Multilinguality

CS 6120 Natural Language Processing

Northeastern University

Si Wu

Logistics

- Data and experimental design due tonight
- The dates of the next two quizzes are on the class website
  - 11/7, 11/18
- Next Tuesday, Prof. Blevins will give a guest lecture in the other session
  - The lecture will be recorded, so we won't have class that day. You can watch the recording at home.
- Today:
  - Multilinguality
    - Prof. Blevins will also talk about this topic!

Diversity of human languages

- There are approximately 7000 language being spoken today
  - No official data about how many written languages
- Approx. 140 language family according to *Ethnologue*
- Many of these statistics are not including **sign languages**, which are actually legitimate human languages by linguistic definition!

Multilingualism and multi-culturalism

- According to UNESCO, estimated **50% to 66% of the world population use two or more language in their daily life**
  - One can argue that multilingualism should be a norm!
- Many countries have more than 1 official languages
  - One of them is our friendly neighbor/neighbour Canada!
- Languages frequently borrow and adopt words from one another
- Even when speaking the same language, they can have different cultures:
  - UK vs USA, Mexico vs Spain, French vs Canadian French…
- Within a language, there are many different dialects and sociolects

Photo from https://www.hongkonghike.com/7-hong-kong-road-signs-that-are-accidentally-hilarious/

Quick note on sign languages

According to the National Association of the Deaf:

- 90% of the deaf population has two hearing parents
- Most of those parents do not know sign language…

Sign language is a language by linguistic definition.

→ Good resource to learn more about Sign Language Processing:
https://research.sign.mt/

Code switching

- The practice of alternating between two or more languages (or dialects) in a conversation
- People do it for many reasons: such as group identity, solidarity and gratitude, unconscious effort, to fit it, etc.

Examples of code-switching in Hong Kong:

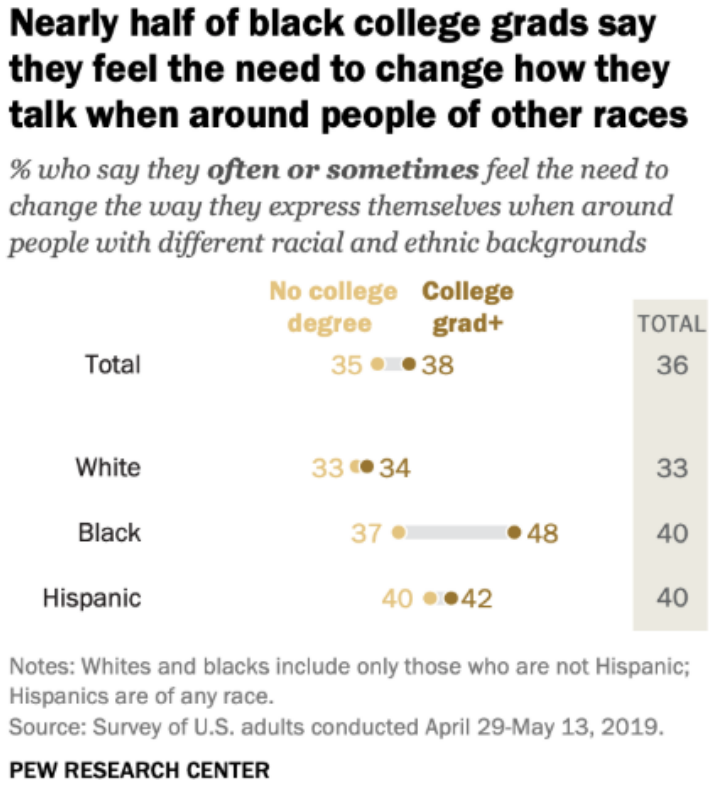
去canteen食飯；('go to the canteen for lunch')

我唔sure；('I'm not sure')

幫我check一check啊；('Help me search/check for it’)

Code switching (cont.)

- While code-switching is not inherently good or bad, it can reveal deep-rooted sociocultural dynamics.
  - For example, individuals may feel the pressure to conform to dominant cultural or linguistic norms in academic or professional settings.



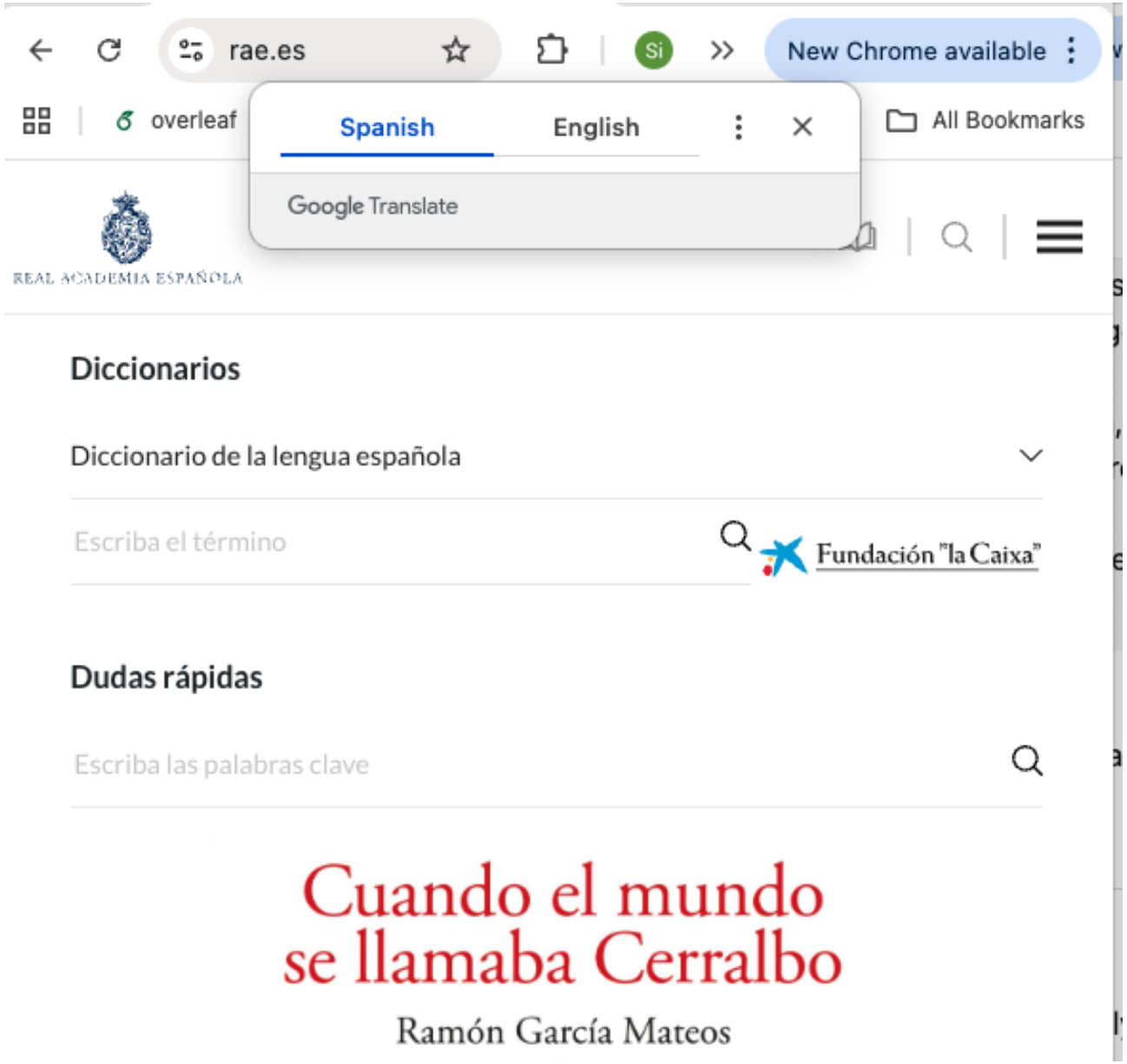
To be, or not to be…Black: The effects of racial codeswitching on perceived professionalism in the workplace ☆

Courtney L. McCluney [a], Myles I. Durkee [b], Richard E. Smith II [b], Kathrina J. Robotham [b], Serenity Sai-Lai Lee [c]

[a] 397 Ives Hall, ILR School, Cornell University, Ithaca, NY 14853, United States of America
[b] 530 Church Street, Department of Psychology, University of Michigan, Ann Arbor, MI 48109, United States of America
[c] The Wharton School of the University of Pennsylvania, PA 19104, United States of America

Human translators to machine translation

- Before the age of machines, languages were translated only by humans.
- Today, while human expert translators are still the ones translating important materials like books and literature, *much of the text online are actually translated by machines!*
- The automatic translation feature is everywhere:
  - Your browser will suggest to translate a website when you are visiting it in a language you don’t normally read
  - Your social media apps: Instagram, Facebook, TikTok, will suggest to translate, probably based on your language settings
  - Automatic live translation of speech: YouTube translates subtitle, etc.

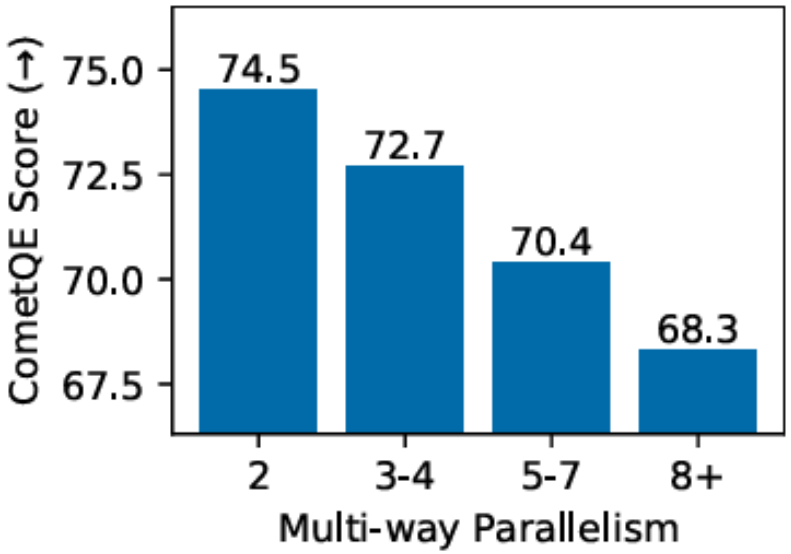


Visiting the website of a Spanish dictionary: Royal Spanish Academy (Real Academia Española)

and Chrome suggests to translate this webpage to English

Large amount of text online is translated

- Content on the web is often translated into many languages
- The quality of these translation is often low and likely created using machine translation
- Machine generated content dominates the translations in lower resource languages



A Shocking Amount of the Web is Machine Translated: Insights from Multi-Way Parallelism

Brian Thompson,[1] Mehak Preet Dhaliwal,[*2] Peter Frisch,[1] Tobias Domhan,[3] Marcello Federico[1]
[1]AWS AI Labs [2]UC Santa Barbara [3]Amazon Alexa
brianjt@amazon.com

Europe and Asia dominate most of the internet resources

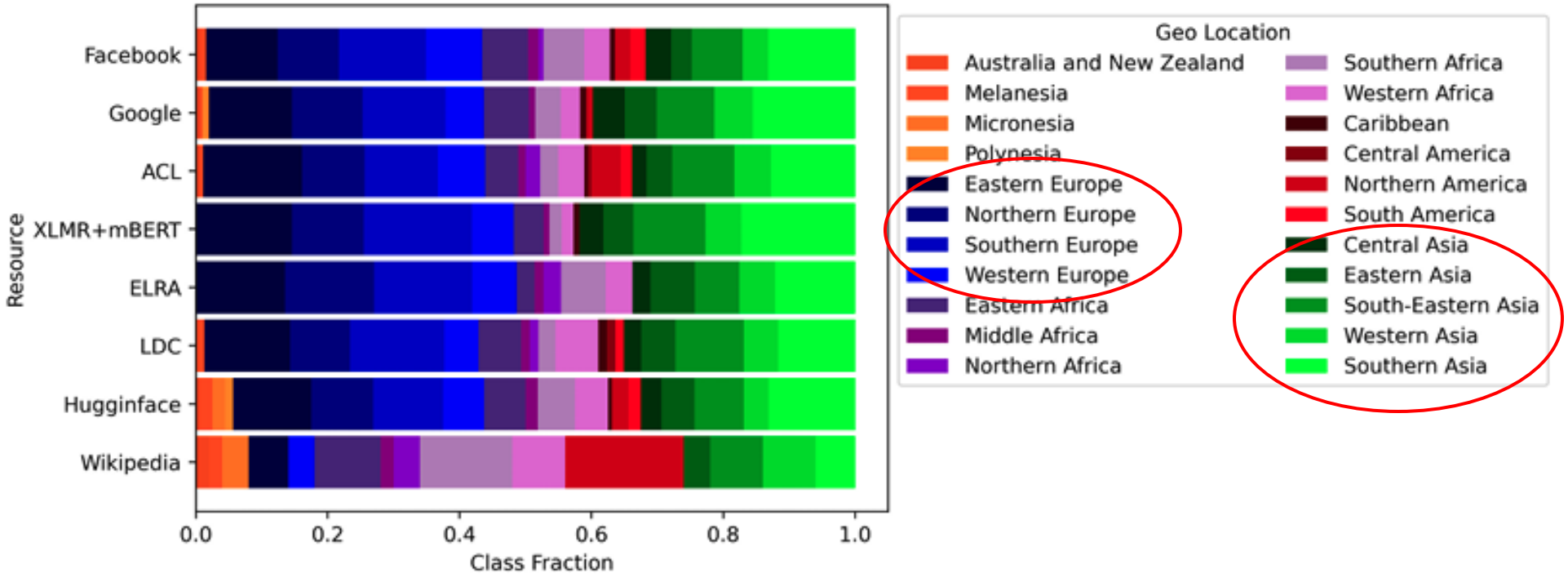


Figure 14: By Geographical Location of the Language Origin

Figure from Ranathunga and Silva (2022)

Resource distribution by language families

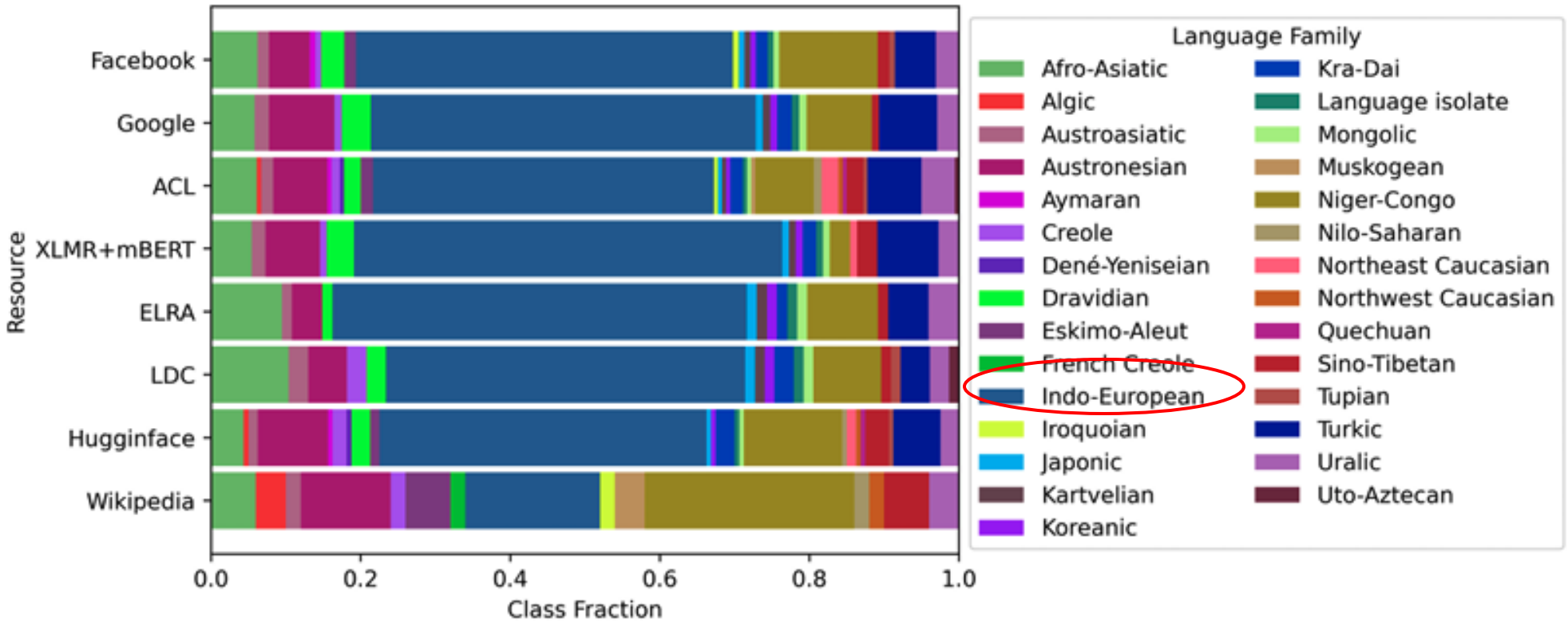


Figure 15: By Language Families

Figure from Ranathunga and Silva (2022)

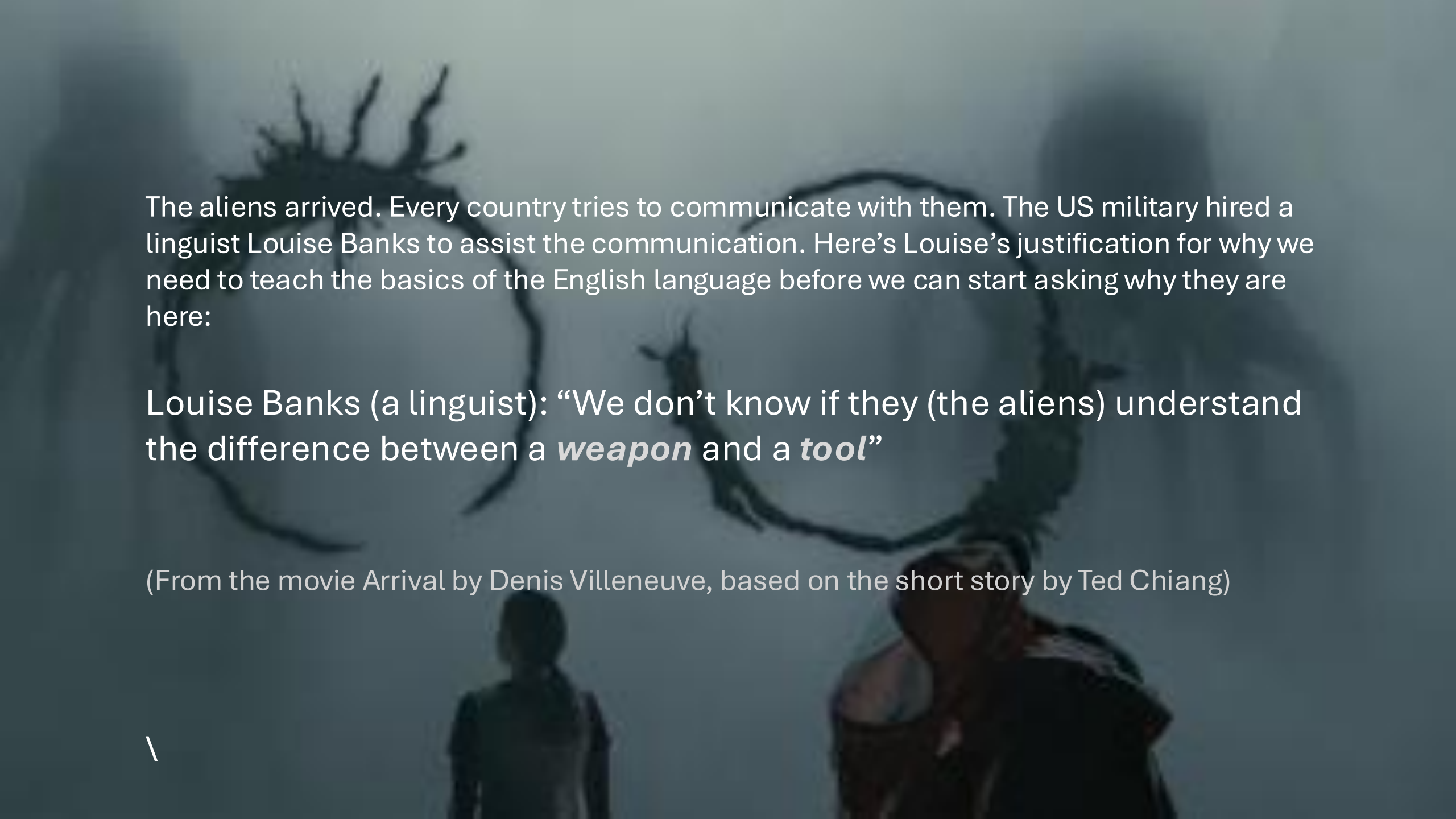
Despite all this, most LLMs are still English-centric

- LLMs can understand non-English language even when not intentionally multilingual because during pretraining, it has seen the entire internet, and that includes non-English content
- On the other hand, just because it’s a multilingual model, doesn’t mean it performs equally well on all languages.

Why is translation hard?

Translation is hard for human

- Not everyone is an expert translator
- A good translation requires:
  - Understanding of both languages: sentence structure, grammar, word sense disambiguation, etc.
    - Word-for-word is generally a bad translation
  - Deep understanding of both cultures
    - Awareness of cultural implications and nuance
  - Knowledge of idioms, figurative language, dialects, etc.
  - For literature specifically, it is a form of language art. Certain things just can’t be translated perfectly.



The aliens arrived. Every country tries to communicate with them. The US military hired a linguist Louise Banks to assist the communication. Here's Louise's justification for why we need to teach the basics of the English language before we can start asking why they are here:

Louise Banks (a linguist): "We don't know if they (the aliens) understand the difference between a ***weapon*** and a ***tool***"

(From the movie Arrival by Denis Villeneuve, based on the short story by Ted Chiang)

\

The complexity of meaning within a language

Definition from Webster dictionary

Weapon: something (such as a club, knife, or gun) used to injure, defeat, or destroy

Tool: (one of the definition) a device or implement, especially one held in the hand, used to carry out a particular function.

→ Tool can be considered the **hypernym** of weapon

The complexity of meaning across languages: **lexical divergence**

Between English and Chinese, **nouns** have less of this problem, but **verbs** are tricky…

Here's a good example

The Chinese word 打 （dǎ）generally means "beat" or "hit", the concept of beat and hit is somehow different from the English idea of beat and hit:

打架 : "hit frame" → fight
打鼓 : "hit drum" → play a drum

打篮球 : "hit basketball" → play basketball
打电话 : "hit telephone" → make a phone call
打开 : : "hit open" → open

Language family

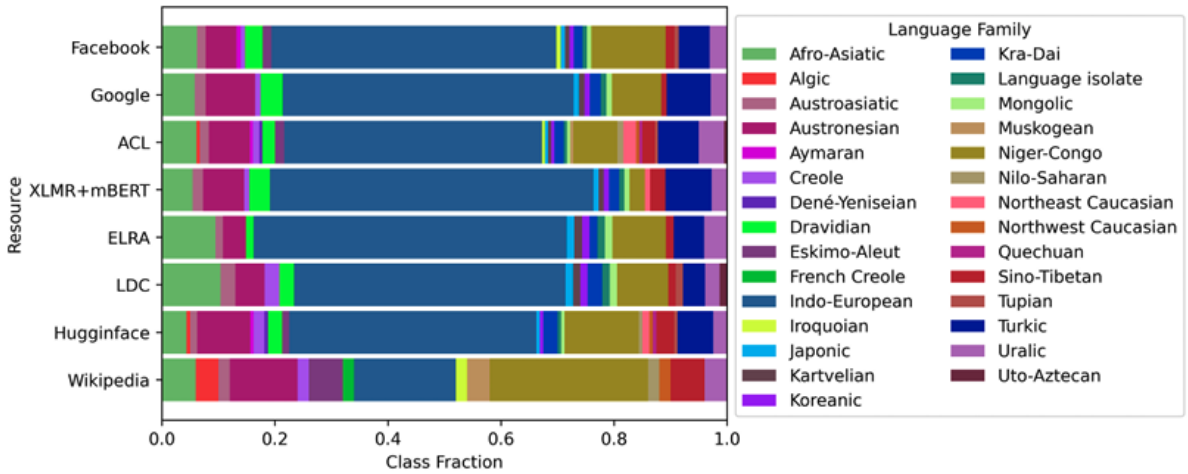


Figure 15: By Language Families

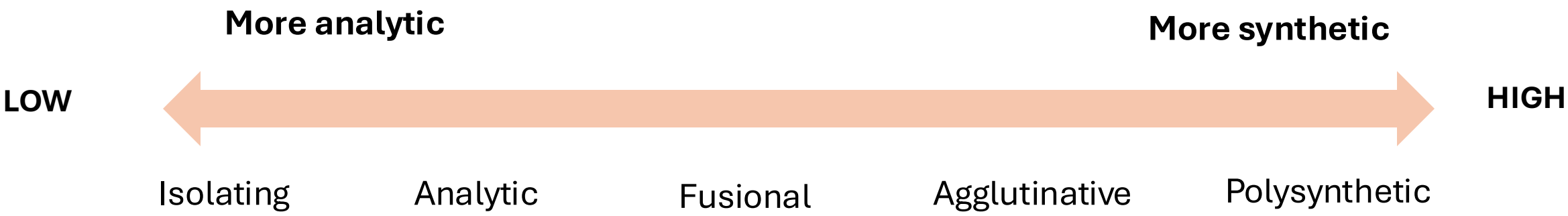
Ethnologue 27

Family	No. of languages
Niger–Congo	1,552
Austronesian	1,256
Trans–New Guinea	481
Sino-Tibetan	458
Indo-European	454
Australian	384
Afroasiatic	382
Nilo-Saharan	210
Otomanguean	179
Austroasiatic	167
Kra-Dai	91
Dravidian	85

Glottolog 5.0

Family	No. of languages
Atlantic–Congo	1,410
Austronesian	1,274
Indo-European	586
Sino-Tibetan	514
Afroasiatic	381
Trans–New Guinea	316
Pama–Nyungan	250
Otomanguean	181
Austroasiatic	158
Tai–Kadai	95
Dravidian	85
Arawakan	77

Morpheme to word ratio



Chinese

1 word 1 morpheme

English

Russian

A couple morphemes per word

Why is machine translation hard?

- Out of domain
- Amount of training data
- Low-frequency words
- Long sentences
- Decoding
- In the corpus:
  - Misaligned sentences
  - Misordered words
  - Wrong language
  - Untranslated sentences
  - Short segments

Six Challenges for Neural Machine Translation

Philipp Koehn
Computer Science Department
Johns Hopkins University
phi@jhu.edu

Rebecca Knowles
Computer Science Department
Johns Hopkins University
rknowles@jhu.edu

On the Impact of Various Types of Noise on Neural Machine Translation

Huda Khayrallah
Center for Language & Speech Processing
Computer Science Department
Johns Hopkins University
huda@jhu.edu

Philipp Koehn
Center for Language & Speech Processing
Computer Science Department
Johns Hopkins University
phi@jhu.edu

- Today, another problem is that evaluation for MT has not kept up with the translation quality achieved by LLMs
  - Human eval is obviously the best, but too expensive
  - And for low-resource languages, it's even more difficult

Low-resource languages

- We struggle to build effective MT systems for most human languages because the majority of them are considered low-resource languages
- Languages are “low-resource” due to many different reasons, including:
  - Sociocultural: limited written tradition, oral dominance, therefore less digital and textual data.
  - Political: some langauges lack official recognition or state support, therefore minimal investment in language documentation or tech development
  - Economy: communities that speak low-resource languages often have limited access to digital infrastructure, funding for linguistic/NLP research, education, data creation/curation, etc.

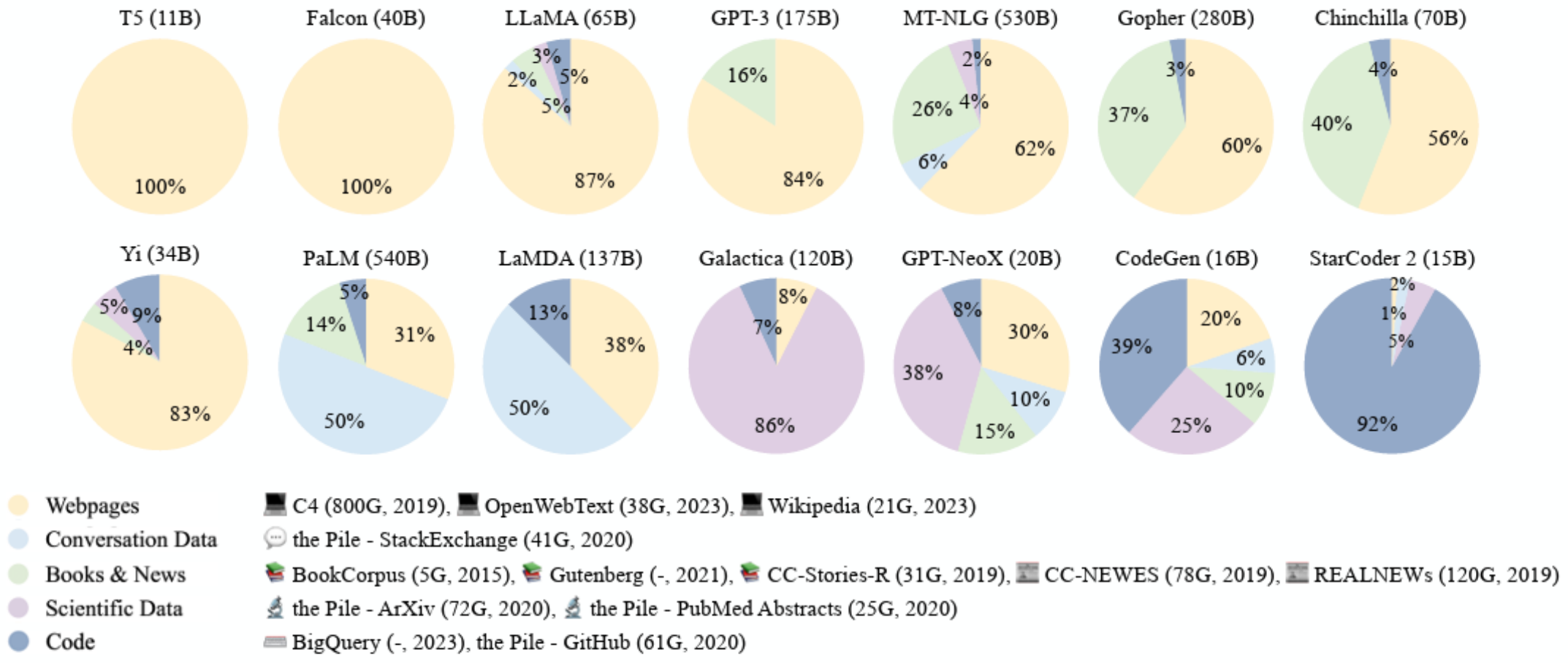


Fig. 6: Ratios of various data sources in the pre-training data for existing LLMs.

Zhao, et. al., 2023, A Survey of Large Language Models.
https://arxiv.org/pdf/2303.18223

Syntactic divergence

- Syntactic divergence: differences in sentence structure and grammar between languages

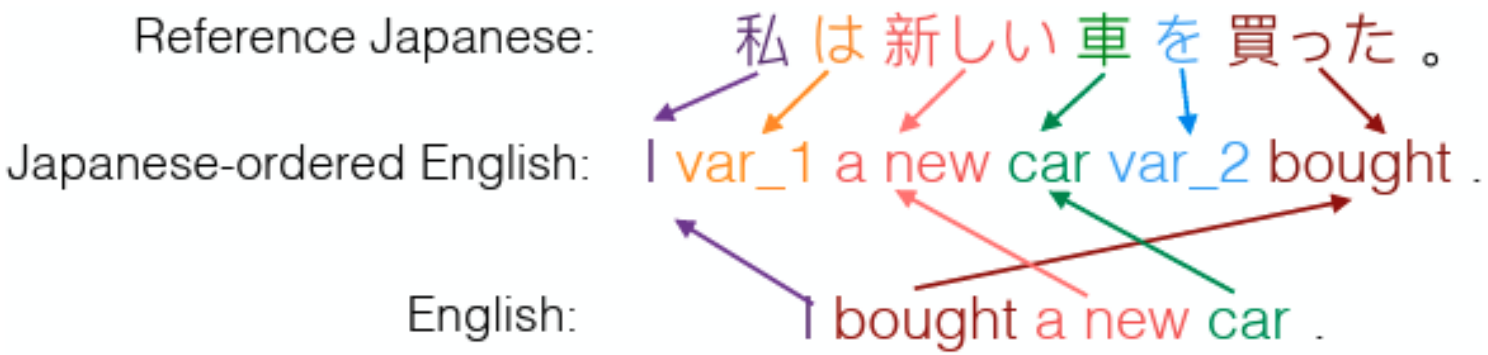


Figure from Zhou et al., 2019, Handling syntactic divergence in low-resource machine translation, https://aclanthology.org/D19-1143.pdf

Machine translation models

IBM models

- A series of models published in 1988-1990
- 6 in total, increasingly more complex
- It starts with lexical alignment to using HMM
- Using Expectation-Maximization (EM) algorithm

Statistical machine translation

- For IBM models, you can't have multiple translations for a word.
  - One to many is ok, but not many to one.
- Use statistical models like IBM, but richer structures
  - Phrases, syntax
  - Better translation fluency and accuracy
- Phrase-based SMT: break sentences into phrases
  - Learn phrase tables and **pivot tables** (mappings like cotton candy → (fr) barbe Ă papa (daddy's beard))
  - Many to many is now possible
- Downside: lots of heavy feature engineering
  - E.g. Moses uses a log linear model, and the weights requires heavy feature engineering

Neural machine translation models

- You've already learned all about this in previous lectures
- RNNs, LSTMs, encoder-decoder models, transformers
- Much more fluent and context-awrare translations
- Less manual feature design
- Attention mechanism in transformers really improve the long-term dependency problem in the RNNs.
- Downside: training is slow and still struggle with low-resource languages.

Transformer encoder-decoder

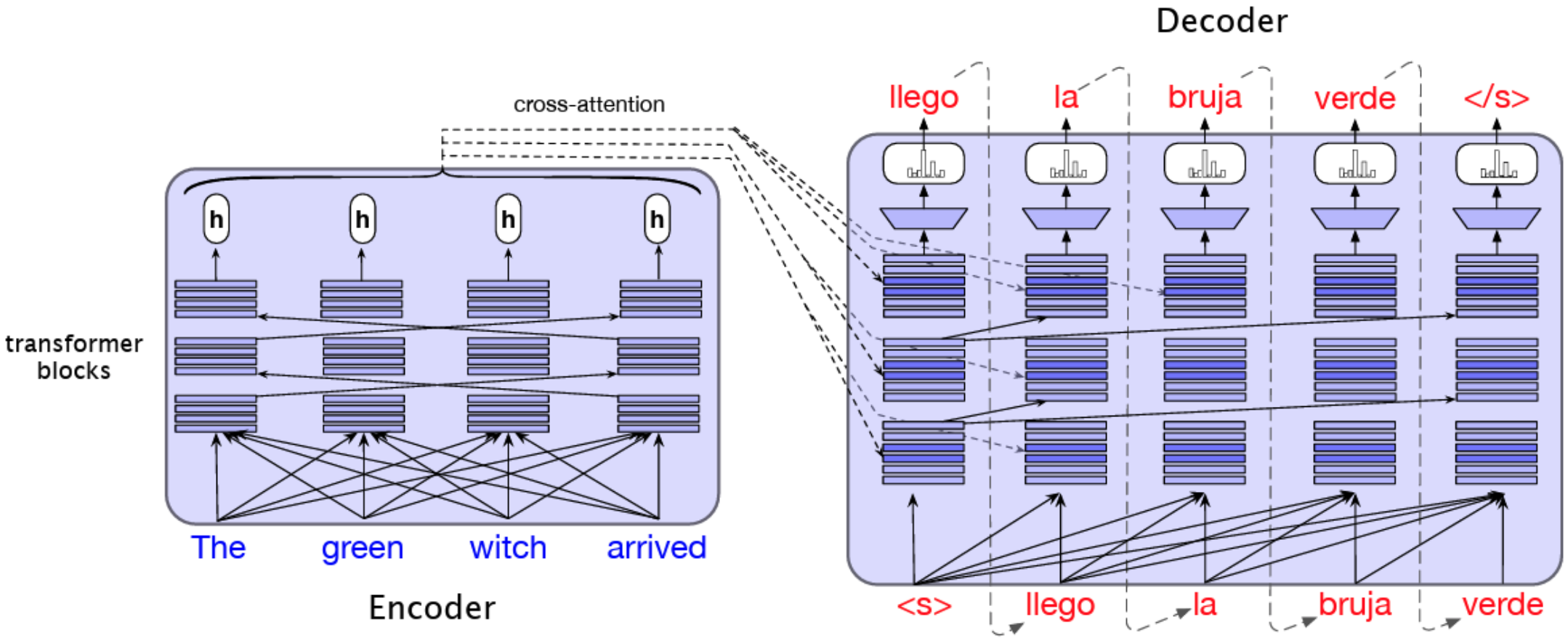


Figure from older version of Jurafsky and Martin chapter 10

Revisit minimum bayes risk decoding

Minimum bayes risk decoding

- Works better than beam search and temperature scaling
- Often used on machine translation (Kumar and Byrne, 2004), speech recognition
- High-level idea: instead of choosing the most probable, choose the one likely to have least error (low risk).

- **Risk**, here, means, rather than picking the most probable sequence, we instead do some *risk assessment*:
  - According some metrics (BLEU, chrF, BERTScore),
  - Comparing to some known good translation (minimizing expected loss/risk)

Minimum bayes risk decoding

- In practice, we don't know the perfect set of translation for a given sentence, we instead, we choose the candidate translation which is most similar with some set of candidate translation
  - First beam search or sampling
  - Then pairwise similarity
- Essentially, we are approximating the enormous space of all possible translations U with a smaller set of possible candidate translations Y.
- Given this set of possible candidate translation Y, and some similarity function $util$, we choose the best translation which is the most similar to the other candidate translations.

$$\hat{y} = \underset{y \in \mathscr{Y}}{\operatorname{argmax}} \sum_{c \in \mathscr{Y}} \operatorname{util}(y, c)$$

Tokenization

- What happens if we tokenize Spanish with English tokenizer
- Also some languages simply do have more morphemes
- Recall different tokenization methods from previous lecture
- There are language-specific vs. language-agnostic tokenization
  - Language-specific: user must specify the language before tokenization
  - Language-agnostic: don’t need explicit language ID
    - E.g., SentencePiece
    - Used by multilingual LMs like mT5, mBERT
- And then there are tokenizer-free models that operate directly on bytes or characters
  - Great for low-resource language

Multilinguality and LLMs

- Many LLMs are inevitably and “incidentally” multilingual
- The internet is made up of multilingual data
- E.g. GPT3, GPT4, Claude, Gemma, LlaMA 2
- Their tokenizers and training pipelines are optimized for English

Some **intentionally** multilingual MT models

- BLOOM
- No Language Left Behind (NLLB)
- mT5
- mBART
- XLM, XLM-R
- mBERT

Llama 3 is capable of multilingual understanding but it’s not intentionally multilingual

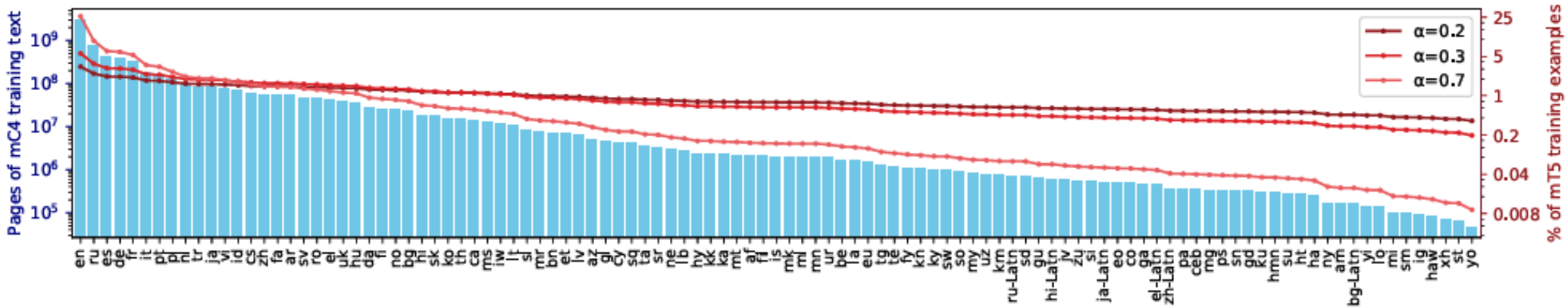


Figure 1: Page counts per language in mC4 (left axis), and percentage of mT5 training examples coming from each language, for different language sampling exponents α (right axis). Our final model uses α=0.3.

Model	Architecture	Parameters	# languages	Data source
mBERT (Devlin, 2018)	Encoder-only	180M	104	Wikipedia
XLM (Conneau and Lample, 2019)	Encoder-only	570M	100	Wikipedia
XLM-R (Conneau et al., 2020)	Encoder-only	270M – 550M	100	Common Crawl (CCNet)
mBART (Lewis et al., 2020b)	Encoder-decoder	680M	25	Common Crawl (CC25)
MARGE (Lewis et al., 2020a)	Encoder-decoder	960M	26	Wikipedia or CC-News
mT5 (ours)	Encoder-decoder	300M – 13B	101	Common Crawl (mC4)

Xue et al., 2021

Curse of multilinguality

- The performance degrades when a fixed-capacity model is trained on an increasing number of languages
  - In other words, if the model size or training resource size stays the same, adding more languages to a multilingual model will cause the performance of each language to drop
- Prof. Blevins will talk about her work on this topic next Tuesday

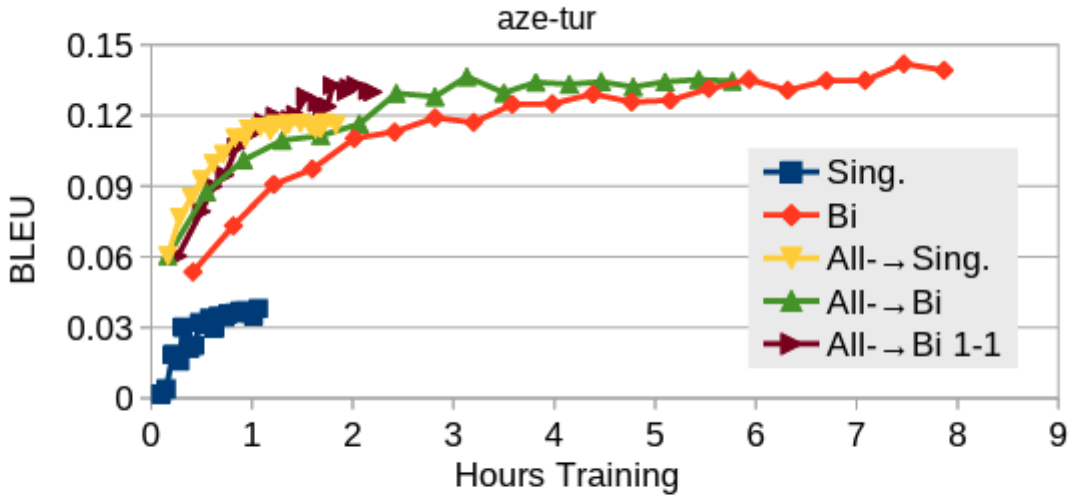
Breaking the Curse of Multilinguality with Cross-lingual Expert Language Models

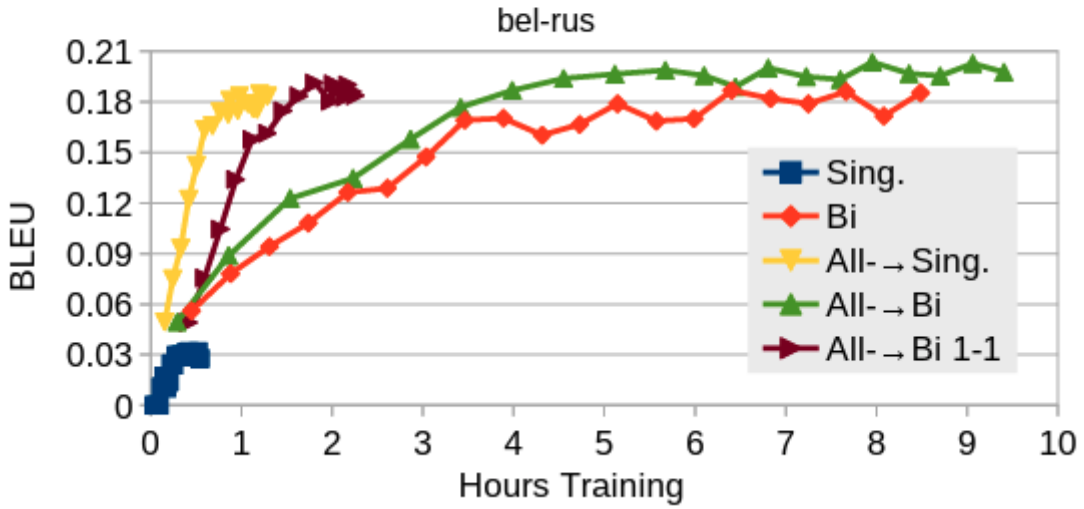
Terra Blevins[1†] **Tomasz Limisiewicz**[2*] **Suchin Gururangan**[1] **Margaret Li**[1]
Hila Gonen[1] **Noah A. Smith**[1,3] **Luke Zettlemoyer**[1]
[1]Paul G. Allen School of Computer Science and Engineering, University of Washington
[2]Faculty of Mathematics and Physics, Charles University in Prague
[3]Allen Institute for Artificial Intelligence

How to include a new language or low-resource languages

- Train a new model? fine-tune?
- Does this new language have a **similar language** that's already in the language coverage of this particular model?
  - Is this similar language a high-resource language? Hopefully yes
- How much data do you have for this language:
  - If not a lot, can you create synthetic data using a paraphraser? It actually really helps low-resource languages (Khayrallah et al., 2020).
  - Can also create more data using **back-translation:** translate back to the original language as data augmentation
    source sentence → target sentence → new source sentence

Adapting low resource languages





[Neubig and Hu, 2018, Rapid Adaptation of Neural Machine Translation to New Languages]

Improve MT on low-resource language

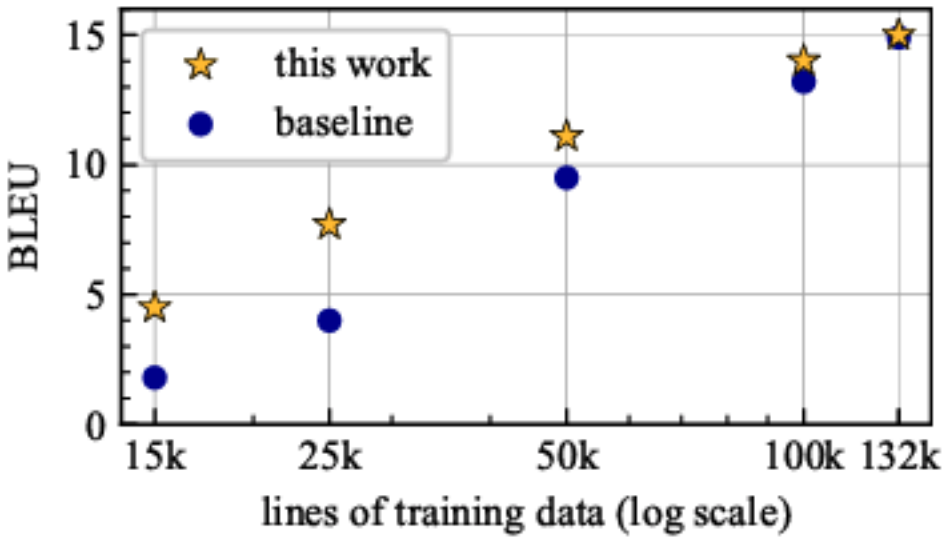


Figure 2: Bengali-English data ablation. Improvements of 2.7, 3.7, 1.6, and 0.8 BLEU at the 15k, 25k, 50k, and 100k subsets are statistically significant.

Data augmentation:

- Paraphraser: what kind of paraphraser matters too.
- Back-translation

Khayrallah et al., 2020, Simulated Multiple Reference Training Improves Low-Resource Machine Translation

Evaluation for machine translation

Eval metrics

- BLEU, chrF: n-gram based
  - BLEU: word level (usually 1 to 4 grams)
    - Includes a brevity penalty to prevent very short translations from scoring high.
    - Good for languages with clear word boundaries like English and Spanish
    - Sensitive to word order: because of the 2+ grams
  - chrF: character level.
    - More robust to morphological variation and small spelling differences
    - better for languages without clear word boundaries like Chinese, Japanese, and Thai
- COMET: using embedding to measure semantic similarity
- Human eval: costly but gold standard

If you are interested in MT, check out the Conference on Machine Translation (WMT)

Language model	Input tok.	Output tok.	Cost
Aya23	4.4 M	0.7 M	4.1 $
Claude-3.5	5.5 M	1.0 M	31.9 $
CommandR-plus	4.4 M	0.7 M	23.4 $
Gemini-1.5-Pro	3.9 M	0.6 M	40.3 $
GPT-4	5.9 M	1.0 M	240.4 $
Llama3-70B	5.0 M	0.7 M	5.1 $
Mistral-Large	6.0 M	1.1 M	37.0 $
Phi-3-Medium	5.9 M	1.1 M	4.5 $

Table 4: Number of input and output tokens and estimated pricing for translating the full WMT24 test set without test suites. The Gemini model refused to translate Icelandic, and the estimate is therefore lower. Pricing for the open models Aya23 and Llama3 was estimated via together.ai.

Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, Vilém Zouhar

Recurrent tasks

Translation tasks

- General machine translation task (former News task)
- Biomedical translation task
- Multimodal translation task
- Unsupervised and very low resource translation task
- Lifelong learning in machine translation task
- Chat translation task
- Life-long learning in machine translation task
- Machine translation using terminologies task
- Sign language translation task
- Robustness translation task
- Triangular machine translation task
- Large-scale multilingual machine translation task