Decoding the Polyglot: Understanding Cross-Lingual Learning in Language Models

Ali Basirat

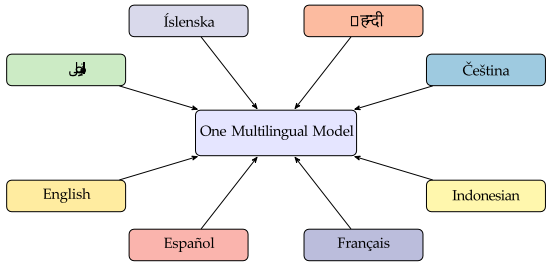
Centre for Language Technology

alib@hum.ku.dk

UNIVERSITY OF COPENHAGEN

Multilingual NLP

- Goal: Build one model capable of handling many languages.



One model capable of handling many languages.

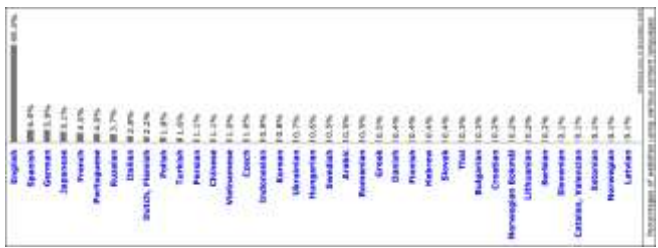
Traditional Approaches to Multilingual NLP I

- Universal Grammar (linguistic hypothesis):
  - Search for shared structural principles across languages.
  - Inspired early multilingual parsing and rule-based systems.
- Massively Parallel Data:
  - Train translation models using aligned corpora (e.g., Europarl, UN).
  - Relied on explicit word/phrase alignment and bilingual dictionaries.
- Multilingual Word Embeddings (2010s):
  - Align monolingual embedding spaces via linear maps or joint training.
  - Examples: MUSE, multilingual fastText.
- Multilingual Pretraining (late 2010s):
  - Shared encoder across languages with subword vocabularies.
  - Examples: mBERT, XLM, XLM-R.

Traditional Approaches to Multilingual NLP II

Key Limitations:

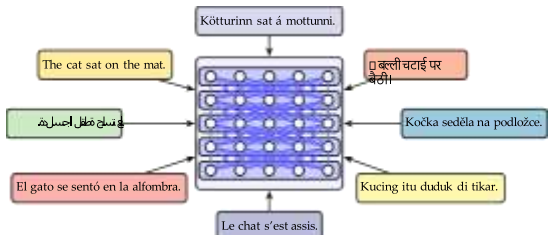
- Data imbalance: over 7,000 languages exist, but fewer than 100 have substantial digital corpora; over 90% of web text is in just 10 languages.[1]
- Typological diversity: huge variation in morphology (isolating vs. agglutinative), syntax (SVO vs. SOV), and script (Latin, Arabic, Devanagari, Cyrillic, etc.).



[1]The State and Fate of Linguistic Diversity and Inclusion in the NLP World (Joshi et al., ACL 2020); W3Techs (2025)

Modern Approaches – Multilingual LLMs

- Training on multilingual corpora enables Large Language Models (LLMs) to understand and generate text in many languages.
- But what mechanisms allow this multilingual capability to emerge?



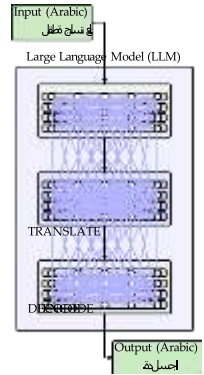
Large Language Model (LLM). Train to predict $P(w_t \mid w_1, \ldots, w_{t-1})$

Common Approaches to Studying Multilingualism in LLMs

- Behavioral Investigation
  - Evaluate outputs across languages (translation, transfer, code-switching).
  - Risk: Illusion of reasoning from surface patterns.
- Mechanistic Interpretability
  - Trace internal computations (layers, heads, neurons).
  - Tools: Logit lens, activation patching.
- Representational Analysis
  - Probe hidden states for cross-lingual alignment.
  - Methods: Probes, similarity metrics.

Dominance Hypothesis

- Behavioral evidence suggests that models translate inputs into a dominant language, process them, and translate back.
- Whether such a mechanism exists in internal representations remains unclear.



Don't Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs (Zhang et al., EMNLP 2023)

Working Language

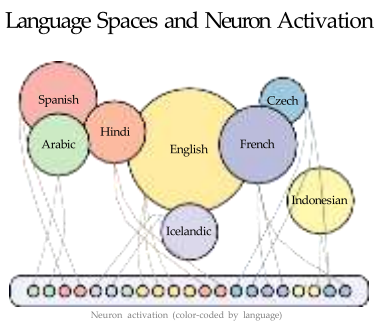
- Mechanistic interpretation of intermediate activations suggests that certain aspects of language are processed through a working language, while others are handled in the target language.

Layer	القطقل	في	تسلاج	?
1	<bos>	the	cat	...
2	the	sat	..	mat
3	cat	sat	on	the
4	the cat	mat	...	...
5	القطقل	sat	on	mat
6	القطقل	تسلاج	on	...
7	القطقل		في	..تسلاج
8	القطقل		في	الجسلدقتسلاج
9	القطقل		في	الجسلدقتسلاج
10	القطقل		في	الجسلدقتسلاج

Do Llamas Work in English? On the Latent Language of Multilingual Transformers (Wendler et al., ACL 2024)

The Language Space

- Representational analysis investigates the neural activations in an LLM.
- Language space: subregion of a model layer's intermediate representations.
- Enables us to investigate model dynamics from an information theoretic perspective.



Language Space – An Evaluation Testbed

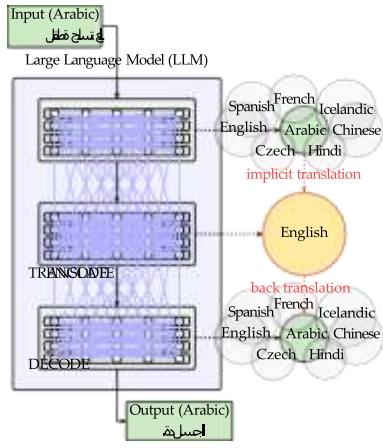
The study of language spaces opens path to answer questions like:

- Multilingual Learning Strategists
- Linguistic investigation from an LLM perspective
- Computational cognitive aspects of multilingual learning

Language Dominance and Language Space

- Implicit translation entails a shift from a source language space to a dominant language space.

Input ⇒ Mediating Lang ⇒ Output



Language Dominance in Multilingual Large Language Models (Shani & Basirat, BlackboxNLP 2025) [BEST PAPER]

Language Dominance Metric

- Goal: Quantify the extent to which the embedding space of a target language T is activated when processing tokens from a source language S.
- Formulation: Expected posterior probability of T given token representation $\tilde{\mathbf{h}}$ from S's embedding space:

$$D(S \rightarrow T) = \mathbb{E}_{\tilde{\mathbf{h}} \sim P(\tilde{\mathbf{h}}|S)} P(T \mid \tilde{\mathbf{h}})$$

- Interpretation: Higher $D(S \rightarrow T)$ indicates stronger alignment of S's representations with T's space.

Language Dominance in Multilingual Large Language Models (Shani & Basirat, BlackboxNLP 2025)

Test Models

- BLOOM – a 1.7B-parameter multilingual language model trained on 46 languages, designed for balanced performance across diverse scripts and language families.

- mGPT – a 1.3B-parameter multilingual GPT-style model trained on Wikipedia and other web corpora across 61 languages, optimized for general-purpose multilingual text generation.

Test Languages

- Paralle Universal Dependencies: parallel corpora, balanced, and high-quality samples across all languages.
- Includes both seen and unseen languages during pre-training.

Language	Family	Genus	ISO	Train
Arabic	Afro-Asiatic	Semitic	ar	✓
Czech	Indo-European	Slavic	cs	–
English	Indo-European	Germanic	en	✓
French	Indo-European	Romance	fr	✓
Hindi	Indo-European	Indo-Aryan	hi	✓
Icelandic	Indo-European	Germanic	is	–
Indonesian	Austronesian	Malayo-Poly.	id	✓
Portuguese	Indo-European	Romance	pt	✓
Spanish	Indo-European	Romance	es	✓

Results – Language Dominance

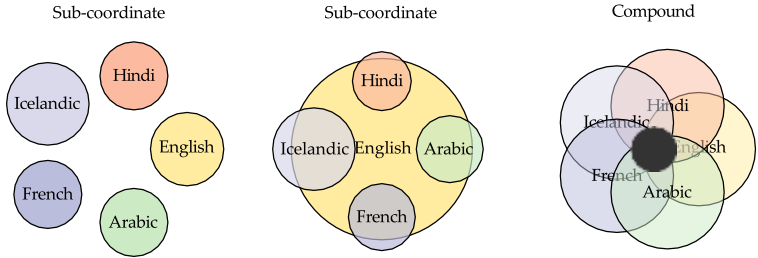
- Dominance scores are balanced (10–15%), suggesting shared, distributed representations in middle layers.
- No strong evidence of a single dominant language was found.
- Dominated tokens are mostly function words or orthographically similar items; content words remain in their own language space.

Alignment with Human Multilingualism I

- Coordinate Learning: distinct linguistic environments and distinct language processing systems.
- Sub-coordinate learning: late acquisition and mental translation into a dominant language.
- Compound learning: shared linguistic environments and universal understanding of language.

Three types of bilingualism (D'Acierno and Rosaria, English as a Foreign Language 1990)

Alignment with Human Multilingualism II



Conceptual models of language spaces representing three type of human bilingualism.

Multilingual Learning Strategies in Multilingual Large Language Models (Basirat, MRL 2025)

Linguistic Information in Language Spaces I

Variational-usable information to probe the discriminative aspects of the activations:

- Activation vector $\mathbf{h}^k$ taken from a layer k is passed through a PCA to reduce its dimensions.
- The amount of variational-usable information of the PCA-reduced activation $\hat{\mathbf{h}}^k$ is measured based on a target task (Y) (language or UPOS indetification)

$$I_{nv}(Y; \hat{\mathbf{h}}^k) = 1 - \frac{H(Y/\hat{\mathbf{h}}^k)}{H(Y/\Phi)}$$

A Theory of Usable Information under Computational Constraints (Xu et. al., ICLR 2020)

Linguistic Information in Language Spaces II

Gradient-weighted Class Activation Mapping (Grad-CAM) to probe the generative aspects of the activations:

- For a feature h_i^k at layer k, importance is computed as:

$$c_i^k(t_j) = \text{ReLU}\left(h_i^k(t_j) \cdot \frac{\partial f(t_{j+1})}{\partial h_i^k(t_j)} \right)$$

- $h_i^k(t_j)$: activation value for input token t_j
- $f(t_{j+1})$: model logit for the next-token prediction
- Layer Differentiation Rate: the proportion of features in a layer that contribute significantly to predicting a target word within a target category.

Decoding layer saliency in language transformers (Hou and Castanon, ICLR 2022)

Test Languages

- 1000 aligned sentences across multiple languages.
- 21 typologically diverse languages.
- Data comes form Parallel Universal Dependencies (PUD).

Language	ISO	Family	Size	mBERT	mGPT	BLOOM	XLMR
Arabic	ar	Afro-Asiatic	20K	✓	✓	✓	✓
Chinese	zh	Sino-Tibetan	21K	✓	✗	✓	✓
Czech	cs	IE Slavic	18K	✓	✗	✗	✓
English	en	IE Germanic	21K	✓	✓	✓	✓
Finnish	fi	Uralic	15K	✓	✓	✗	✓
French	fr	IE Romance	25K	✓	✓	✓	✓
Galician	gl	IE Romance	25K	✓	✗	✗	✓
German	de	IE Germanic	21K	✓	✓	✗	✓
Hindi	hi	IE Indo-Aryan	23K	✓	✓	✓	✓
Icelandic	is	IE Germanic	18K	✓	✗	✗	✓
Indonesian	id	Austronesian	19K	✓	✓	✓	✓
Italian	it	IE Romance	25K	✓	✓	✗	✓
Japanese	ja	Japonic	28K	✓	✓	✗	✓
Korean	ko	Koreanic	16K	✓	✓	✗	✓
Polish	pl	IE Slavic	18K	✓	✓	✗	✓
Portuguese	pt	IE Romance	24K	✓	✓	✓	✓
Russian	ru	IE Slavic	19K	✓	✓	✗	✓
Spanish	es	IE Romance	23K	✓	✓	✓	✓
Swedish	sv	IE Germanic	19K	✓	✓	✗	✓
Thai	th	Kra-Dai	22K	✓	✓	✗	✓
Turkish	tr	Turkic	17K	✓	✓	✗	✓

CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (Zeman et al., CoNLL 2017)

Test Models

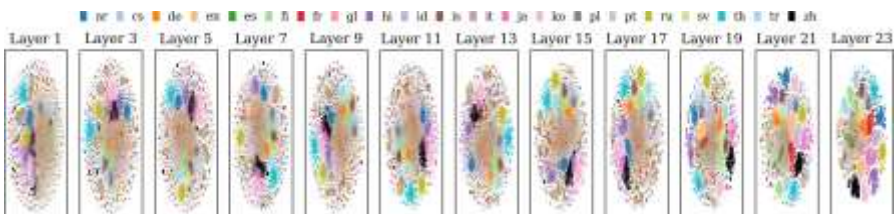
- Multilingual LLMs of different size, architecture, and language coverage.
- Encoder-only and decoder-only models.
- Freely available models from Huggingface.

LLM	Size	#Layers	#Features	Lang. Diversity	Lang. Coverage
BLOOM	1.7B	24	1536	46	17%
mGPT	1.3B	24	2048	61	28%
mBERT	172M	12	768	104	100%
XLMR-base	270M	12	768	100	100%
XLMR-large	550M	24	1024	100	100%

Results – Alignment with Human Multilingualism I

Coordinate learning: Language-specific feature space:

- Dominates multilingual representation in LLMs.
- Decoder-only models develop strongly separated language spaces.

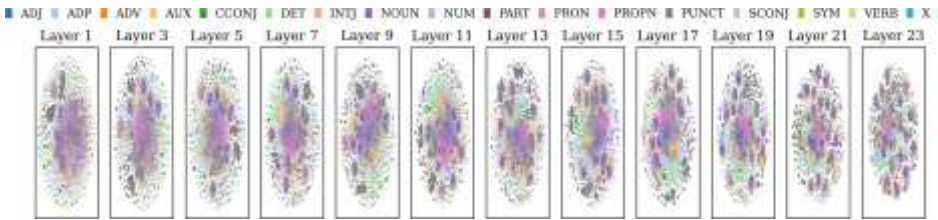


Language spaces in the intermediate layers of BLOOM.

Results – Alignment with Human Multilingualism II

Compound learning: Universal feature space:

- Weakly observed in encoder-only models.
- Universal aspects of language are processed within language spaces (supporting coordinate learning).

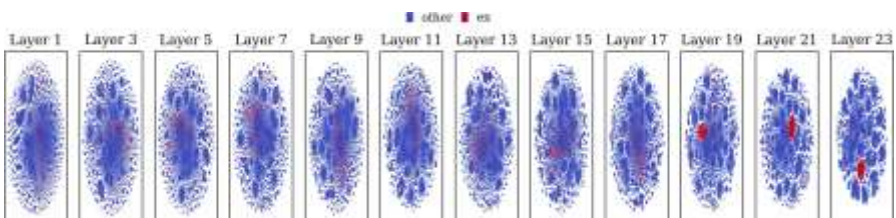


BLOOM's intermediate feature spaces representing UPOS tags.

Results – Alignment with Human Multilingualism III

Sub-coordinate learning: Mediating language:

- No evidence in decoder-only models.
- Only weak signs in encoder-only models.



English vs non-English language spaces in the intermediate layers of BLOOM.

Linguistic Theories I

- Expensive assessment of language universals through fMRI scans.
- Language models activations correlate with human brain activations.
- Could they then be a test bed for linguistic theories?

The neural architecture of language: Integrative modeling converges on predictive processing (Schrimpf et al., PNAS 2021)

Alignment of Brain Embeddings and Artificial Contextual Embeddings in Natural Language Points to Common Geometric Patterns (Goldstein et al., Nature 2024)

Linguistic Theories II

Grammatical Gender:

- A nominal classification system.
  - Masculine/Feminine (Arabic and Italian)
  - Masculine/Feminine/Neuter (German, Greek, and Russian)
  - Common/Neuter (Danish and Swedish)
- Looks arbitrary and language dependent.
  - German: das Mädchen ('the$_{\text{NEUT}}$ girl') – neuter despite referring to a female.
  - French: le soleil ('the$_{\text{MASC}}$ sun') vs. German: die Sonne ('the$_{\text{FEM}}$ sun')
- RQ: Are there some universal principles behind the assignment of grammatical gender to nouns?
- RQ: Is grammatical gender a semantic or syntactic concept?

Linguistic Theories III

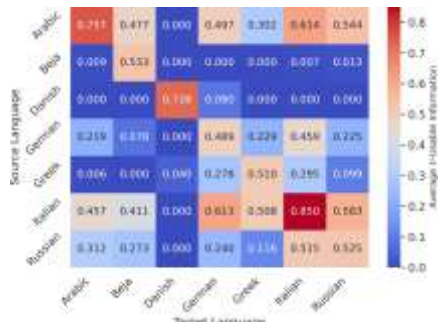
Universalism supports transfer learning

- If the concept of grammatical gender is universal then its knowledge should be transferable across languages.
- Static word embeddings provide some evidence on the universality of grammatical gender.

Cross-lingual Embeddings Reveal Universal and Lineage-Specific Patterns in Grammatical Gender Assignment (Veeman et al., CoNLL 2020)

Linguistic Theories IV

- Universal pattern of grammatical gender is also evident in large language models.
- Transfer learning across layers:
  - LLMs encode grammatical gender together with other semantic features.
  - They knowledge of grammatical gender is transferable (to some extend) to unseen languages.



Universal Patterns of Grammatical Gender in Multilingual Large Language Models (Schröter & Basirat, MRL 2025)

Conclusions

- Not a significant mediation by a dominant language.
- Language-specific processing is more evident.
- LLMs a test beds for assessing linguistic theories.

Thank You for Your Attention!