

MUNI
FI

Low-Resource Machine Translation

Edoardo Signoroni

- MT is the task of translating a sentence from a source language to the corresponding sentence in the target language.
- Nowadays, it is done with **Neural** Machine Learning systems trained on **parallel corpora**.
- Main issues:
 - Linguistic **ambiguity**
e.g. “ It’s raining cats and dogs. ”
 - **DATA SCARCITY**



MUNI FI

>7000 living languages

plus:

- varieties;
- dialects;
- slangs;
- code-switching;
- code-mixing;
- ... and more

but most of these are “Left-Behinds” or **Low-resource languages**

since the biggest MT system online supports a grand total of

243 (or 3.47%)

What are LRLs?

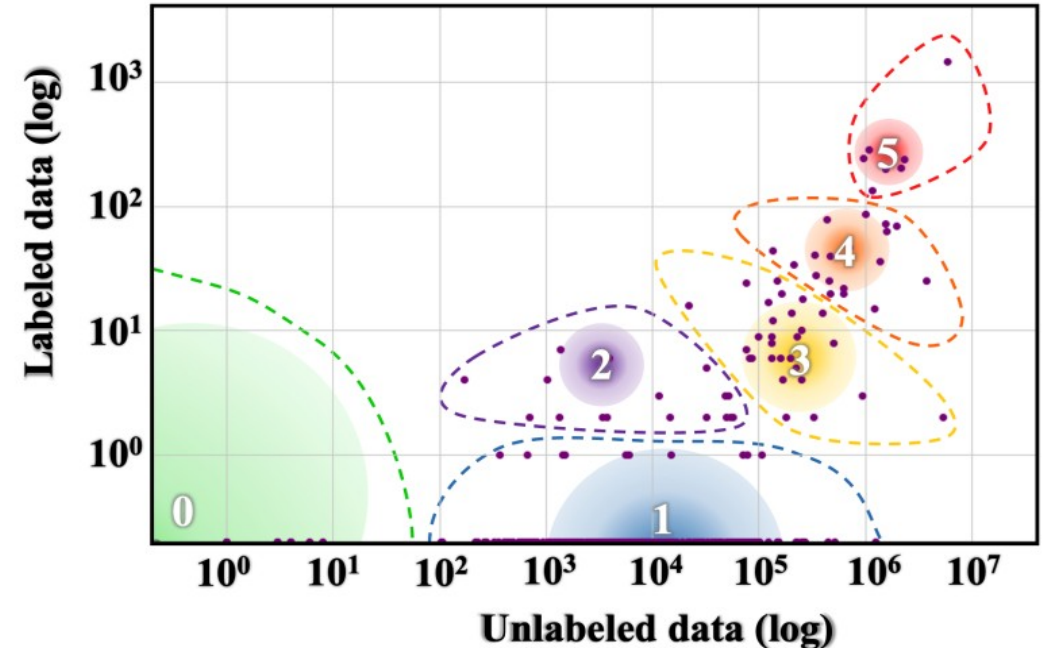


Figure 2: Language Resource Distribution: The size of the gradient circle represents the number of languages in the class. The color spectrum VIBGYOR, represents the total speaker population size from low to high. Bounding curves used to demonstrate covered points by that language class. Blasi et al. (2022), Joshi et al. (2020)

What are LRLs?

Joshi et al. 2020 define LRLs, in incremental classes:

“Have exceptionally limited resources, and have rarely been considered in language technologies.”

“Have some unlabelled data; however, collecting labelled data is challenging.”

“A small set of labelled datasets has been collected, and language support communities are there to support the language.”



What are LRLs?

Table 1. Language Categories Identified by Joshi et al. [93] and Number of Languages per Class

Class	Description	Examples	# langs	#Speakers	% of Total Langs
0	Have exceptionally limited resources, and have rarely been considered in language technologies.	Slovene, Sinhala	2,191	1.2B	88.38%
1	Have some unlabelled data; however, collecting labelled data is challenging.	Nepali, Telugu	222	30M	5.49%
2	A small set of labeled datasets has been collected, and language support communities are there to support the language.	Zulu, Irish	19	5.7M	0.36%
3	Has a strong web presence, and a cultural community that backs it. Have been highly benefited by unsupervised pre-training.	Afrikaans, Urdu	28	1.8B	4.42%
4	Have a large amount of unlabeled data, and lesser, but still a significant amount of labelled data. have dedicated NLP communities researching these languages.	Russian, Hindi Italian, Czech	18	2.2B	1.07%
5	Have a dominant online presence. There have been massive investments in the development of resources and technologies.	English, Japanese	7	2.5B	0.28%

“adapted” from Ranathunga et al. (2023) and Joshi et. al (2020)

MUNI FI

What are LRLs?

For MT:

No standard definition.

Usually LR pair if the size of the parallel corpora is **<500k sentences**

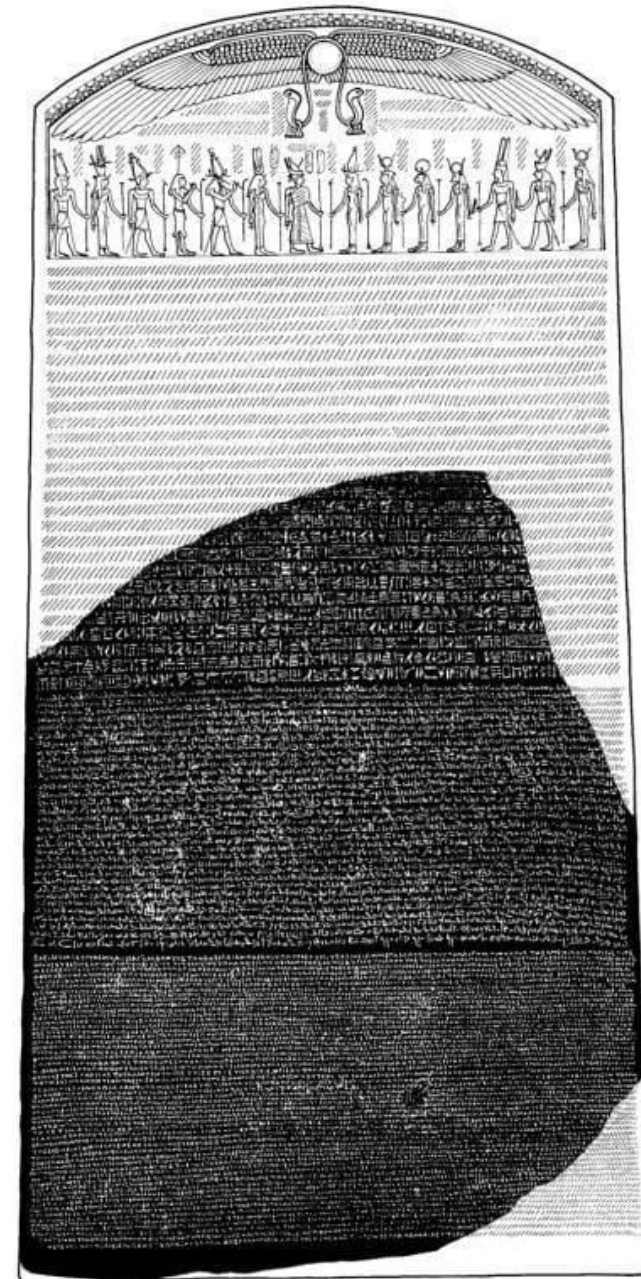
and **extremely LR** below **100k** pairs

if **no data** is available, we enter the **zero-shot** setting

WMT22 deu-dsb 40k sents 500k words

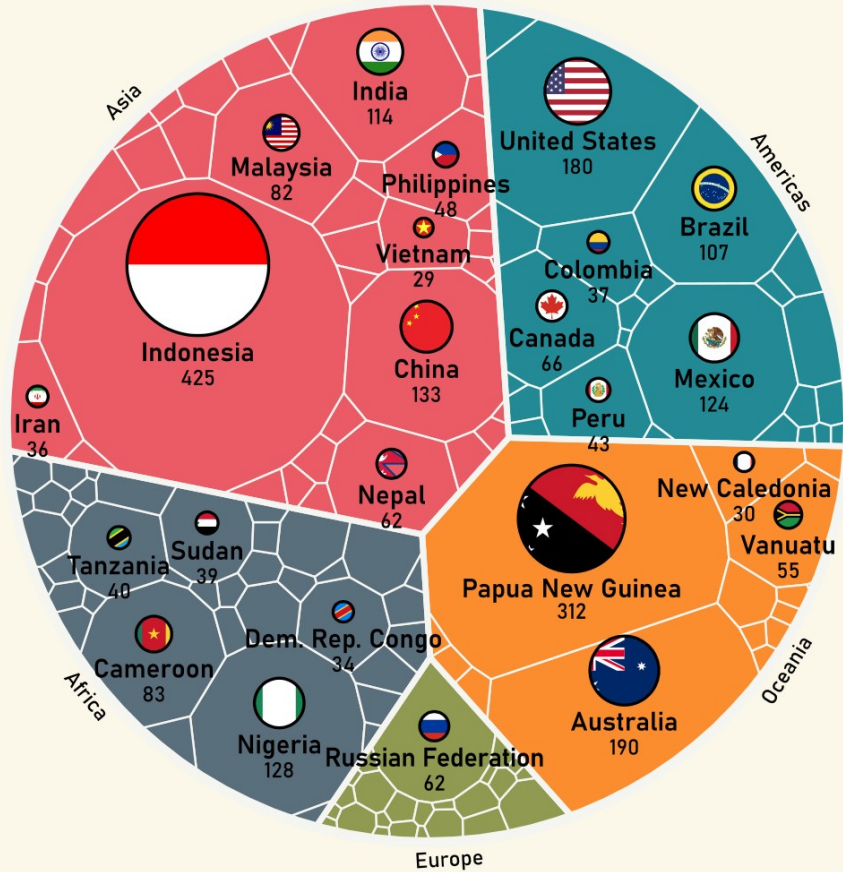
The Good Soldier Švejk 200k words

New Testament 185k words



Global Endangered Languages (2023)

3,078 endangered languages analyzed by continent and country



The top 25 countries account for 2,484 endangered languages (~80%), whereas the Rest of the World accounts for 594 endangered languages (~20%)

Top 10



Why work on LRLs?

decreasing the **digital divide**

<http://labs.theguardian.com/digital-language-divide/>

dealing with **inequalities of information access** and **production**

mitigating **cross-cultural biases**

deploying NLP technologies for **underrepresented** languages

understanding **cross-linguistic differences**

preserving **linguistic diversity**

~3000 (43%) are endangered

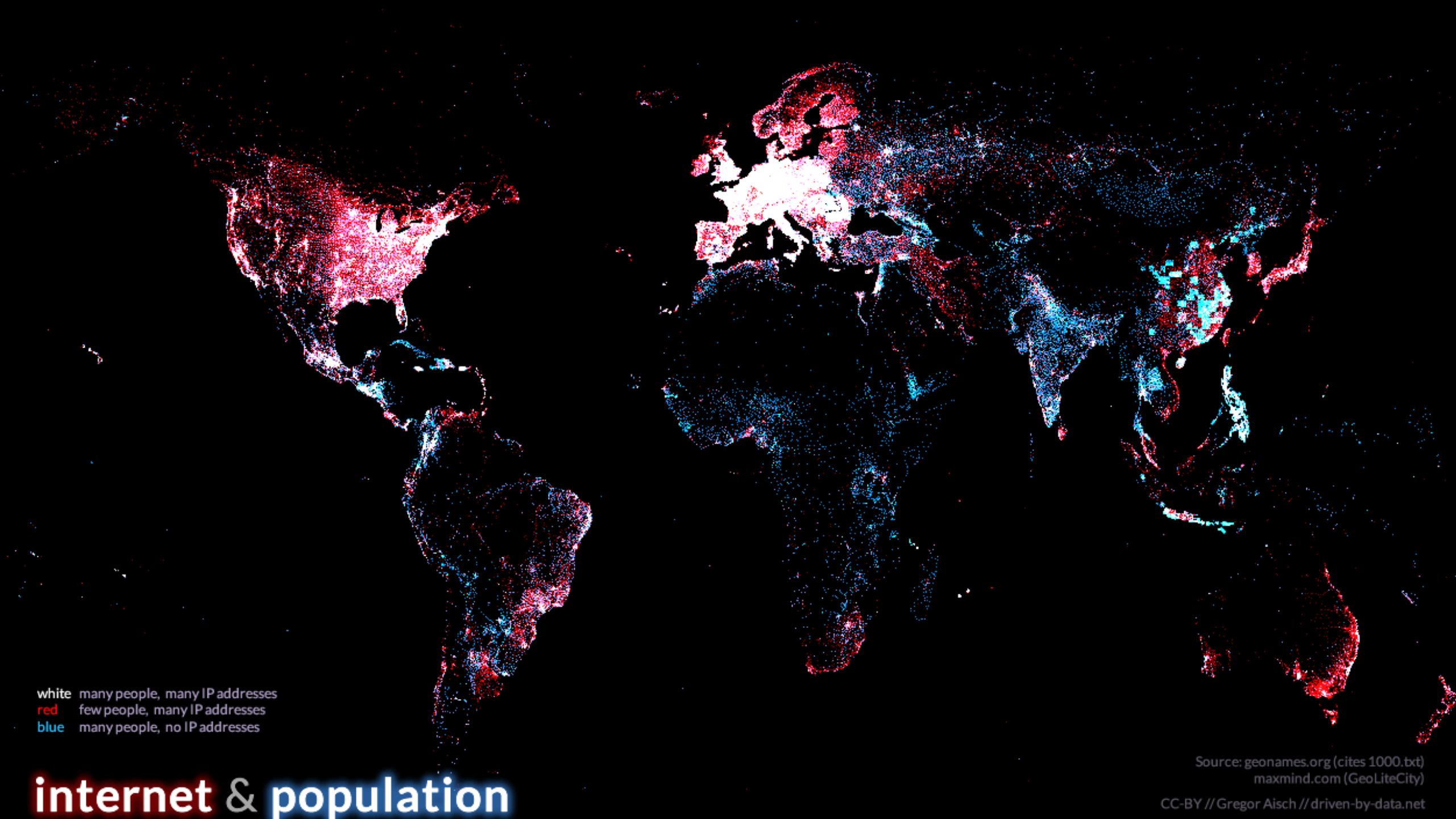
90% of all languages will be extinct within 100 years;

in the best case scenario, only 50% will survive,

and just 10% are considered safe during the next century

https://www.endangeredlanguages.com/about_importance/

Given this variability, always **highlight clearly the languages you are working on** (Bender Rule & Data Statements)



white many people, many IP addresses
red few people, many IP addresses
blue many people, no IP addresses

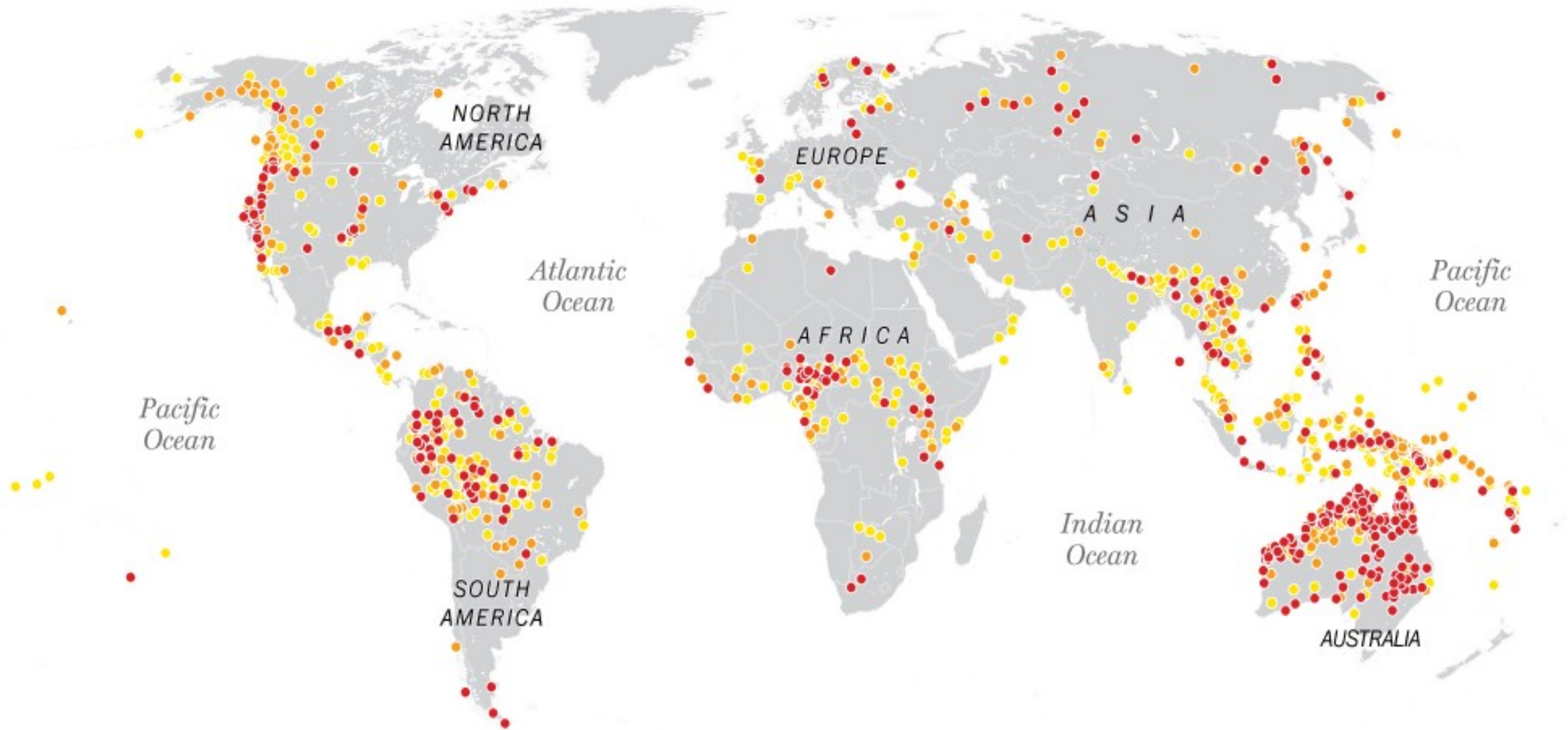
internet & population

Source: geonames.org (cites 1000.txt)
maxmind.com (GeoLiteCity)

CC-BY // Gregor Aisch // driven-by-data.net

At risk languages

● Critically endangered ● Seriously endangered ● Endangered



MUNI

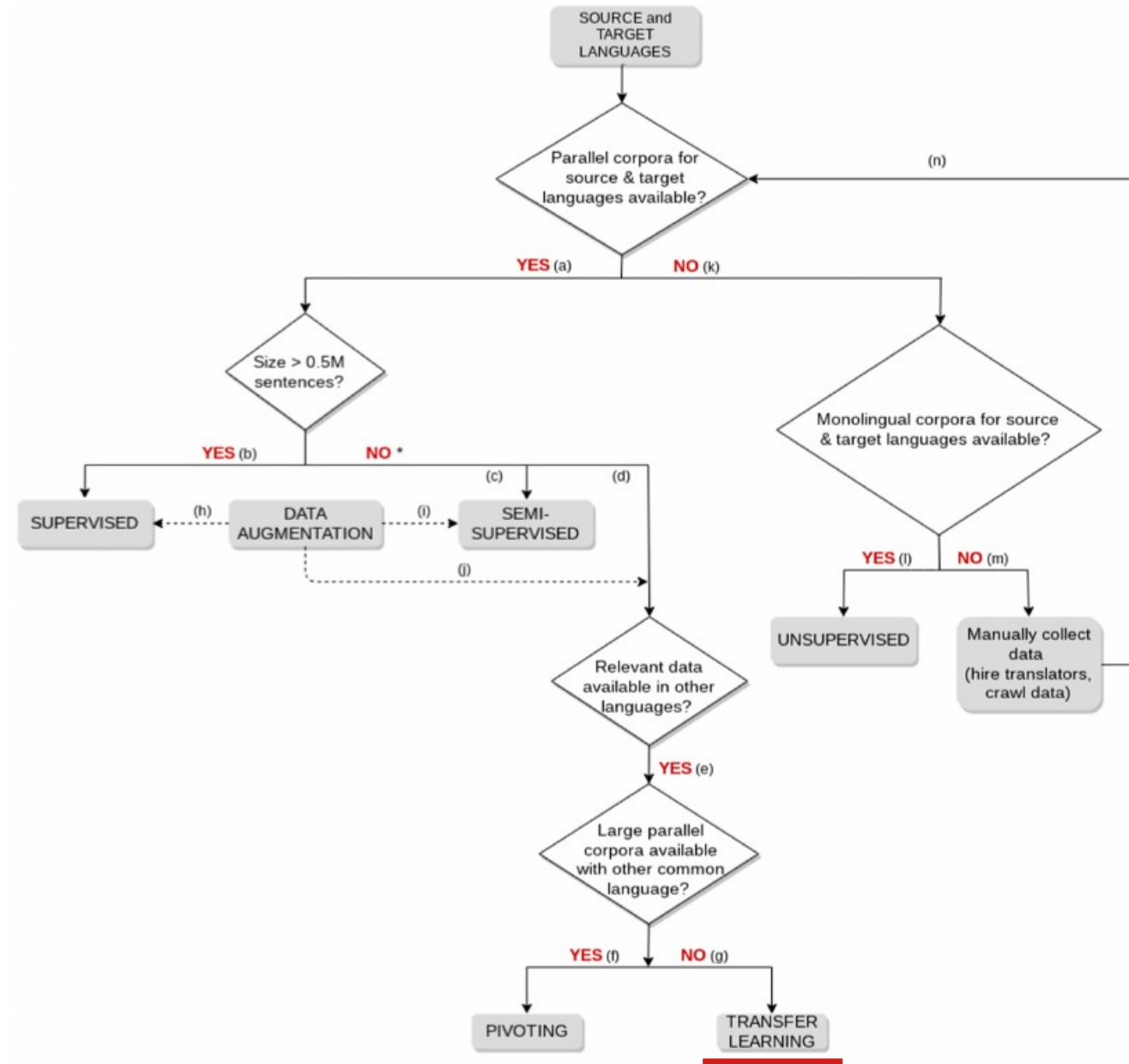
How is it done?

FI

Currently, the **state-of-the-art** for HRLs is **NMT**

Several approaches have been proposed for LRLs, too

But most of the impact can be obtained with **careful** and **clever use of the data** we have

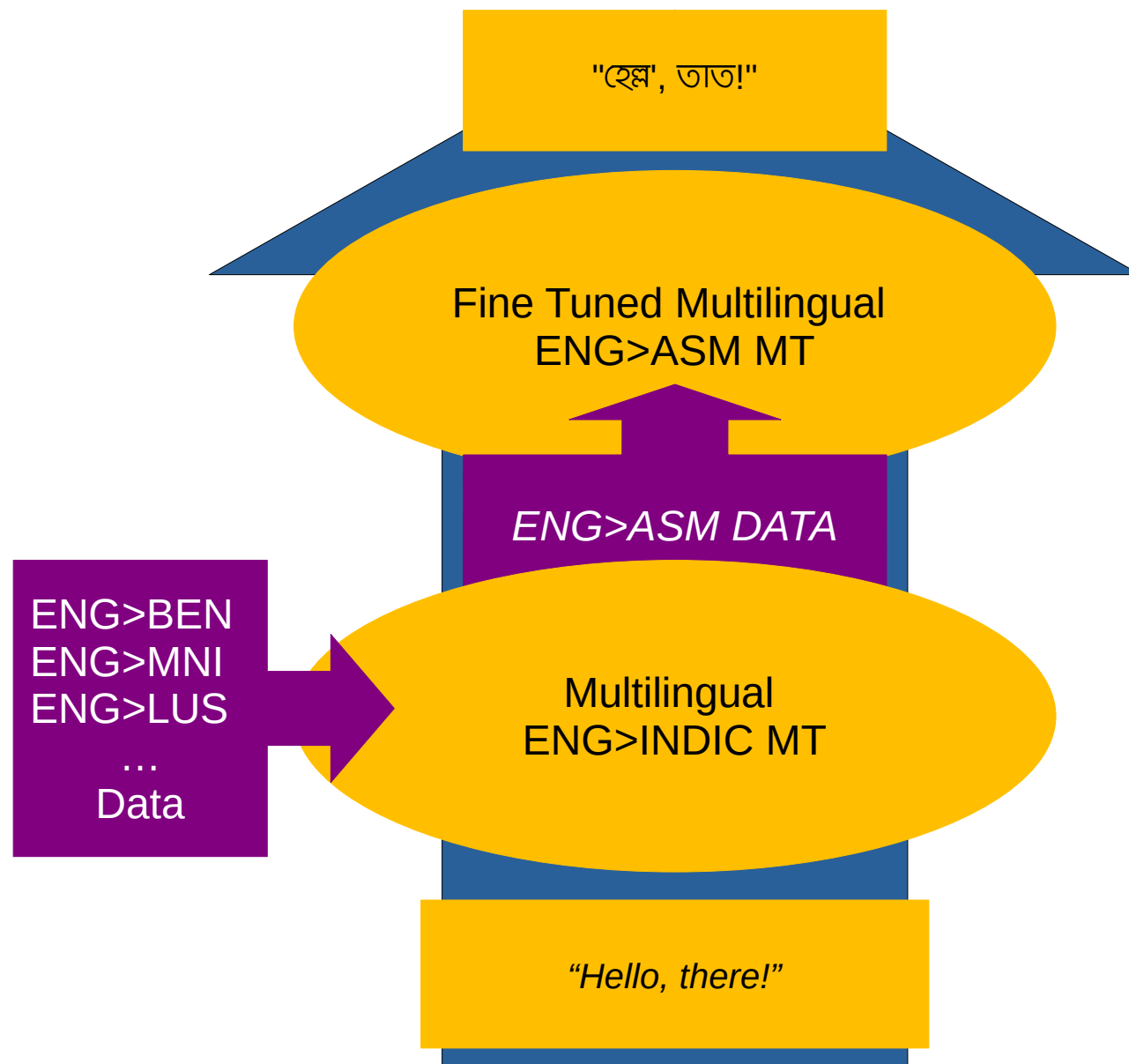


* Assuming monolingual corpora are also available

MUNI FI

How is it done?- Current Methods

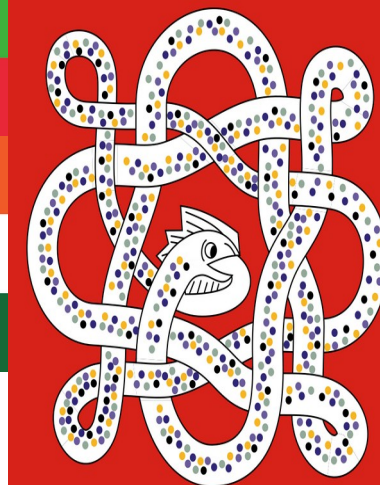
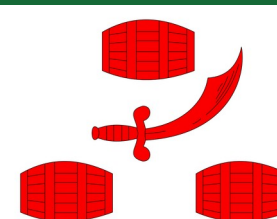
- **Multilingual NMT transfer** learning is the current state-of-the-art
- Best results **using data from related HRL pairs** and fine-tune pre-trained NMT models to the related data or the small amount of LRLs text available
- issues with **performance** and **equitable access**



MUNI FI

Language Relatedness

- It is beneficial to use **related languages** for transfer between HRLs and LRLs
- However, the extent of this is not clear. Which kind of relatedness is the most helpful?
 - **Genealogical?**
ဗြဟ္မစာအုပ် (Burmese, Tibeto-Burman, Burmese) > মৈতৈলোন (Manipuri, Tibeto-Burman)
 - **Typological?**
हिन्दी (Hindi, Indo-Aryan, SOV) > মৈতৈলোন (Manipuri, SOV)
 - **Writing system?**
বাংলা (Bengali, Bengali script) > মৈতৈলোন (Manipuri, Bengali script)
- How can we better leverage and disentangle these factors?



MUNI FI

An Example: WMT23 Indic LR MT

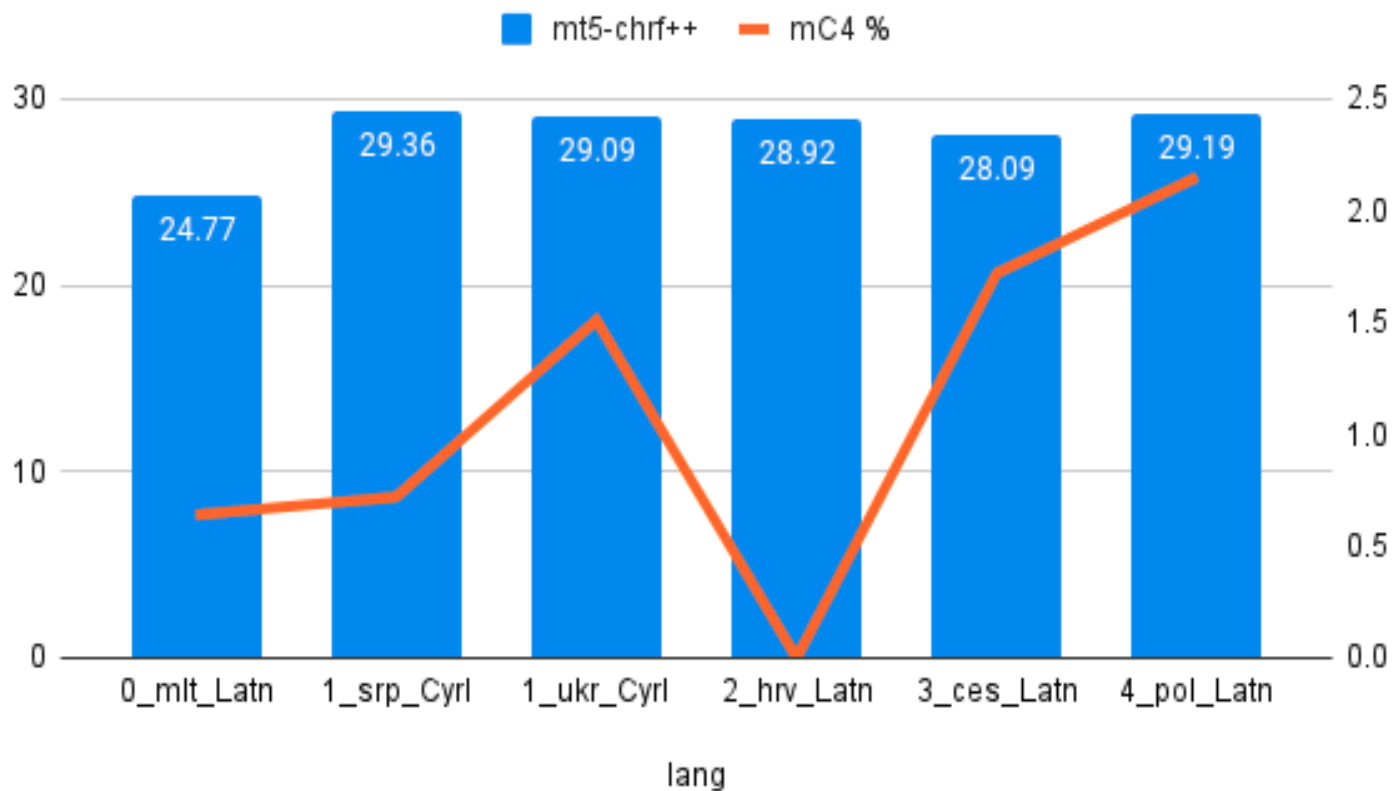
- 4 Low-res Indic languages (*asm,kha,lus,mni*) \leftrightarrow English
- Collated train datasets on a same-script basis (*asm&mni; kha&lus*), and for *all* languages together
- Trained systems on the collated data, and fine-tuned child systems for the single directions
- Best option for *kha;lus>eng*. *Mni>eng* was better with same-script parent



MUNI FT

Zero-shot and relatedness

Fine-grained relatedness



Most of the times, **no data** for the LRL are available
→ **Zero-shot**

Fine-tuning a pre-trained LLM with data from **a related language helps** (e.g. Slavic language into Silesian)

However, the internal, fine-grained relatedness of the language, or its presence in the pre-training data seems not to matter

MUNI FI

Tokenization

- A MT system is a **sequence-to-sequence** model, which takes words in input and generates words as output
- Thus it needs a **vocabulary** of tokens, words in the most simple implementation
- Dealing with **morphological variants** and **variation** leads to huge vocabulary sizes and out-of-vocabulary words, not seen in training

A L U M N U S

A L U M N A E

A L U M N I

MUNI FI

- Text is segmented into **subwords** with data-driven iterative algorithms
- These are combined together to deal with unknown words, but still struggle with **complex morphology, non-standard forms, linguistic diversity, ...**
- *Character, hybrid, token-free, and even pixel-level* approaches have been proposed to overcome such challenges

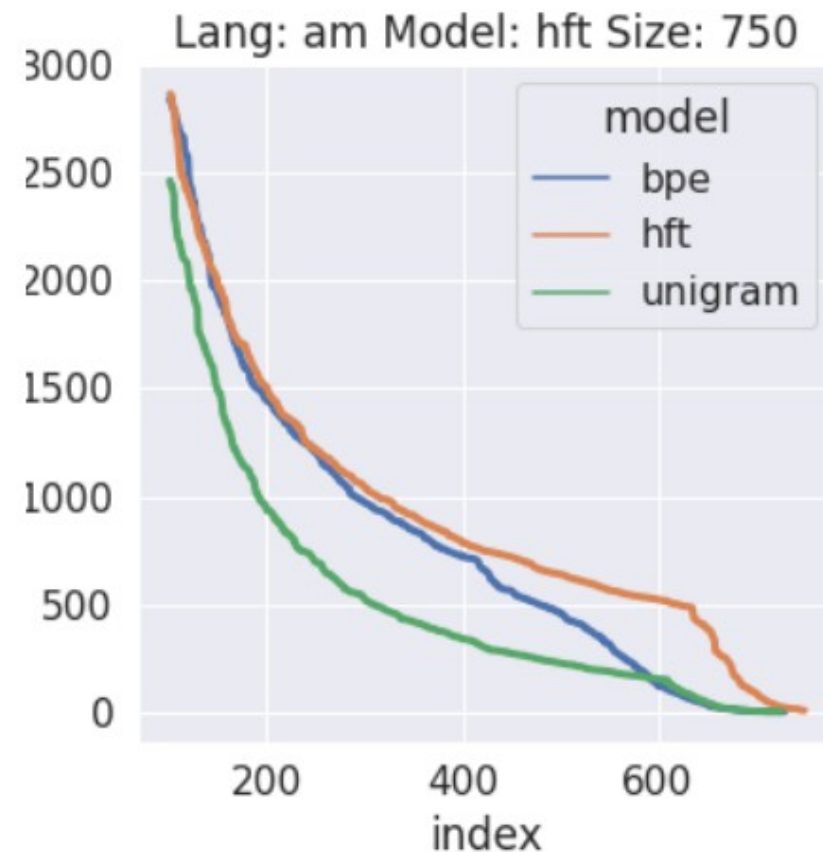
Tokenization

Word Length	2	3	4	5	6
6 letter words					
	sorbus ₁₀	swords ₁₀	sudors ₈		

5 letter words
brows ₁₁ swobs ₁₁ burds ₁₀ drubs ₁₀ sword ₉ words ₉
sorbs ₈ duros ₇ sudor ₇ surds ₇ dross ₆ sords ₆ sorus ₆
sours ₆

4 letter words
bows ₁₀ brow ₁₀ swob ₁₀ budo ₉ buds ₉ burd ₉ drub ₉
dubs ₉ bods ₈ burs ₈ buss ₈ dobs ₈ dows ₈ rubs ₈
subs ₈ urbs ₈ word ₈ wuss ₈ boss ₇ bros ₇ orbs ₇ robs ₇
rows ₇ sob ₇ sorb ₇ sows ₇ wors ₇ dour ₆ duos ₆
duro ₆ ouds ₆ suds ₆ surd ₆ udos ₆ urds ₆ dors ₅ doss ₅

Tokenization impacts the quality of downstream **NMT**, especially for LRLs, thus choosing its parameters carefully is crucial.

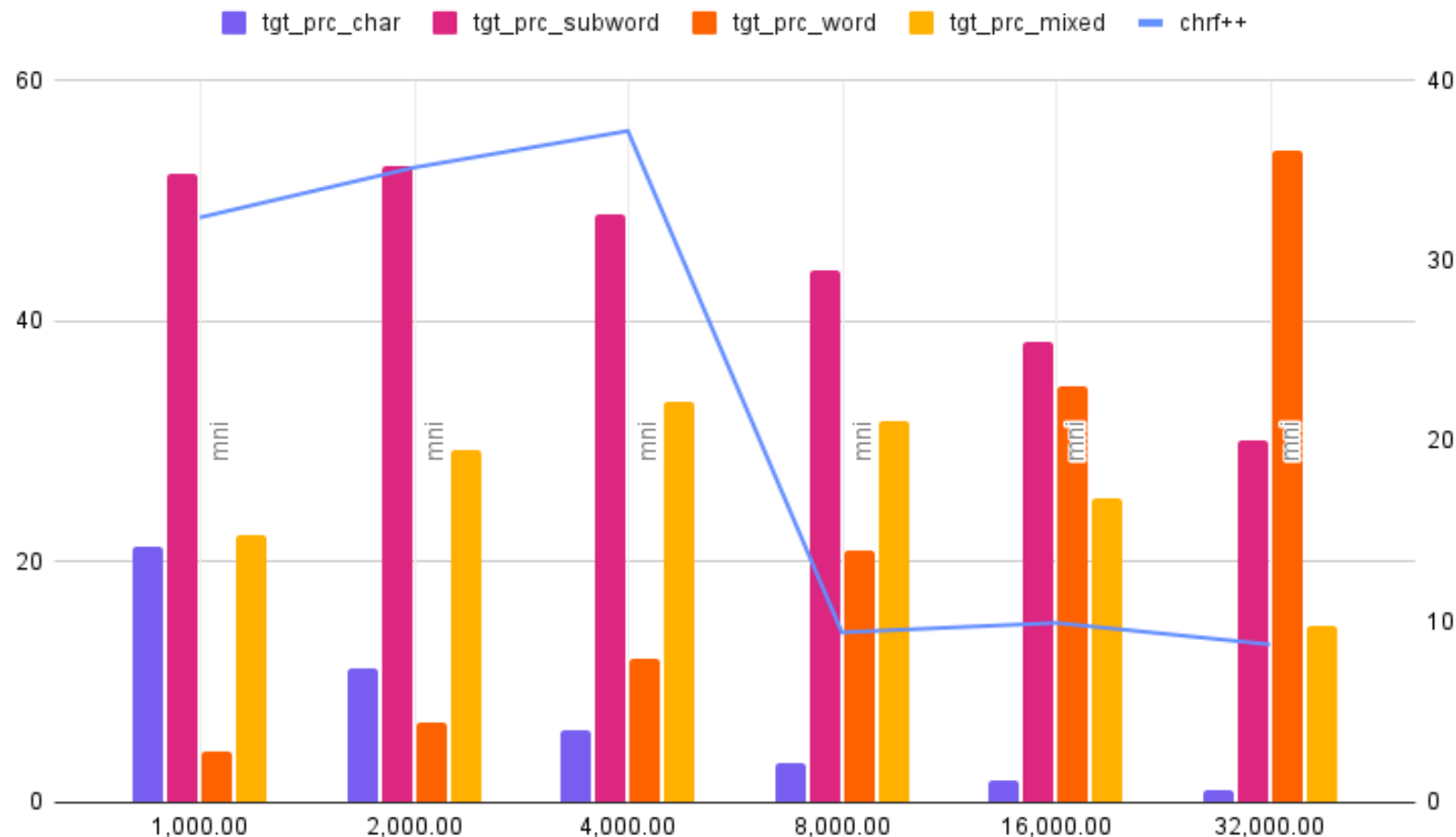


An Example: WMT22 LR MT

- 2 LRLs: *Lower & Upper Sorbian*
- by using a custom our custom HFT tokenizer to obtain more frequent and thus better represented tokens, we outperformed the default bpe approach using only the given LR (40k, 450k) parallel corpora

	DSB-DE	DE-DSB	DSB-HSB	HSB-DSB
t-bpe	27.92	22.74	72.01	69.71
t-hft	34.20	30.86	72.21	70.71
t-opt-bpe	29.75	25.06	71.37	69.50
t-opt-hft	35.46	31.12	71.83	68.95
t-bpe-dd	33.02	28.54	73.47	71.98
t-hft-dd	38.42	33.53	73.53	71.59

Tokenization Impact on NMT

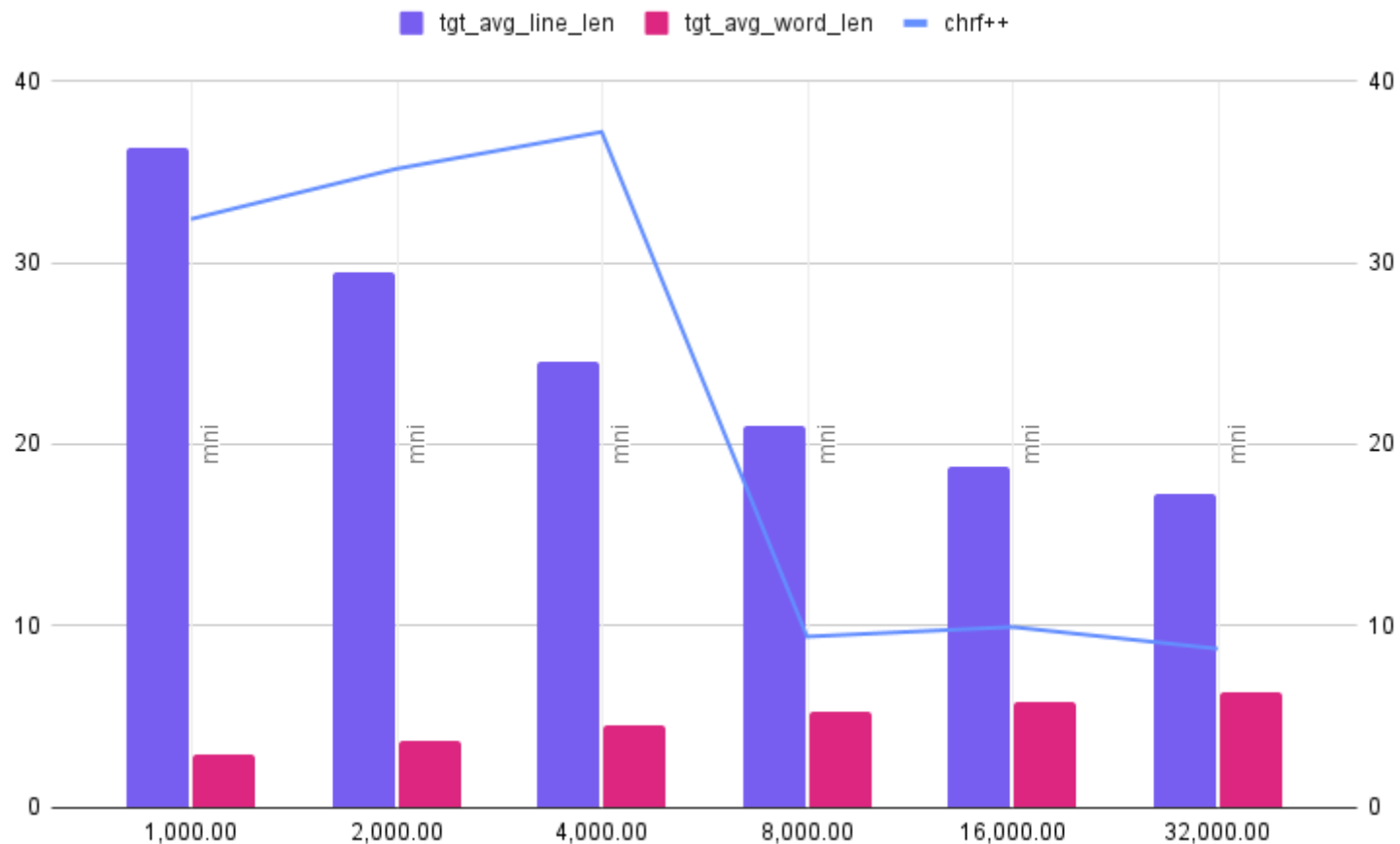


As we set an higher value for the vocabulary size, we get:

- Less characters
- Slightly more subwords, and more mixed-use tokens at first
- More full words
- But also less quality

A more “balanced” mix of characters, subwords, and words generalizes better to unseen data than a word-heavy vocabulary

Tokenization Impact on NMT



As we set an higher value for the vocabulary size, we naturally get longer tokens and shorter lines

250 __A mb as s ad or __M s . __N ik ki _H al ey , __U n it ed _S t at es _P er m an ent _R ep res ent at ive __to __the __U n it ed __N ation s

500 __A mb ass ad or __M s . __N ik ki _H al ey , __U n it ed _S t at es _P er m an ent _R ep res ent at ive __to __the __U n it ed __N ations

1k __A mb ass ad or __M s . __N ik ki _H al ey , __U n it ed _S t at es _P er m an ent _R ep res ent at ive __to __the __U n it ed __N ations

2k __A mb ass ad or __M s . __N ik ki _H al ey , __U n it ed _S t at es _P er m an ent _R ep res ent at ive __to __the __U n it ed __N ations

4k __A mb ass ad or __M s . __N ik ki _H al ey , __U n it ed _S t at es _P er m an ent _R ep res ent at ive __to __the __U n it ed __N ations

8k __A mb ass ad or __M s . __N ik ki _H al ey , __U n it ed _S t at es _P er m an ent _R ep res ent at ive __to __the __U n it ed __N ations

16k __A mb ass ad or __M s . __N ik ki _H al ey , __U n it ed _S t at es _P er m an ent _R ep res ent at ive __to __the __U n it ed __N ations

32k __A mb ass ad or __M s . __N ik ki _H al ey , __U n it ed _S t at es _P er m an ent _R ep res ent at ive __to __the __U n it ed __N ations

MUNI

FI

- Pre-trained models use huge vocabularies to account for all of the training data, and require heavy computational resources
- If carefully tuned, “traditional” trained-from-scratch systems can achieve meaningful representation at a fraction of the computational size and cost, even in extremely LR conditions
- In particular, smaller vocabulary sizes, most often lead to:
 - better MT performance
 - Smaller model size
 - Faster training times

Smaller Vocabularies

Source	Target	Voc. Size	ChrF	Params	sec/epoch
eng	akk	1k	34.425	7.885	27.987
		2k	26.764	8.393	23.200
		4k	20.438	9.398	21.514
		8k	19.530	11.373	21.622
		16k	12.472	15.104	21.857
		32k	13.603	21.115	25.844
dsb	deu	1k	35.755	7.889	15.123
		2k	38.011	8.397	13.627
		4k	42.621	9.417	13.219
		8k	45.434	11.434	13.313
		16k	45.468	15.387	13.402
		32k	45.248	22.784	15.807
eng	mni	1k	36.171	7.885	16.452
		2k	38.952	8.395	14.905
		4k	40.878	9.400	14.203
		8k	10.604	11.366	14.63
		16k	11.090	15.03	14.839
		32k	9.685	20.363	17.276

MUNI Automated Metrics for LR MT

FI

- Automated metrics allow for low-cost, fast comparison of system
- Two types are relevant for LR MT:

Lexical Overlap

- They compare the sequence similarity between the proposed translation and one or more references
- BLEU (Papineni et al. 2002), ChrF (Popovic 2015, 2017)

Neural Metrics

- Fine-tuned LMs on human judgements that predict a score based on a given input of source, translation, and reference.
- COMET, xCOMET (Rei et al. 2020)



MUNI Automated Metrics for LR MT

FI

- While **Neural Metrics** are the state-of-the-art; they **perform poorly** in for LRLs
- **Fine-tuning COMET** models to LRLs was shown to be promising: IndicCOMET (Sai B et al. 2023); AfriCOMET (Wang et al. 2023)
- If this is not possible, **ChrF(++)** was deemed the best back off metric

MUNI FI

Some Conclusions

- Working on LRLs is important for several **linguistic, social, and democratic** reasons
- Multilingual NMT approaches involving transfer learning are currently the state-of-the-art for LRLs-MT
- but they still have various issues regarding their performance and equitable access
- Careful tuning of the parameters and clever use of the training data goes a long way to alleviate the problems of LR MT
- Some best practices, such as highlighting the LRLs studied and using fitting metric to evaluate the output of MT are also important