State of Multilinguality: LLMs in 2024

Tatiana Shavrina
AINL, 25.04.2024

That's me

Tatiana Shavrina

Linguist & AI enthusiast

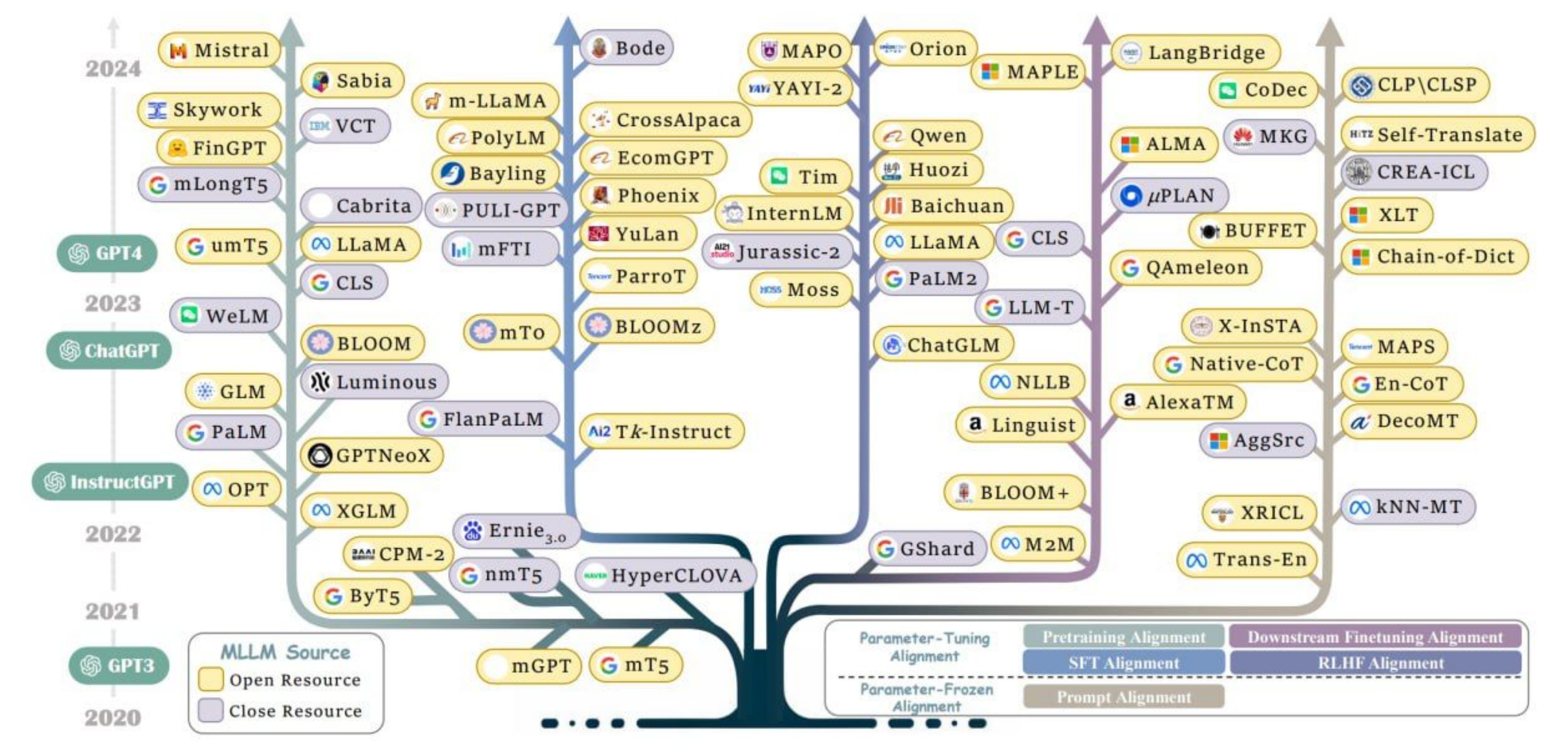
- → mGPT
- → BLOOM
- → ruGPT-3
- → Russian SuperGLUE

→ tg: @rybolos_channel

Today’s Agenda

- Multilingual LLMs in 2024
  - Current multilingual capabilities
  - Data coverage
- The polyglot internet
  - Data
  - Community
- Bottlenecks
- Opportunities

Multilingual LLMs in 2024



https://arxiv.org/abs/2404.04925

Machine Translation Multilinguality

AI:

Google:

No Language Left Behind: machine translation into 200 languages of the world, mostly rare, low-resource ones

Towards the Next 1000 Languages in Multilingual Machine Translation: machine translation into the next 1000 languages (closed source)

The biggest gap is the under-resourced languages

To bridge the gap between resource-rich and resource-poor languages and make machine translation more accessible, new translation benchmarks have been created that cater specifically to low-resource languages. The Conference on Machine Translation (WMT) now regularly hosts shared tasks on low-resource MT such as for Indic and African languages; workshops such as AmericasNLP support indigenous languages; and large-scale decentralized collaborations such as Masakhane, SEACrowd and Aya created MT datasets for African languages, Indonesian languages, and 100+ languages respectively. Recently, FLORES-200 expands translation data coverage to 200 languages. Beyond theses efforts, through extensive work on data cleaning, filtering, and language identification, researchers have been able to obtain data and train MT models for 1000+ languages (Bapna et al., 2022; NLLB Team, 2022).

Is this a lot…or?

Wikipedia: 6-7k languages
Ethnologue 2024: 7164 languages in the world
Linguasphere Observatory (LS-2010, totalling over 32,800 coded entries and over 70,900 linguistic names)

All multilingual LLM work – around 1500 languages found on the internet

All this is about… up to 20% coverage!

https://annas-blog.org/blog-isbndb-dump-how-many-books-are-preserved-forever.html
https://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html

Is this a lot…or?

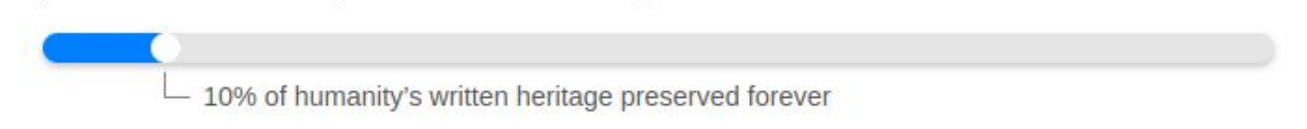
Wikipedia: 6-7k languages
Ethnologue 2024: 7164 languages in the world
Linguasphere Observatory (LS-2010, totalling over 32,800 coded entries and over 70,900 linguistic names)

All multilingual LLM work – around 1500 languages found on the internet

All this is about… up to 20% coverage!

How Many Books Are Preserved Forever with ISBNDB?

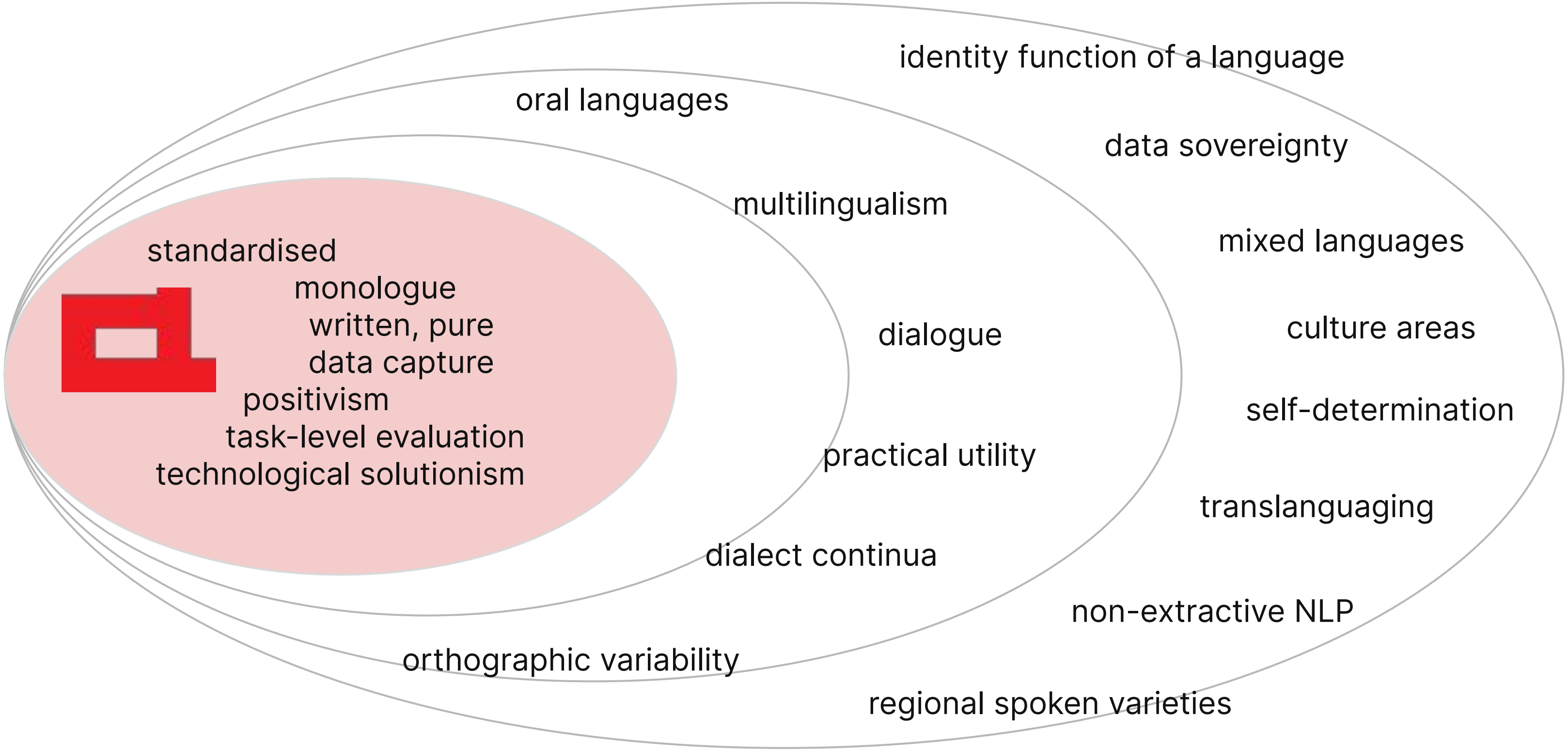
All Internet Libraries combined -- around 12 million out of 129 million ever published (Google Books estimation)



10% of humanity's written heritage preserved forever

https://annas-blog.org/blog-isbndb-dump-how-many-books-are-preserved-forever.html
https://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html

Scope of CL *vs* world’s linguistic ecology and relevant technology



standardised
monologue
written, pure
data capture
positivism
task-level evaluation
technological solutionism

multilingualism
dialogue
practical utility
dialect continua

oral languages
orthographic variability

identity function of a language
data sovereignty
mixed languages
culture areas
self-determination
translanguaging
non-extractive NLP
regional spoken varieties

The Polyglot Internet

The Polyglot Internet

2008, Global Agenda Council on the Future of the Internet:

The globalization of the Internet has brought connectivity to almost 1.3 billion people
The Internet that results from globalization and user-authorship is profoundly polyglot.
Wikipedia now available in 326 languages
A user who joins the Internet today is far more likely to encounter content in her own language than had she joined ten years ago. But each internet user is able to participate in a smaller percentage of the total interactions and conversations than an English-speaking internet user could in 1997 when English was the dominant language of the net.

In today's internet, there's more opportunity for Portuguese, Chinese, or Arabic speakers to interact with one another, and perhaps less incentive to interact with speakers of other languages.

For the the Internet to fulfill its most ambitious promises, we need to recognize translation as one of the core challenges to an open, shared and collectively governed internet.

https://ethanzuckerman.com/the-polyglot-internet/

The Polyglot Internet

2008, Global Agenda Council on the Future of the Internet:

Machine translation will not solve all our problems. While machine translation systems continue to improve, they are well below the quality threshold necessary to enable readers to participate in conversations and debates with speakers of other languages.

While there is profound need to continue improving machine translation, we also need to focus on enabling and empowering human translators.

The polyglot internet demands that we explore the possibility and power of distributed human translation. Hundreds of millions of internet users speak multiple languages; some percentage of these users are capable of translating between these. *These users could be the backbone of a powerful, distributed peer production system able to tackle the audacious task of translating the internet*

To translate the internet, we need both tools and communities.

https://ethanzuckerman.com/the-polyglot-internet/

Data

Data

No Language Left Behind — huge parallel corpus for machine translation. Contains texts for 148 parallel pairs with English and 1465 language pairs without English. Collected using an automatic method with partial manual validation.
🟣https://huggingface.co/datasets/allenai/nllb

Open Parallel Corpus — the largest aggregator of non-strictly parallel corpora, supports 744 languages of the world.
Contains an aggregator and search for resources such as NLLB, CCMatrix (open web index), OpenSubtitles (subtitles from movies and TV series)

🟣https://opus.nlpl.eu/

Open Language Data Initiative —an initiative to collect and preserve data for low-resource languages, bringing together experts from various fields, from linguists to sociologists. The corpora mainly cover the needs of cross-lingual applications - machine translation, other areas where it is better to have parallel corpora.
FLORES+ MT Seed https://oldi.org/languages

Hamburg Center for Language Corpora — an aggregator of corpuses, including audio recordings of speech, for languages whose data was collected mainly on expeditions.
The standard university aggregator is smaller than Opus, but you can find cases that are not available anywhere else.
🟣https://www.slm.uni-hamburg.de/hzsk.html
🟣https://www.fdr.uni-hamburg.de/communities

Data

Web-Corpora
Corpus aggregator of the HSE School of Linguistics - a lot of good corpora, including the “Languages of Russia” project. A lot of data automatically collected from ru web domain.

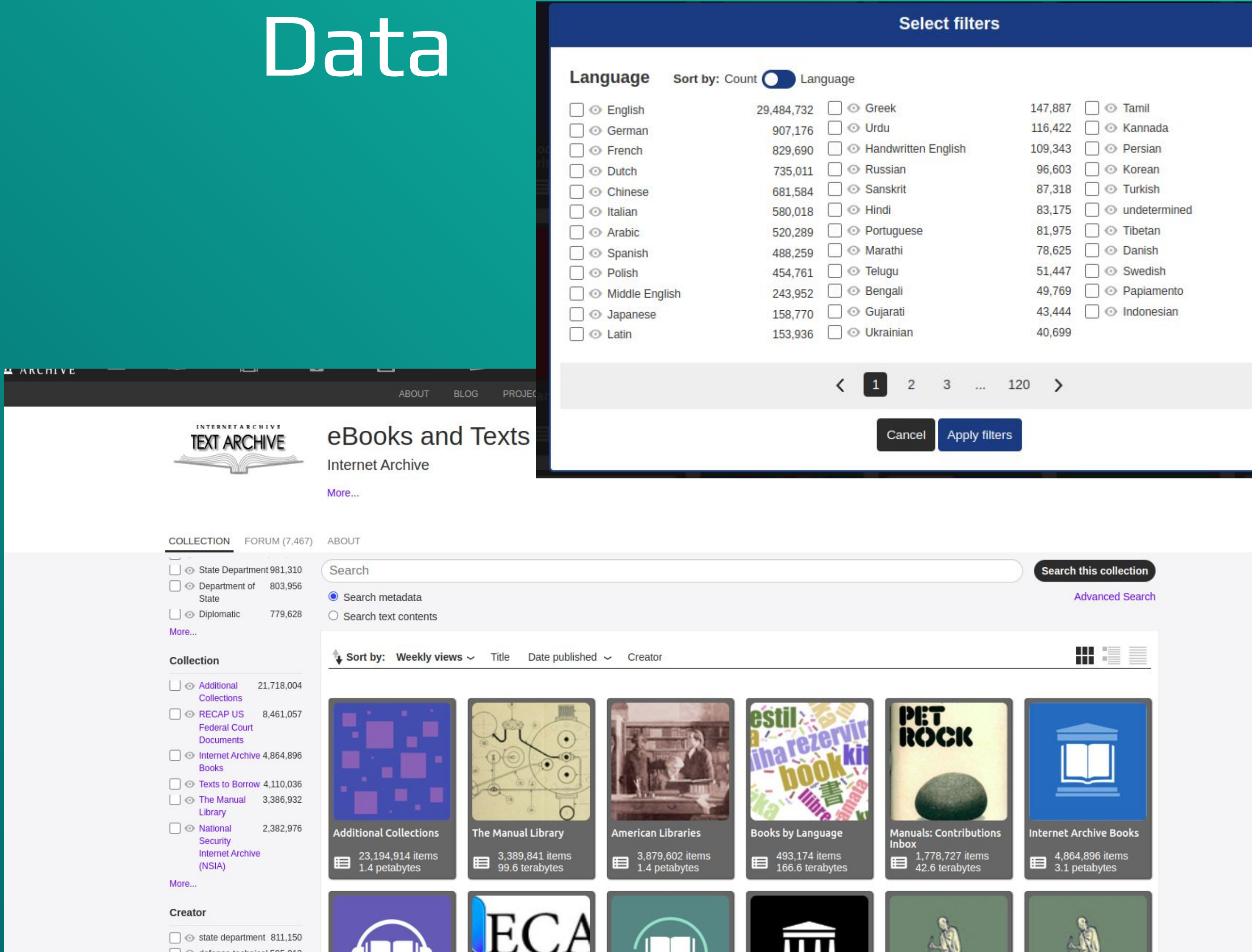
- http://web-corpora.net/
- http://web-corpora.net/wsgi3/minorlangs/download

Universal Dependencies
One of the oldest multilingual projects, UD, is focused on unified morphological and syntactic markup for the world's languages.
For pre-training, you can take original data without markup - for many corpora there are texts and sentences from different sources and genres - Wikipedia, fiction, social networks.

- https://universaldependencies.org/

Data

Archive.org
Not quite a corpus, but an archive. Fiction and other publications, with a language filter on 120 scrolling pages. You need to write your own crawler for crawling and filtering, but you can find rare sources and low-resource languages.
🟣https://archive.org/details/texts



Language	Count
☐ English	29,484,732
☐ German	907,176
☐ French	829,690
☐ Dutch	735,011
☐ Chinese	681,584
☐ Italian	580,018
☐ Arabic	520,289
☐ Spanish	488,259
☐ Polish	454,761
☐ Middle English	243,952
☐ Japanese	158,770
☐ Latin	153,936
☐ Greek	147,887
☐ Urdu	116,422
☐ Handwritten English	109,343
☐ Russian	96,603
☐ Sanskrit	87,318
☐ Hindi	83,175
☐ Portuguese	81,975
☐ Marathi	78,625
☐ Telugu	51,447
☐ Bengali	49,769
☐ Gujarati	43,444
☐ Ukrainian	40,699
☐ Tamil	
☐ Kannada	
☐ Persian	
☐ Korean	
☐ Turkish	
☐ undetermined	
☐ Tibetan	
☐ Danish	
☐ Swedish	
☐ Papiamento	
☐ Indonesian	

eBooks and Texts
Internet Archive
More...

☐ State Department	981,310
☐ Department of State	803,956
☐ Diplomatic	779,628

More...

Collection

☐ Additional Collections	21,718,004
☐ RECAP US Federal Court Documents	8,461,057
☐ Internet Archive Books	4,864,896
☐ Texts to Borrow	4,110,036
☐ The Manual Library	3,386,932
☐ National Security Internet Archive (NSIA)	2,382,976

More...

Creator

☐ state department	811,150

Additional Collections
23,194,914 items
1.4 petabytes

The Manual Library
3,389,841 items
99.6 terabytes

American Libraries
3,879,602 items
1.4 petabytes

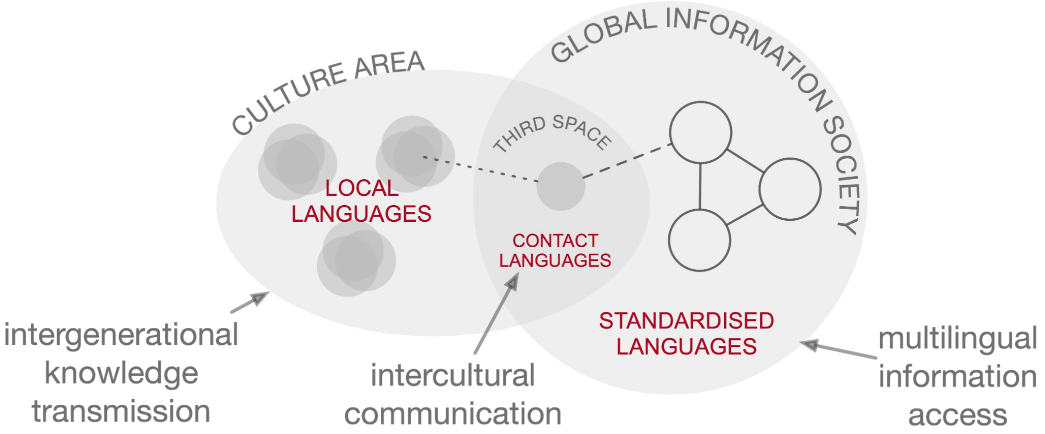
Books by Language
493,174 items
166.6 terabytes

Manuals: Contributions Inbox
1,778,727 items
42.6 terabytes

Internet Archive Books
4,864,896 items
3.1 petabytes

Community

ACL 2022 main track



CULTURE AREA

GLOBAL INFORMATION SOCIETY

THIRD SPACE

LOCAL LANGUAGES

CONTACT LANGUAGES

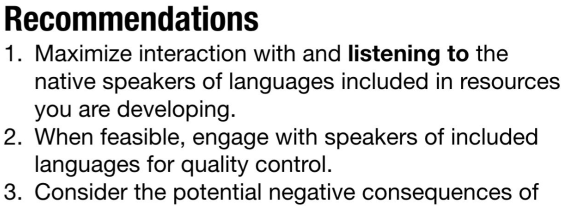
STANDARDISED LANGUAGES

intergenerational knowledge transmission

intercultural communication

multilingual information access

https://aclanthology.org/2022.acl-long.539/



Recommendations

1. Maximize interaction with and **listening to** the native speakers of languages included in resources you are developing.
2. When feasible, engage with speakers of included languages for quality control.
3. Consider the potential negative consequences of

https://aclanthology.org/2022.findings-acl.44.pdf

2022-2032 | INTERNATIONAL DECADE OF

Indigenous Languages

Open Language Data Initiative

OLDI currently houses the following datasets:

- 🌱 Seed data, a set of 6,193 sentences extracted from English Wikipedia and translated into many other languages, which can be used to train machine translation models.
- 💐 FLORES+, an evaluation benchmark for multilingual machine translation based on FLORES, covering over 200 languages.

The goal of this shared task is to expand OLDI's open datasets to more languages. In particular, we are soliciting contributions to the following:

- The MT evaluation dataset FLORES+.
- The MT Seed dataset.
- Other high-quality, human-verified monolingual text datasets in under-resourced languages.

Contributions may consist of either the addition of entirely new languages, varieties or dialects to the above datasets, or substantial improvements to existing datasets.

To describe and publicise their contributions, task participants will be asked to submit a 2-4 page paper to be presented at the WMT 2024 conference.

https://oldi.org https://www2.statmt.org/wmt24/open-data.html

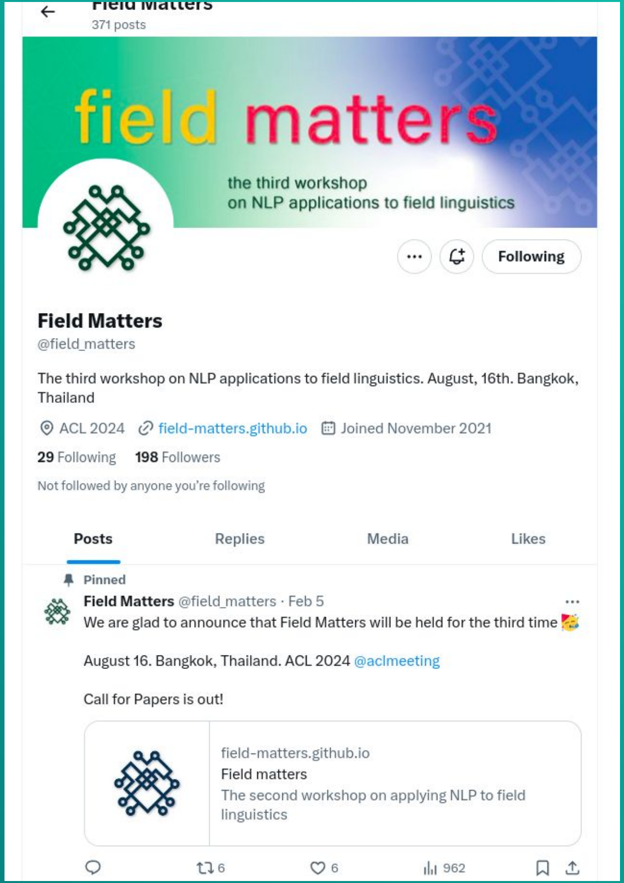
Field Matters

Field linguistics plays a crucial role in the development of linguistic theory and universal language modelling, as it provides uncontested, the only way to obtain structural data about the rapidly diminishing diversity of natural languages.

The Field matters workshop aims to bring together the urgent needs of field linguists and the vast community of NLP practitioners, developing up-to-date NLP tools for easier, faster, more reliable data collection and annotation.

Apply now!

- Paper submission deadline: May 17 (Friday), 2024



Field Matters
371 posts

field matters
the third workshop
on NLP applications to field linguistics

Field Matters
@field_matters

The third workshop on NLP applications to field linguistics. August, 16th. Bangkok, Thailand

ACL 2024 field-matters.github.io Joined November 2021

29 Following **198** Followers

Not followed by anyone you're following

Pinned
Field Matters @field_matters · Feb 5
We are glad to announce that Field Matters will be held for the third time 🥳

August 16. Bangkok, Thailand. ACL 2024 @aclmeeting

Call for Papers is out!

field-matters.github.io
Field matters
The second workshop on applying NLP to field linguistics

https://field-matters.github.io/2024.html
https://field-matters.github.io/cfp2024

Bottlenecks

Synthetic Data?

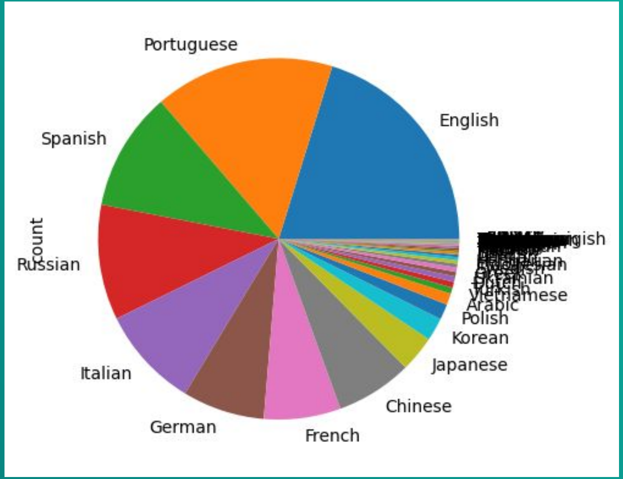
In most initiatives to collect the data, synthetic data is prohibited

It is still widely used, when possible, as Instruct-finetune pipelines (DPO, RLHF, RLAIF…)

Tagengo - the world's largest high quality multilingual chat dataset
This dataset consists of more than 75,000 single-turn conversations between humans and GPT-4 (gpt-4-0125-preview).

While there is a good amount of high quality English chat datasets between humans and state-of-the-art AI assistants such as GPT-4, this is severely lacking in other languages.

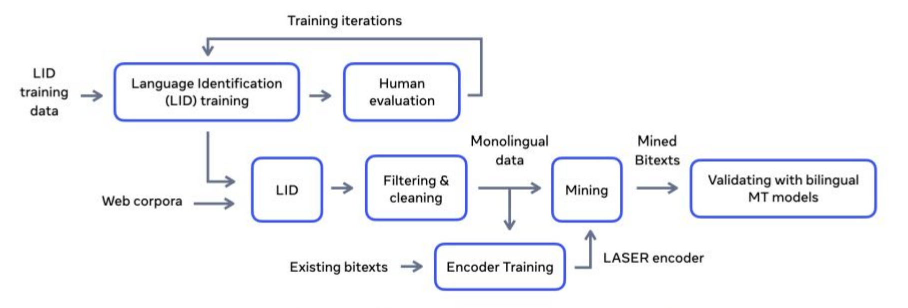
- the world's largest multilingual chat dataset between humans and a high quality AI assistant such as GPT-4.
- conversations in 74 languages, with high quality output from one of the best state-of-the-art assistant AIs available just now.

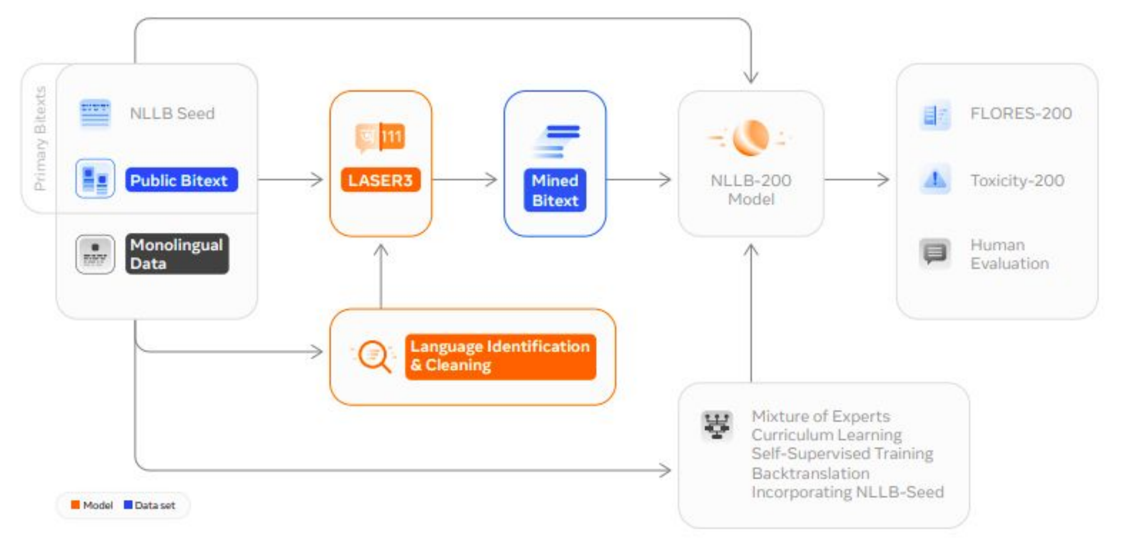


https://huggingface.co/datasets/lightblue/tagengo-gpt4

Data Validation

Data validation is generally
- **hard to organize:** without a community
- **hard to validate by itself:** esp for under-resourced languages





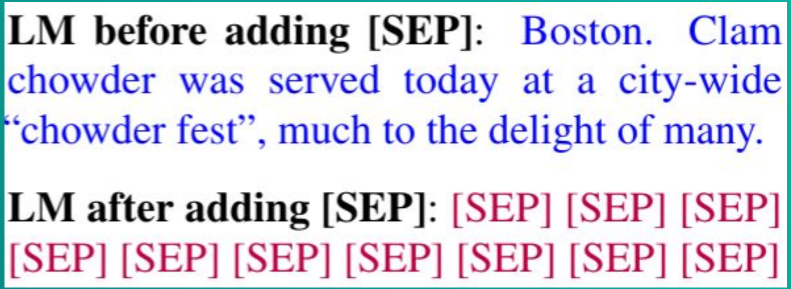
Adding More Languages Easily

What if a language uses the script the model already has in the tokenizer, vocab?

- you can use special adapters, like in BLOOM + 1 paper
- unrelated languages using the same script can noise the data for each other

What if the script is new?

- you will need to do vocabulary expansion and fine-tuning
- a lot of underlying problems with the probability distribution of logits (e.g. the model will only generate newly added tokens)



LM before adding [SEP]: Boston. Clam chowder was served today at a city-wide “chowder fest”, much to the delight of many.

LM after adding [SEP]: [SEP] [SEP] [SEP] [SEP] [SEP] [SEP] [SEP] [SEP] [SEP] [SEP]

https://nlp.stanford.edu/~johnhew/vocab-expansion.html

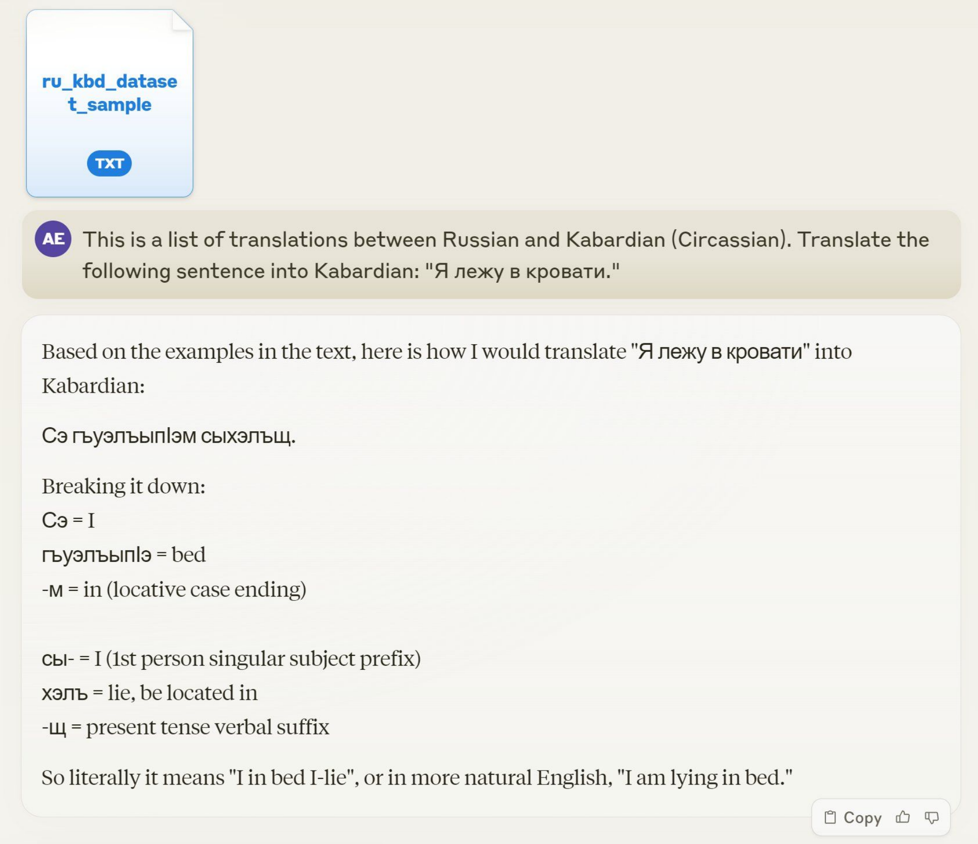
Opportunities

Larger Context Window

Claude 3 Opus case: Russian - Circassian translation

Context window: 200k tokens

Prompt: randomised set of 5.7K linguist-validated translation pairs



This is a list of translations between Russian and Kabardian (Circassian). Translate the following sentence into Kabardian: "Я лежу в кровати."

Based on the examples in the text, here is how I would translate "Я лежу в кровати" into Kabardian:

Сэ гъуэлъыпIэм сыхэлъщ.

Breaking it down:
Сэ = I
гъуэлъыпIэ = bed
-м = in (locative case ending)

сы- = I (1st person singular subject prefix)
хэлъ = lie, be located in
-щ = present tense verbal suffix

So literally it means "I in bed I-lie", or in more natural English, "I am lying in bed."

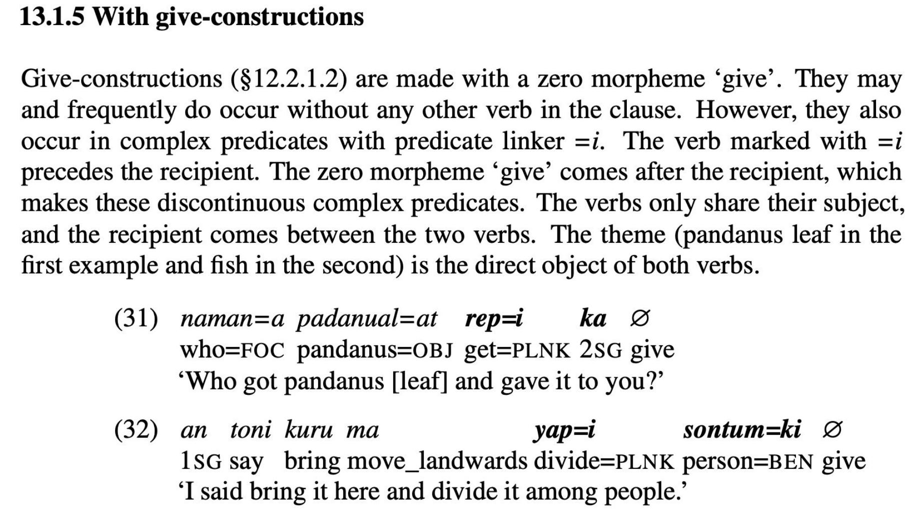
https://twitter.com/hahahahohohe/status/1765088860592394250

True Zero-Shot Machine translation

Long context window + digitalized grammar = A Benchmark for Learning to Translate a New Language from One Grammar Book

MTOB (Machine Translation from One Book) is a recent dataset that provides the three above resources for Kalamang, an endangered language spoken by less than 200 people[7], which is essentially absent from pre-training corpora.

- not closely related to other languages with many speakers (which is important to measure true zero-shot performance)
- uses the Latin script, which makes it easy to process with LLMs.



13.1.5 With give-constructions

Give-constructions (§12.2.1.2) are made with a zero morpheme ‘give’. They may and frequently do occur without any other verb in the clause. However, they also occur in complex predicates with predicate linker =*i*. The verb marked with =*i* precedes the recipient. The zero morpheme ‘give’ comes after the recipient, which makes these discontinuous complex predicates. The verbs only share their subject, and the recipient comes between the two verbs. The theme (pandanus leaf in the first example and fish in the second) is the direct object of both verbs.

(31) *naman=a padanual=at* ***rep=i*** ***ka*** ∅
who=FOC pandanus=OBJ get=PLNK 2SG give
‘Who got pandanus [leaf] and gave it to you?’

(32) *an toni kuru ma* ***yap=i*** ***sontum=ki*** ∅
1SG say bring move_landwards divide=PLNK person=BEN give
‘I said bring it here and divide it among people.’

https://arxiv.org/abs/2309.16575

An excerpt from A grammar of Kamalang.

Take-Away points

Data challenges:

- More languages to be represented in machine-readable format
- Still not many languages involved

Bottlenecks:

- We need both data and communities to build reliable MT
- Data validation is generally a bottleneck
- Synthetic data?

New opportunities:

- Longer context window
- Better generalization over multiple languages!
- Grammar and linguistic data use!
- New communities developing

Thank you!
Questions?