Fairness in Multilingual and Multimodal Models

Alexander Fraser

Technical University of Munich, CIT (Computer Science)

Chair for Data Analytics & Statistics

Workshop on Bias in Large Language Models
Heilbronn
2026-06-23

Prof. Dr. Alexander Fraser - Chair for Data Analytics & Statistics

Research at the intersection of **Natural Language Processing** and **Machine Learning**

Historically machine translation and morphological generation

These days mostly focusing on research on multilingual foundational models (e.g., character-level representations, **linguistic knowledge**, **cross-lingual transfer**, domain adaptation, long-context language models, **multilingual alignment**, **human alignment**, …)

BTW, these slides will be available on my personal web page after the talk (alexfraser.github.io)

Fairness in Multilingual and Multimodal Models

Fairness in both textual and multimodal AI requires solving alignment problems simultaneously across languages and cultures

The gap between English-centric AI and the world's multilingual reality is huge!

We'll talk briefly about multilingual alignment, cover three issues involving fairness (cross-lingual hate speech detection, moral reasoning, gender bias in text-to-image models) and then talk about language equality in multilingual models

Large Language Models are Multilingual, but…

GPT and other Large Language Models (LLMs) are often trained on multilingual corpora

However, the main focus on research in LLMs is to try to achieve Artificial General Intelligence (whatever that is)

No one is arguing that multilingual capabilities are necessary for this, so many companies just use old translation benchmarks or similar as a “nice-to-have”

But actually, these abilities are pretty amazing! Even models that see no parallel data learn to translate and can be instructed in one language and carry out a task (like summarization) in another language

In fact, three languages can be involved! (See, e.g., Lai, Mesgar, Fraser - Findings ACL 2024)

Translation and cross-lingual transfer are both thought to be so-called “emergent” abilities (Wei et al 2022), meaning that at a certain amount of computation/data these abilities suddenly appear

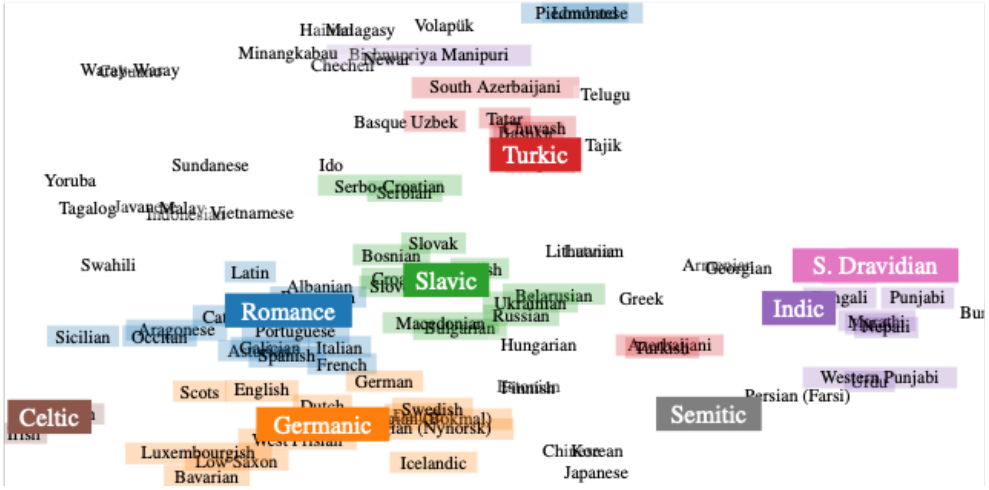
But the amount of data seen for each language is quite uneven! (We’ll come back to this later)

Multilingual LLMs and Multilingual Alignment

We've been interested in multilingual LLMs since mBERT (which was introduced in the BERT paper)

One particular concept we are interested in is Language Neutrality (Libovicky, Rosa, Fraser, Findings EMNLP 2020)

This is the result of encoding 100000 sentences in each of the languages listed, and then clustering them



Multilingual LLMs and Multilingual Alignment

Understanding Cross-Lingual Alignment -- A Survey
Katharina Hämmerl, Jindřich Libovický, Alexander Fraser
Findings ACL 2024

If you have an encoder-only model, language neutrality applies in a straightforward fashion

If you have an encoder-decoder model, you want language neutrality in the encoder, and language to be part of the embedding in the decoder (so that you don't get, e.g., "off-target" outputs in the wrong language)

If you have a decoder-only model, it seems to be the case that lower layers should be language neutral, while upper layers should not be, but we have work in progress trying to verify this (and to determine where to draw the border)

The survey discusses multiple objectives in cross-lingual alignment

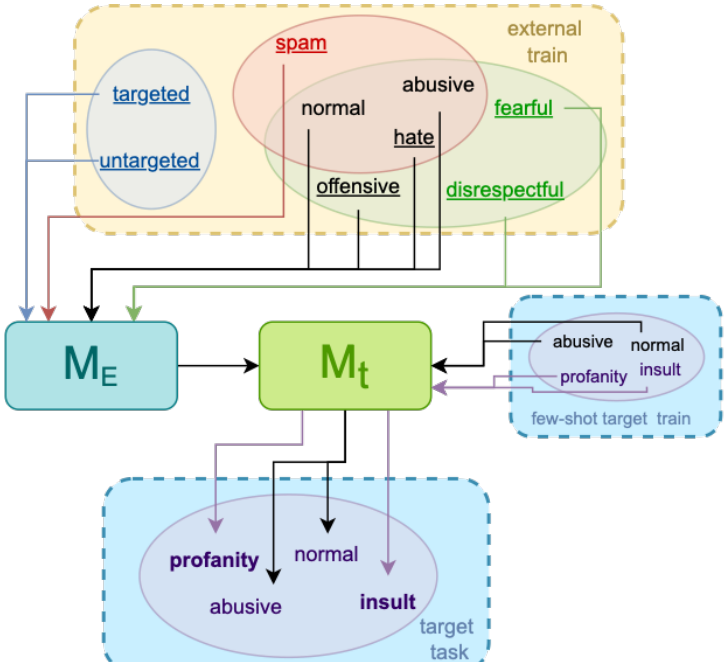
- Main ideas: maximizing full-vector similarity, minimizing language-specific signals, optimizing for zero-shot transfer or task-specific projection
- Overall: selective alignment tailored to downstream tasks often leads to better multilingual performance

Cross-lingual transfer in practice - hate speech detection

Labeled hate speech datasets are overwhelmingly English-centric; collecting and annotating data for low-resource languages is expensive and slow

Speakers of low-resource languages are systematically less protected by automated content moderation

We previously worked on differing hate speech definitions. Can we address this?



Cross-Lingual Transfer Learning for Hate Speech (Bigoulaeva, Hangya, Fraser 2021)

Starting point: train on high-resource source language (English), transfer to low-resource target language (here: German) with no target-language labels

Used cross-lingual word embeddings to bridge the language gap; achieved competitive performance without any target-language annotations

Key challenge identified: label inconsistency across datasets. Different corpora define hate speech differently, which hinders cross-lingual transfer

Extended work (journal 2023): addressed these label issues through modification and bootstrapping strategies for zero-shot cross-lingual transfer; showed good performance can be achieved even across very different hate speech definitions

Fine-Grained Transfer — Label-Specific Soft Prompt Tuning (Ghorbanpour, Hangya, Fraser, NAACL 2025)

The diversity of hate speech types (racism, sexism, threats, etc.) makes coarse transfer insufficient

We introduce label-specific soft prompt tuning: rather than treating all harmful content as one category, we learn separate soft prompts per label type

This allows fine-grained transfer - the model learns which aspects of harmful content transfer well across languages and which are more culturally specific

Results: consistently outperforms standard fine-tuning baselines on harmful content detection across languages

Data-Efficient Cross-Lingual Transfer via Nearest Neighbor Retrieval (Ghorbanpour, Dementieva, Fraser, EMNLP 2025)

Assumption: a very small number of labeled examples in the target language are available

Method: use these few labeled examples to retrieve the most relevant instances from a large multilingual hate speech pool via nearest-neighbor retrieval

Key results across eight languages:

- Consistently outperforms models trained only on target-language data
- In most cases, surpasses state-of-the-art
- Highly data-efficient: as few as 200 retrieved instances can be sufficient
- Scalable: the retrieval pool can be expanded; the method adapts readily to new languages

Takeaways: Hate Speech Detection

Overall:

- Studies on cross-lingual transfer, fine-grained label transfer, data-efficient retrieval, and LLM prompting all try to address hate speech detection for low resource languages
- Poor cross-lingual alignment is not just a benchmark problem, it leaves real communities vulnerable

Key remaining challenges:

- Very low-resource languages with almost no labeled data
- Cultural specificity of hate speech definitions
- LLMs are promising for the future (but not yet best; also issues with safety alignment)

Multilingual Alignment vs Human Alignment

The languages in a multilingual model can be viewed as sub-models

We as a field have just started to understand how, for instance, it is possible that for the same question posed in English and Arabic, we can get two different answers

One interesting thing we can try to study with multilingual models is moral reasoning across languages

Here we are trying to use language as a proxy for culture

Speaking Multiple Languages Affects the Moral Bias of Multilingual Models
(Hämmerl, Deiseroth, Schramowski, Libovický, Rothkopf, Fraser, Kersting — Findings ACL 2023)

Motivation

Multilingual models are trained on vastly unequal amounts of data per language

This raises a fairness question beyond task performance: do models impose English-centric or Western-centric moral norms on other languages?

Two failure modes to worry about:
- The model encodes English moral norms and projects them onto all other languages
- The model encodes random, incoherent moral beliefs in lower-resource languages

Both could cause real harm in cross-lingual deployment

Theoretical Background: Moral Foundations Theory

Moral Foundations Theory (Haidt & Joseph 2004): five foundational moral dimensions
- Care/Harm, Fairness/Reciprocity, Authority/Respect, In-group/Loyalty, Purity/Sanctity

Their relative importance varies across cultures and political contexts

The Moral Foundations Questionnaire (MFQ) has been administered to humans in many countries — giving us a human baseline to compare against

Key question: do multilingual models reflect these cultural differences, or flatten them?

The MoralDirection Framework

We apply the MoralDirection framework (Schramowski et al. 2022) to multilingual models

Originally developed for monolingual English models; we extend it to Arabic, Czech, German, Chinese, and English

MoralDirection identifies a subspace of model weights corresponding to a sense of "right" and "wrong"

Models tested: monolingual models for each language, plus multilingual models (XLM-RoBERTa)

Additional novel contribution: training multilingual sentence transformers by translating existing training data, without requiring a teacher-student setup

Experiment 1: MoralDirection Scores Across Languages

We score action verbs (e.g., kill, steal, love, thank) in each language using MoralDirection

Findings:
- Models do encode a moral dimension in all tested languages
- Scores sometimes align between mono- and multilingual models for the same language — but not always
- Differences between languages exist, but do not cleanly correspond to known cultural differences

The models are highly reliant on lexical cues, making them vulnerable to word sense disambiguation failures (e.g., XLM-R confusing German “schätzen” — "to treasure" — with "to estimate")

Experiment 2: Parallel Corpora (OpenSubtitles)

We test model behavior on parallel Czech-English and German-English subtitle data

Same semantic content, different language — do models score sentences consistently?

Finding: scores on parallel sentences are often inconsistent across languages, suggesting the moral signal is language-specific rather than meaning-specific (because the same statement is judged differently depending on which language it is expressed in)

Experiment 3: Moral Foundations Questionnaire

We ask our models the MFQ questions and compare results to human responses from different countries

Finding: models do not reliably reproduce human cultural differences
- Sometimes a language's model scores align with human data from that culture, but this is inconsistent
- No model reliably distinguishes, say, Arabic cultural moral intuitions from German ones in the way humans do

The models capture some moral signal, but it does not map well onto real cross-cultural variation

Takeaways and Fairness Implications

Multilingual models do encode moral biases, but these are not culturally faithful

Deploying these models cross-lingually may impose dominant-language moral norms on speakers of other languages, or produce incoherent outputs in lower-resource languages

Using language as a proxy for culture is imperfect but interesting

Open challenge: how do we build multilingual models that genuinely respect cultural moral diversity rather than projecting a single dominant norm?

Multilingual Alignment vs Multimodal Fusion

What about multimodality?

This is an even bigger challenge!

How likely is it that we can achieve “fusion” between a picture of a dog barking and the English sentence “The dog is barking.”, if we can’t even ensure that the German translation has the same representation?

Let’s talk about a concrete issue involving multilinguality, multimodality and fairness:

Gender bias in multilingual text-to-image models

Gender bias in multilingual text-to-image models

Fairness in Multimodal Models - Gender Bias in Multilingual Text-to-Image Generation

Friedrich, Hämmerl, Schramowski, Brack, Libovický, Kersting, Fraser — ACL 2025

In this work, we assume we want 50/50 gender distribution across professions

- Text-to-image models don't do this
- They also don't capture the empirical real world distribution well
- Our techniques can be applied to any desired distribution
  - In some applications, the ability to specify this distribution will be useful

en
doctor

de
Doktor

Bias in Multilingual Text-to-Image Models

Text-to-image (T2I) models are now widely used and increasingly multilingual, broadening global access

We show that multilingual T2I models suffer from substantial gender bias

A striking illustration: using the same model, seed, and prompt structure, the German “Doktor” and the English “doctor” produce images that are visually nearly identical except for the apparent gender of the person depicted

The expectation that results should be consistent across languages does not hold

Grammatical gender is an added complication

A key obstacle in multilingual bias evaluation: languages differ fundamentally in how they encode gender

- Languages with gendered nouns: Arabic, German, Spanish, French, Italian
- Languages with gendered pronouns only: English, Japanese
- Languages with no grammatical gender: Korean, Chinese

English "doctor" is gender-neutral; German requires a choice: Doktor (masculine/generic), Doktorin (feminine), Doktor*in (gender-star, inclusive)

MAGBIG — Our Benchmark

We introduce MAGBIG: Multilingual Assessment of Gender Bias in Image Generation

Coverage: 20 adjectives, 150 occupations across 6 categories (healthcare, engineering, hospitality, etc.), translated into 9 languages with native speaker supervision

Prompt types per language: masculine, feminine, gender-neutral/indirect, and (for German) gender-star convention

3,630 prompts total; over 1.8 million images evaluated across 5 multilingual T2I models

Key design principle: control for linguistic differences so that observed biases are intrinsic to the models, not artifacts of translation

Results: Strong Bias, and It Varies by Language

All models show strong skews toward male-presenting images for professional occupations, even with gender-neutral prompts

Bias is not consistent across languages: the same model produces different gender distributions depending on which language the prompt is written in

The generic masculine in languages like German and French does not function as truly gender-neutral, instead it strongly primes male-presenting outputs

Even languages without grammatical gender (Korean, Chinese) show occupation-linked bias, suggesting the models carry English-centric stereotypes into other language spaces

Shallow Strategies are Largely Ineffective

We explore prompt engineering as a mitigation strategy: indirect prompts (avoiding the occupation noun), gender-star convention in German (Doktor*in)

Findings:
- Indirect prompts partially reduce bias but significantly hurt text-image alignment. The model no longer reliably generates the intended occupation
- The gender-star convention is rare in training data and often poorly tokenized, leading to inconsistent results
- No mitigation strategy achieves both fair gender distribution and accurate prompt comprehension
- Surface-level prompt engineering is not a viable solution, bias is deeply entrenched in the models

Takeaways and Fairness Implications

Multilingual T2I models magnify rather than merely reflect gender stereotypes, and they do so inconsistently across languages

Non-English speakers are not getting equivalent outputs — which language you use affects what image you see, even for the same underlying concept

Future work: training data with more diverse language representations, better tokenization of gender-inclusive language conventions, and bias-aware training objectives

MAGBIG is publicly available

LLMS are multilingual, but how multilingual are they?

mBERT supports 104 languages

Many closed models (e.g. GPT-4o) support around this number too

Open-weight-only models (e.g., META Llama3.1 8B) also, a few support 200 or more (but usually badly)

Unfortunately, there are (arguably) **no open-data or truly open-everything/open-source massively multilingual models**

(Apertus is promising here)

But how multilingual are big LLMs really? (For instance, ChatGPT?)

As an example, consider tokenization

Heavy optimization for English leads to significantly longer token sequences and higher representation costs for low-resource languages (real performance decreases!)

Language Death and Languages on the Web

Approximately one language death every two weeks

7,000+ languages spoken worldwide, nearly 40% are considered endangered
- UNESCO identifies about 2,500 languages as endangered
- SIL International classify about 1,500 as dying

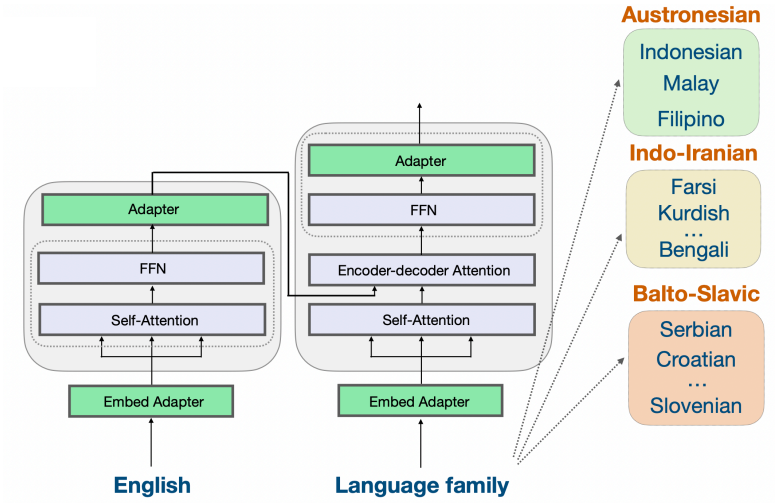
By the end of the 21st century, somewhere from 50% to 90% of current languages could become extinct if no preservation efforts are made

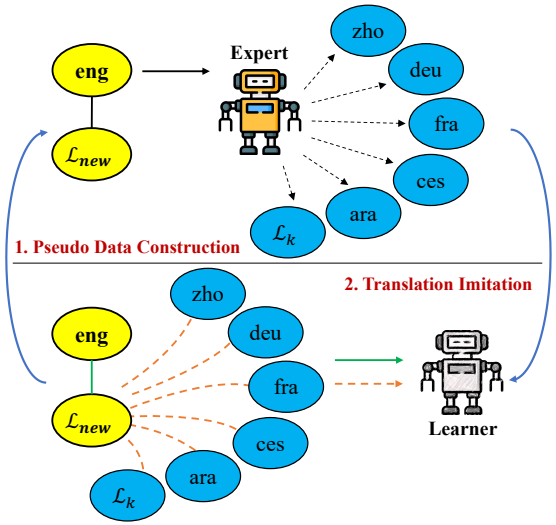
IMO technology is primarily currently playing a role in boosting the visibility and prestige of lower-resource languages, important for ensuring that **children actually use these languages**!

Statistics from a recent Common Crawl (html data only):
- English is 1.08 billion pages and is 43% of the data
- Next 5 (other UN languages - Arabic + German + Japanese) about 31% total
- 23% is other 110 detectable languages (including Arabic)
- 3% unknown are (some of) the other approximately 7000 languages

EPICAL (ERC Advanced Grant)

- EPICAL: Evaluating and Programming Intelligent Chatbots for Any Language
- Intelligent chatbots such as ChatGPT work well for a few languages such as English
  - But not for most of the 7099 languages currently spoken on Earth
  - Chatbots are trained on corpora like the Common Crawl
  - 97% of the crawl is top-100 languages, just 3% of the crawl are from the 7000 less-resourced languages
  - High risk, high return idea:
    - First create high quality texts in low-resource languages with the help of chatbots
    - Then use these texts to improve the chatbots (creating a virtuous cycle)
  - Builds on contributions in many areas of machine translation, NLP, machine learning (with low resources)
  - As well as work on many low resource languages, e.g., Upper Sorbian (Germany) and Hiligaynon (Philippines)
- One major problem (and big interest) is morphology…





The Relationship of Tokenization and Orthographic Knowledge

We suppose that LLMs can only generalize to morphology if they can access orthographic knowledge (the characters in their subwords)

This leads us to develop a new benchmark:

CUTE: A Benchmark for LLMs' Understanding of Their Tokens
Lukas Edman, Helmut Schmid, AF, EMNLP 2024

Basic idea: we will prompt LLMs to do orthographic operations and see how they do

We also do a similar task at the word level to see if this type of prompt is understood

The Relationship of Tokenization and Orthographic Knowledge

CUTE: A Benchmark for LLMs' Understanding of Their Tokens
Lukas Edman, Helmut Schmid, AF, EMNLP 2024

LLMs:

State-of-the-art, available (best are Llama3 70B, Command R+). All of these use subwords!

Example tasks:

Swap letters, e.g., given input: **alphabet**, swap "a" and "b". Desired output: **blphbaet**

Sanity check: Swap words, e.g., given input: **the sky is blue**, swap "is" and "the". Desired output: **is sky the blue**

Prompting:

4-shot seems to be enough (see the paper for further details)

The Relationship of Tokenization and Orthographic Knowledge

Reflections on orthography and tokenization in the age of LLMs

- It seems likely that pretraining doesn't lead to learning of orthographic information
- While newer LLMs are better, probably this isn't *emergent*
- It would be great to train state-of-the-art LLMs on the character level
  - Clearly these LLMs will be better at these tasks
  - We suspect they won't be worse on many long-distance tasks, but YMMV
- It's frustrating that there aren't more obvious morphological problems with subword tokenization
  - Are they still learning morphological generalization somehow?
  - Mostly yes, but there are still measurable problems!

Tokenization and Linguistic Knowledge

(ein)pflanzen: *'to plant (in)'*):

word	GPT	ling. sound
einpflanzen	e|inp|fl|an|zen	ein|pflanz|en
eingepflanzt	eing|ep|fl|an|zt	ein|ge|pflanz|t
pflanzte	p|fl|anz|te	pflanz|te
pflanzen	p|fl|an|zen	pflanz|en
pflanztet	p|fl|an|zt|et	pflanz|tet

GPT4 segmentation of (ein)pflanzen

Subword Segmentation in LLMs: Looking at Inflection and Consistency. Marion Di Marco / AF. EMNLP 2024

The Relationship of Tokenization and Linguistic Knowledge

Subword Segmentation in LLMs: Looking at Inflection and Consistency
Marion Di Marco / AF. EMNLP 2024

While we are waiting for state-of-the-art character/byte models, let's take a look at subword tokenization

Can it really be the case that "linguistically bad" tokenization doesn't hurt morphological generalization in LLMs? No, this is wrong, it does hurt!

The Relationship of Tokenization and Linguistic Knowledge

Subword Segmentation in LLMs: Looking at Inflection and Consistency
Marion Di Marco / AF. EMNLP 2024

			DE	SV	FR	IT	ES	PT	FI	HU	CS
freq > 500	highOverlap	zero shot	197	190	196	193	200	200	186	200	182
	lowOverlap	zero shot	189	175*	184*	191	191*	188*	180	182*	179
	highOverlap	one shot	191	194	194	189	200	200	186	200	189
	lowOverlap	one shot	185	185	187	195	191*	197	180	185*	185
freq $\leq \leq 10$	highOverlap	zero shot	188	174	187	196	198	192	180	174	178
	lowOverlap	zero shot	166*	131*	161*	171*	169*	160*	130*	156*	144*
	highOverlap	one shot	189	175	184	195	199	193	185	181	180
	lowOverlap	one shot	172*	140*	172	172*	177*	176*	122*	163*	148*

Table 6: Number of correctly **generated forms** (N=200) contrasting *segmentation consistency*. * marks significant difference between high/low overlap sets (χ-square test with a significance level of α=0.05)

EPICAL shared tasks and workshops

If you know any language activists (for minority languages), please point them to our online workshops on language technologies for language activists!

My group is currently organizing a shared task on Question Answering, Machine Translation and Mathematical Reasoning for Upper Sorbian and for Ukrainian on smaller LLMs (3B parameter limit), please participate!

Conclusion

We talked about language neutrality in multilingual LLMs as well as a followup survey on multilingual alignment

We talked about cross-lingual transfer, studying cultures through multilingual models, studying gender biases in text-to-image models

Then we then talked about language death and my ERC Advanced Grant EPICAL

Finally we talked briefly about linguistic fairness (e.g., in tokenization, ability to deal with rich morphology)

Future Work

We are interested in work on social good in NLP in general, see our survey paper on this

We want the ability to talk with a LLM about language, both in terms of accessing linguistic generalizations and in terms of model editing of linguistic generalizations (initial papers on this with Marion Di Marco, Tsedeniya Temesgen)

We are interested in broadening our work on cross-lingual analysis and generation (e.g. to Question Answering, generating Wikipedia pages, etc)

New team members working on calibration and confidence estimation, both for conventional classification and for generative tasks

Finally, we are also interested in multilingual reasoning models

Thank you, and a big thanks to my research group!