	0.852	0.820	0.791	0.874	0.780	0.824	0.744	0.873
NIHR Series	0.713	0.580	0.824	0.753	0.706	0.885	0.697	0.687	0.634	0.700
eLife	0.538	0.127	0.594	0.715	0.608	0.712	0.812	0.802	0.777	0.861
PLOS Series	0.672	0.309	0.680	0.691	0.635	0.702	0.787	0.843	0.744	0.850
Wiki	0.670	0.429	0.824	0.709	0.607	0.843	0.712	0.619	0.673	0.709
MSD	0.766	0.328	0.784	0.778	0.757	0.867	0.918	0.880	0.863	0.937
Mean ± Std	0.721 ± 0.115	0.414 ± 0.17	0.774 ± 0.1	0.766 ± 0.073	0.711 ± 0.101	0.833 ± 0.092	0.793 ± 0.076	0.786 ± 0.096	0.746 ± 0.075	0.830 ± 0.090

Table 7: Pearson correlation (↑) between human ground-truth readability and each **prompting** and **supervised** readability metric. All numbers are averaged over five runs, and all correlations are statistically significant. denotes RoBERTa-large models. "**-Jar**" means adding a "jargon" term (more details in §4.2). Prompt-based methods are competitive, while still outperformed by fine-tuned models in much smaller sizes.

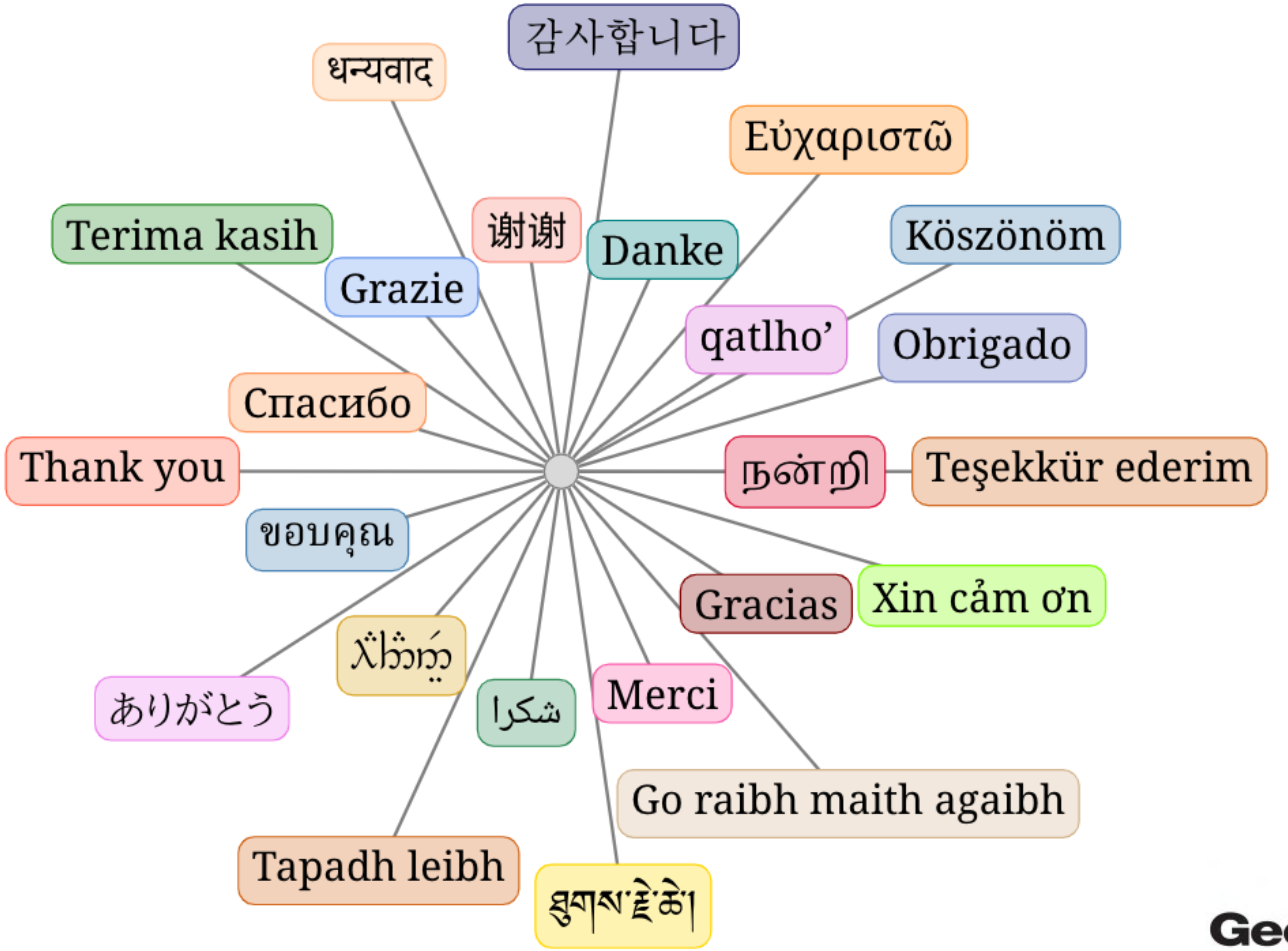

WESTIN
GP
EQUITABLE
TECH


Map
Satellite
Night
Labels
Photos
Search
ONTARIO
QUÉBEC
WASHINGTON
MONTANA
NORTH DAKOTA
MINNESOTA
NEW BRUNSWICK
MAINE
OREGON
IDAHO
WYOMING
SOUTH DAKOTA
WISCONSIN
MICHIGAN
VERMONT
NEW HAMPSHIRE
NEW YORK
MASSACHUSETTS
CT
RI
IOWA
NEBRASKA
PENNSYLVANIA
NJ
NEVADA
UTAH
United States
COLORADO
KANSAS
ILLINOIS
INDIANA
OHIO
MARYLAND
DE
WEST VIRGINIA
MISSOURI
KENTUCKY
VIRGINIA
CALIFORNIA
OKLAHOMA
TENNESSEE
NORTH CAROLINA
ARKANSAS
SOUTH CAROLINA
ARIZONA
NEW MEXICO
MISSISSIPPI
ALABAMA
GEORGIA
Atlanta
TEXAS
Bermuda
BAJA CALIFORNIA
SONORA
LOUISIANA
CHIHUAHUA
COAHUILA DE ZARAGOZA
Golfo de California
FLORIDA
NUEVO LEÓN
Gulf of Mexico
BAJA CALIFORNIA SUR
SINALOA
DURANGO
México
Mexico
TAMAULIPAS
The Bahamas
NAYARIT
SAN LUIS POTOSÍ
Turks and Caicos Islands
Cuba
GUANAJUATO
JALISCO
YUCATÁN
Google
Night-lights imagery by NASA's Earth Observatory
200 km
Terms of Use

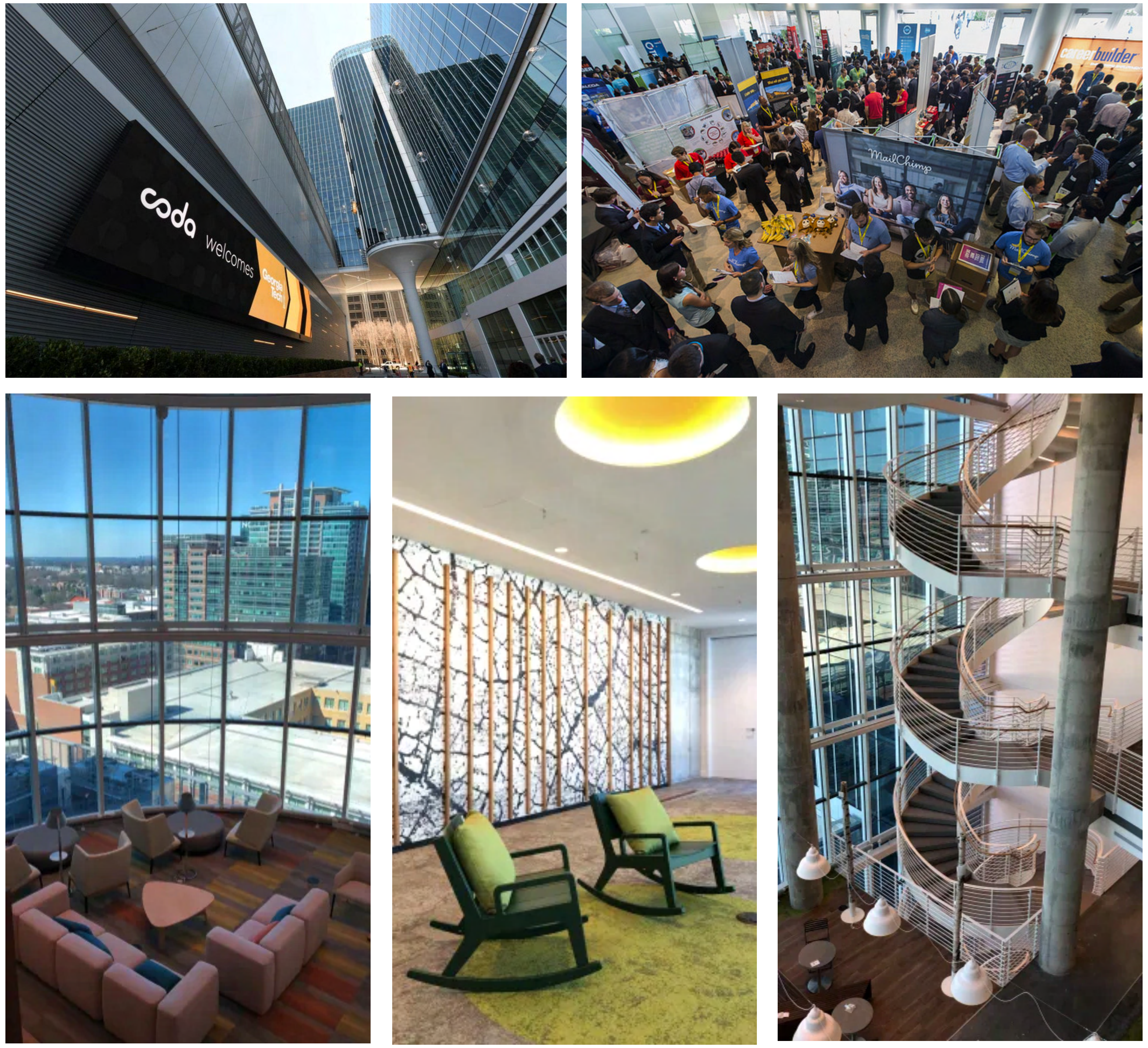

Georgia Tech
at&t

Thank you!

https://cocoxu.github.io/



(image credit: Overleaf!)



(image credit: Georgia Tech)

Georgia Tech Machine Learning
NSF
IARPA
DARPA
NIH
Google
SONY
criteo