# Representing Multiple Languages in Large Language Models

Alexander Fraser

Technical University of Munich, CIT (Computer Science)

Chair for Data Analytics & Statistics

Interdisciplinary School on Machine Learning and Artificial Intelligence for Science
Heilbronn
2025-06-24

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

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

Strong emphasis on multilingual representation learning (large language models) and machine translation

Relevant for today's presentation:

Historically focused on translation of morphologically rich languages

Brief interlude focusing on so-called unsupervised machine translation and multilingual embeddings

These days 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)

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

Located at the new multi-university “Bildungscampus” in Heilbronn

Technical University of Munich (TUM), ETH Zurich, BW Graduate Center Applied AI

TUM’s Heilbronn presence: 15 CS/CE professors and 15 Management Professors

# Many Large Language Models are Multilingual!

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)

There aren't any people 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)

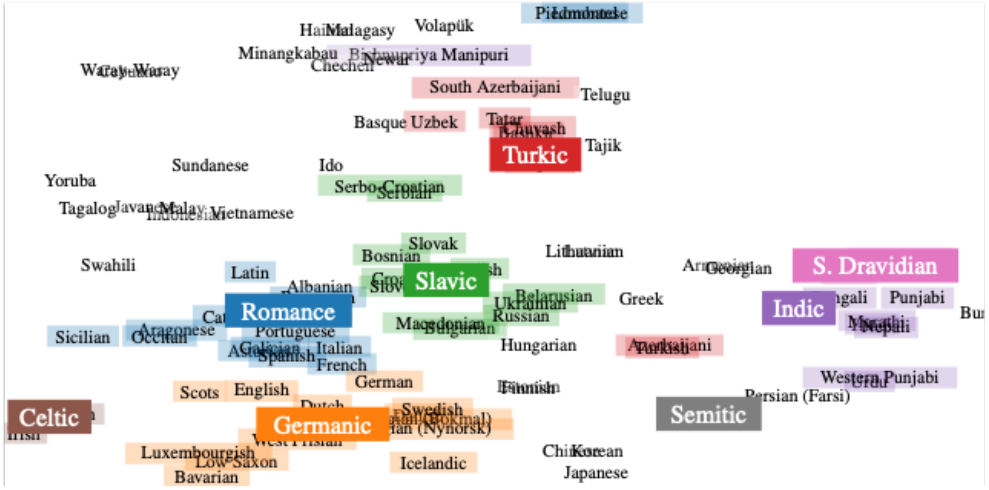
Translation and cross-lingual transfer are both thought to be so-called "emergent" abilities (Wei et al 2022), meaning that at a certain amount of computation/data these abilities suddenly appear

# 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

# Multilingual Alignment vs Multimodal Fusion

What about multimodality?

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?

Now let’s switch gears and talk about language coverage (particularly low resource), and then morphologically rich languages

# 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

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

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

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

Let's talk briefly about tokenization in machine translation, and then come back to this multilinguality issue

# Brief history of machine translation

Before 2000: rule-based systems based on parsing the source language sentence with a grammar

From 2000-2016: “statistical” machine translation, based on a research program from IBM in the 1990s.

- Played a pivotal role in moving natural language processing to be an important area of machine learning, and perhaps a key role in making machine learning mainstream

Statistical Machine Translation is a supervised structured prediction problem:

- Generate a translation word-by-word
- Simultaneously maximize a “translation” model (providing possible rough mappings from source language to target language) and a “language” model (selecting good target language sentences)
- Search problem is difficult and requires inadmissible search heuristics and other optimizations, was also slow

# Present: Somewhat Multilingual Chatbots

Amazingly, ChatGPT is not explicitly trained on parallel data, but can translate quite well

- We are still studying the mechanisms by which this happens in LLMs

For instance, DeepL is still recommended for English to German translation, but in many situations ChatGPT can be a better choice

- Particularly if you need to control the (world) context, or apply global constraints to the output
- Simple example: generating using “du” or “Sie” in German. ChatGPT can even make the style of the language sound more formal if you choose “Sie”
- (However, DeepL also translates powerpoint slides quite well!)

GPT is trained on a very high percentage of English, and a very low percentage of other languages

- This is exactly why most of you use ChatGPT in English (and not your native language if it is not English)

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

# EPICAL: Evaluating and Programming Intelligent Chatbots for Any Language

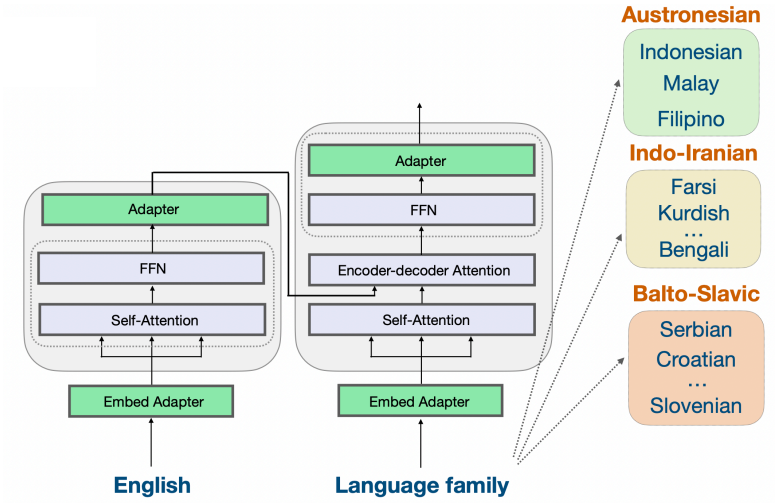
ERC Advanced Grant, recently started (in hiring phase now)

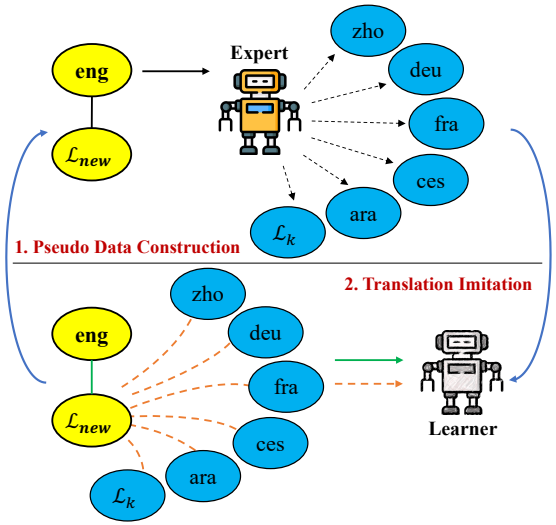
Ties to previous work on translation to morphologically rich languages, multilingual alignment, domain adaptation

Basic idea is to create a virtuous cycle involving native speakers quickly creating text in low resource languages with the help of chatbots, and then improving the chatbots using this text

# 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…



# Neural Machine Translation and Morphology

The era of Neural Machine Translation (NMT) was from 2017 to about 2022

NMT is of course still standard supervised learning - we have millions of sentence pairs for English to German, still solve a (simpler) training problem, to translate still solve a (simpler) structured prediction search problem

From the beginning, it seemed obvious to me that we need access to the character level to solve morphological problems

These models should learn word endings!

For instance: “Ein weißes Haus” vs. “Das Weiße Haus”

Surely the field will agree with me here…

# Why Don’t People Use Character-Based Machine Translation?

Jindrich Libovicky, Helmut Schmid, AF
Findings ACL 2022

This paper is a survey of the work on character-based MT

It briefly covers RNN models and then focuses on transformer models

It seems obvious that character-based models *which learn morphological generalization* should be the solution to all of our problems!

I want to believe!

But they haven’t caught on

BPE and WordPiece are standard in both NMT (focus right now) and in LLMs (later in the talk)

# Why Don’t People Use Character-Based Machine Translation?

Jindrich Libovicky, Helmut Schmid, AF
Findings ACL 2022

This paper traces the development of character-based NMT from the very influential paper of Lee, Cho, Hofmann in 2017 (using CNNs) to our own model, based on downsampling in the decoder

# Why Don’t People Use Character-Based Machine Translation?

Jindrich Libovicky, Helmut Schmid, AF
Findings ACL 2022

But people don’t use these approaches. Why?

Results have been mixed

- We haven’t been able to find measurable morphological generalization as we wanted

Efficiency is a big problem, but downsampling may have helped to solve this

I actually believe that supervision is the main problem

- There isn’t a good way to train on huge amounts of monolingual data (best known way is back translation, but this propagates errors!)

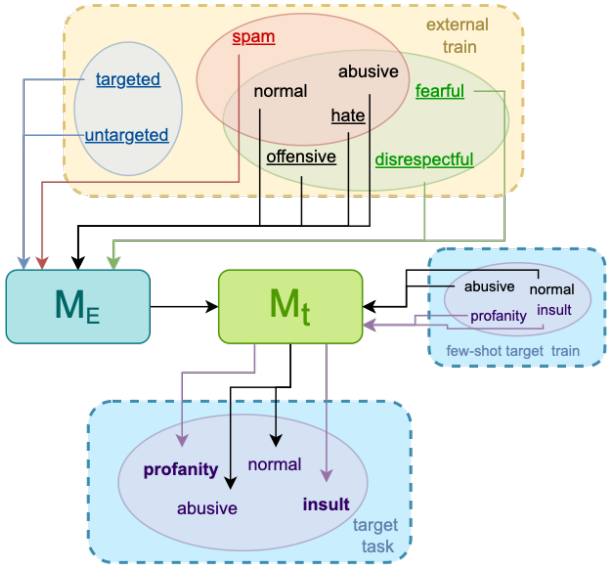
This leads us to other models… Which ones?

# Interlude: Multilingual LLMs and Alignment

NMT was a strong research area in 2017 (the year that the encoder-decoder Transformer was introduced) but arguably BERT (which is a transformer encoder) started to put an end to this

Standard encoder-decoder transformers were state-of-the-art in translation until 2022

At the same time large transformer encoders are still state-of-the-art in many of the cross-lingual classification tasks we work on such as hate speech detection (and presumably many other classification tasks)



# Interlude: Multilingual LLMs and Alignment

Multilingual LLMs seem very promising for solving the morphological generalization problem!

They are trained on massive amounts of text and seem to be able to answer questions about morphology. They also show some degree of morphological generalization.

All known instruction-tuned LLMs are trained using subwords though! But maybe this isn't a problem for morphological generalization?

# 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:

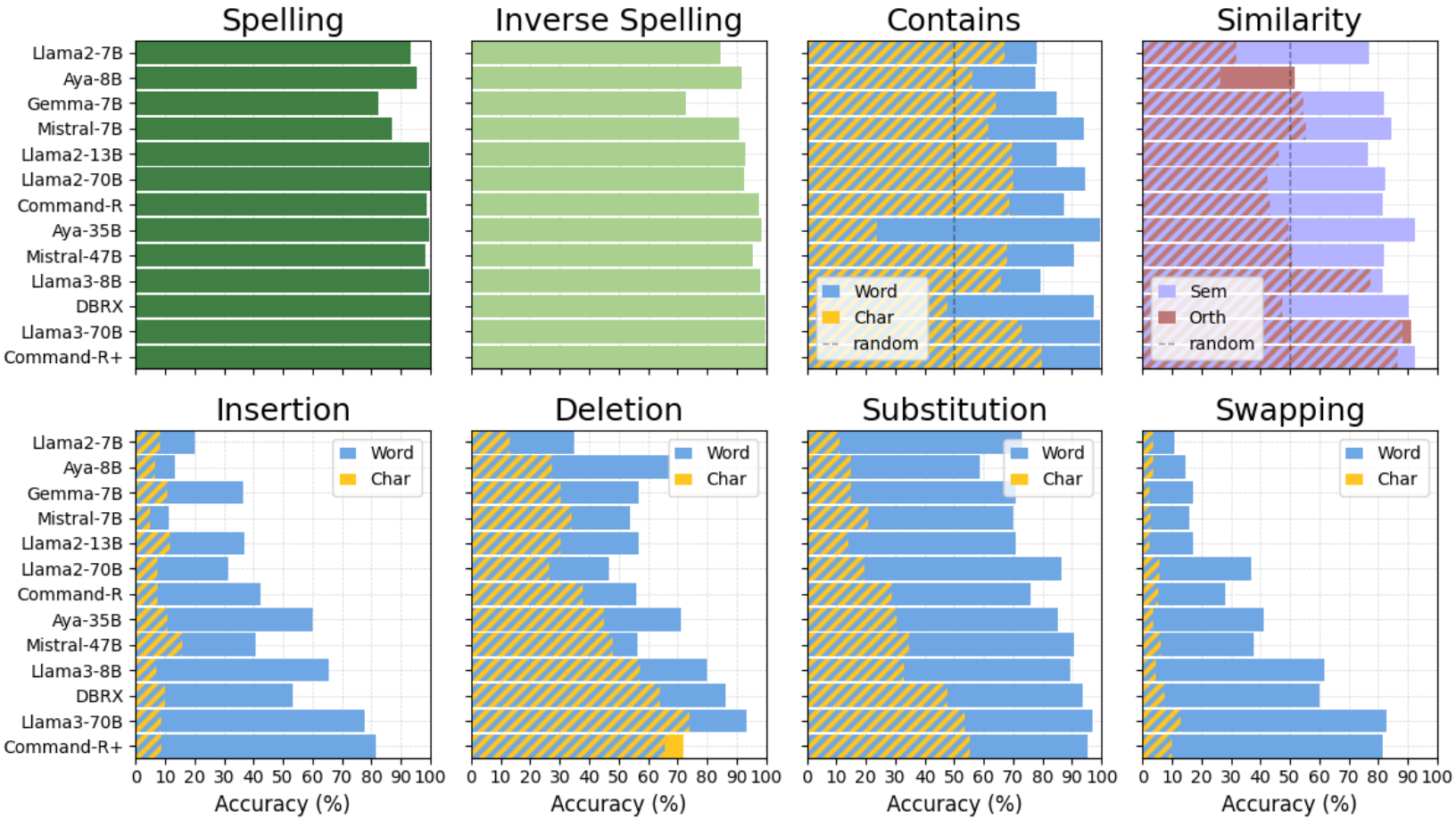
4-shot seems to be enough (see the paper for further details)

# The Relationship of Tokenization and Orthographic Knowledge

CUTE: A Benchmark for LLMs' Understanding of Their Tokens

| | Task | Input | Output |
|---|---|---|---|
| Composition | Spelling | Spell out the word: there | t h e r e |
| | Inverse Spelling | Write the word that is spelled out (no spaces): t h e r e | there |
| | Contains Character | Is there a 'c' in 'there'? | No |
| | Contains Word | Is there a 'the' in 'the sky is blue'? | Yes |
| Similarity | Orthographic Similarity | Which is closer in Levenshtein distance to 'happy'? glad or apply | apply |
| | Semantic Similarity | Which is more semantically related to 'happy'? glad or apply | glad |
| Manipulation | Character Insertion | Add 'b' after every 'e' in 'there' | thebreb |
| | Word Insertion | Add 'is' after every 'the' in 'the sky is blue' | the is sky is blue |
| | Character Deletion | Delete every 'e' in 'there' | thr |
| | Word Deletion | Delete every 'the' in 'the sky is blue' | sky is blue |
| | Character Substitution | Replace every 'e' with 'a' in 'there | thara |
| | Word Substitution | Replace every 'the' with 'is' in 'the sky is blue' | the is sky is blue |
| | Character Swapping | Swap 't' and 'r' in 'there' | rhete |
| | Word Swapping | Swap 'the' and 'is' in 'the sky is blue' | is sky the blue |

# The Relationship of Tokenization and Orthographic Knowledge



# 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 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 ($\chi$-square test with a significance level of $\alpha$=0.05)

# 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 and Machine Translation for Upper Sorbian and for Ukrainian on reasonably sized LLMs (3B parameter limit), please participate!

Hartwig Anzt and others are organizing a workshop on parallel processing for scientific computing (next slide)

Allies at the CommonCrawl, ML-Commons, others are organizing a shared task on web scale language identification - please consider contributing annotations and/or submitting a system (slide after that)


SIAM Conference on Parallel Processing
for Scientific Computing (PP26)
Berlin, March 3-6, 2026


# 1st Workshop on Multilingual Data Quality Signals

Palais des Congrès

Montreal, Canada

**10 October 2025**

## Shared Task

We invite submissions to the first Shared Task on **Language Identification for Web Data**.

### Key dates

**1st Deadline to contribute annotations**: July 7, 2025

**1st Annotations released (train split):** July 14, 2025

# Conclusion

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

We then talked about language death and my ERC Advanced Grant EPICAL

Then we talked about why character-based machine translation hasn't taken over, despite being one obvious answer to tokenization issues

Following this we discussed how to focus on orthographic knowledge in LLMs. We showed that orthographic knowledge is very poor in SOA LLMs

Finally we talked about the relationship of tokenization to simple linguistic knowledge. In a study on ChatGPT4o, we showed that consistent segmentation of the stem is critical

In near future work we will:

Extend our work on orthography to more models and look at multilinguality

Use curriculum learning to fine-tune SOA LLMs to character representations (Towards Reasonably-Sized Character-Level Transformer NMT by Finetuning Subword Systems. Jindřich Libovický, Alexander Fraser. EMNLP 2020)

We are also interested in 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, Marion Di Marco, Tsedeniya Temesgen)

In the very distant future (maybe 2026?), we plan to train a SOA character-based LLM (if we can get enough GPUs)

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