UNIVERSITY OF TORONTO

HARVARD UNIVERSITY

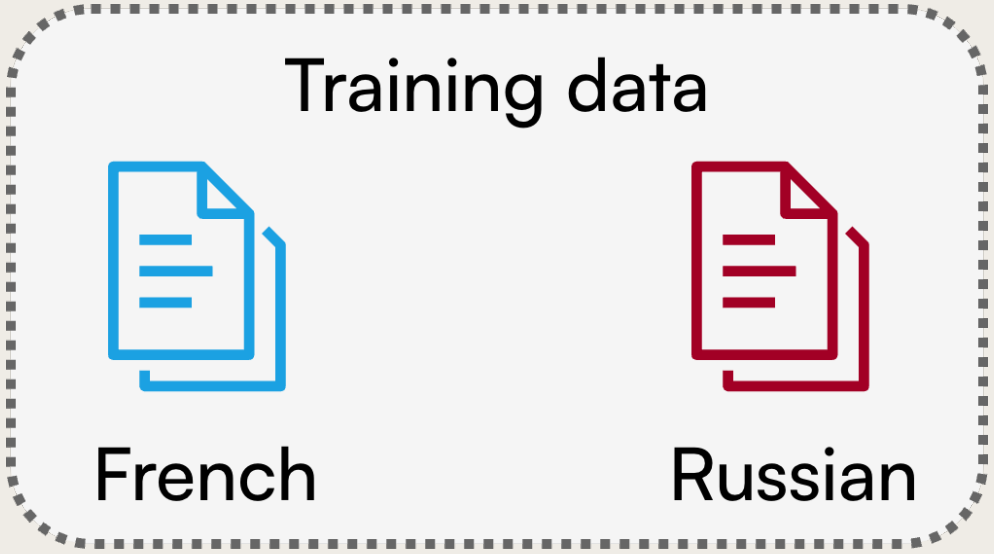
Carnegie Mellon University

UNIVERSITEIT GENT

Modality Matching Matters:
Calibrating Language Distances for Cross-Lingual Transfer in URIEL+

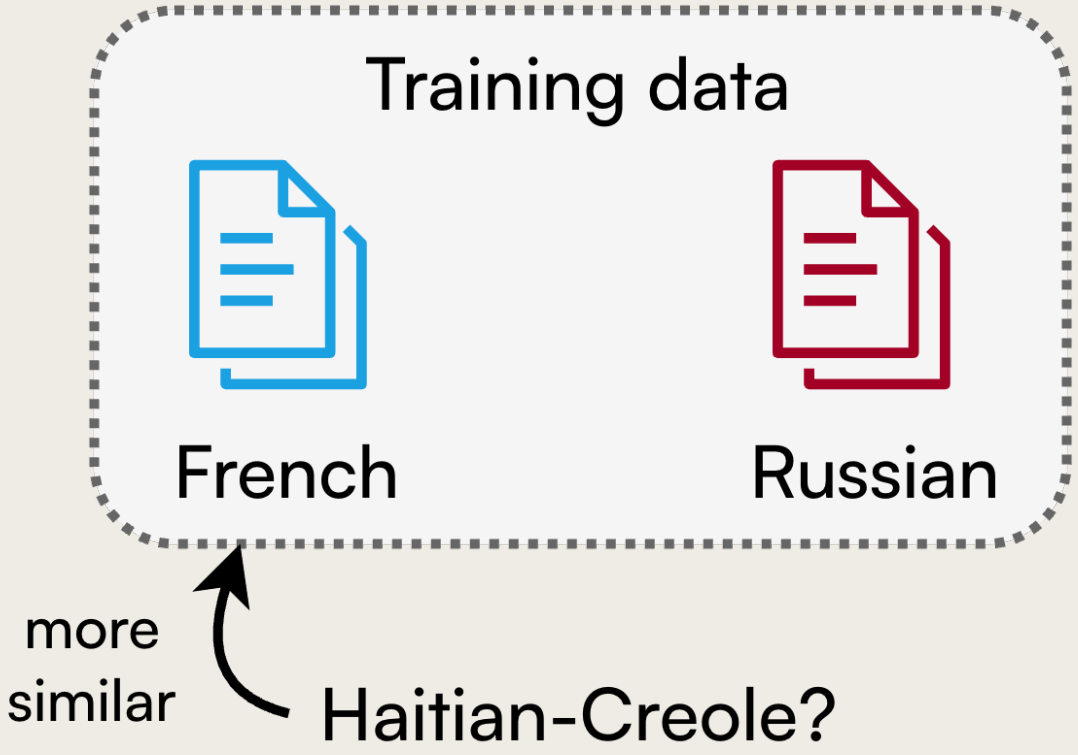
York Ng*, Aditya Khan*, James Lu*, Matteo Salloum, Michael Zhou,
Fiona Hoang, A. Seza Doğruöz, En-Shiun Annie Lee

Imagine training a multilingual LLM...

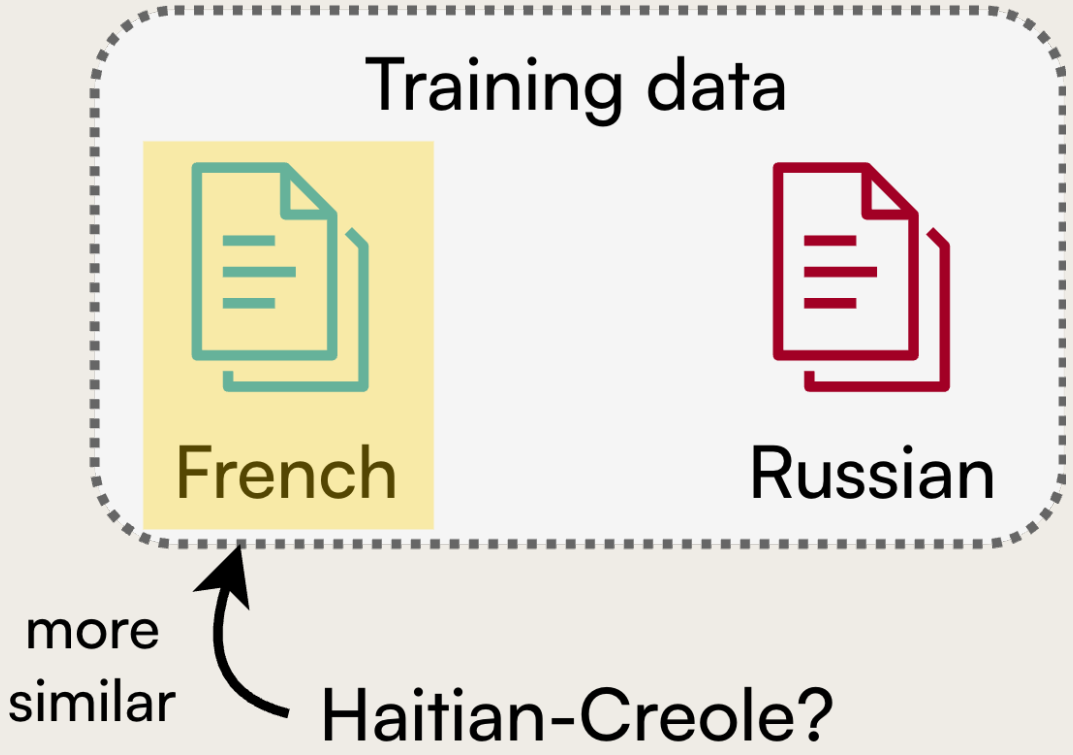


Haitian-Creole?

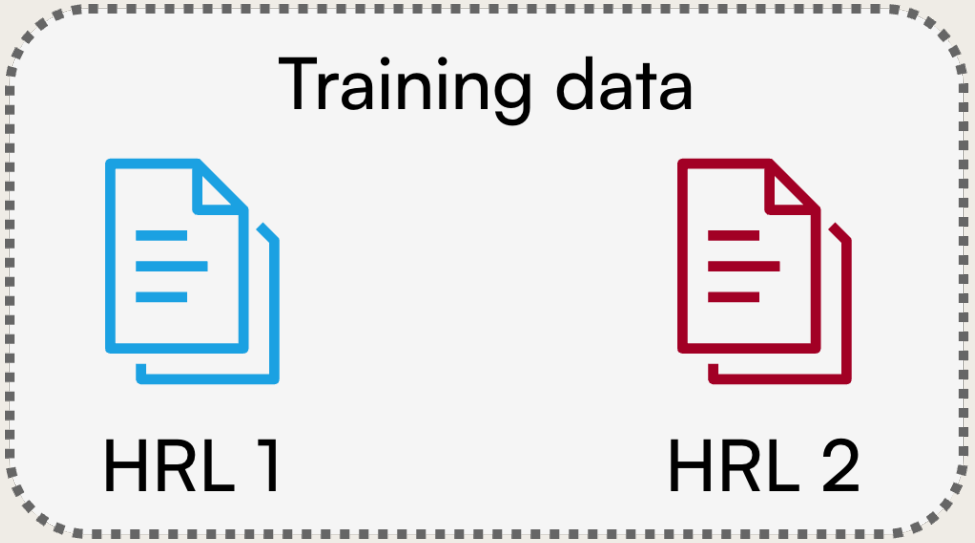
Imagine training a multilingual LLM...



Imagine training a multilingual LLM...



Imagine training a multilingual LLM...



Low-resource language X?

Imagine training a multilingual LLM...

We need a good measure
of language distances.

Language Distances: URIEL+ (Khan et al., 2025)

A **linguistic knowledge base** that calculates distances for 3 **modalities:**

Geographic	Genetic	Typological
Where languages are spoken	Shared ancestry on the language family tree	Difference in structure (e.g. grammar, phonology)

Supports 8k+ languages, each represented as a vector for each modality (constructed from various linguistics data sources).

Language Distances: URIEL+ (Khan et al., 2025)

Language distances are widely used in downstream NLP tasks, e.g.

- Cross-lingual transfer
  (Lin et al., 2019; Ansell et al., 2021; Accou & Poelman, 2026)
- Dependency parsing (Üstün et al., 2020; Glavaš & Vulić, 2021)
- Machine translation (Khiu et al., 2024; Hirak et al., 2026)
- Speech recognition (Adams et al., 2019; Sun et al., 2023)
- LLM regularization (Adilazuarda et al., 2024; Farashah et al., 2026)

being presented at EACL'26!

One Representations, but Three Modalities

URIEL+ represents languages as high-dimensional vectors for all modalities. Language distances = angular distance between vectors.

One Representations, but Three Modalities

URIEL+ represents languages as high-dimensional vectors for all modalities. Language distances = angular distance between vectors.

One Representations, but Three Modalities

URIEL+ represents languages as high-dimensional vectors for all modalities. Language distances = angular distance between vectors.

Geographic

Genetic

	hi	de	en
Indo-European?	1	1	1
West Germanic?	0	1	1
Anglo-Frisian?	0	0	1

en – de: 0.63 en – hi: 0.92

de – hi: 0.90

One Representations, but Three Modalities

URIEL+ represents languages as high-dimensional vectors for all modalities. Language distances = angular distance between vectors.

Geographic

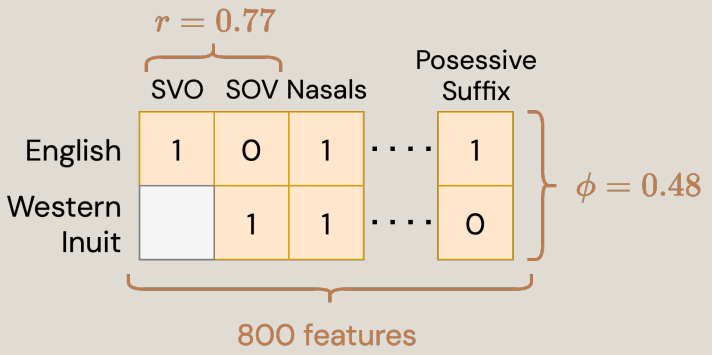
Genetic

	hi	de	en
Indo-European?	1	1	1
West Germanic?	0	1	1
Anglo-Frisian?	0	0	1

en – de: 0.63 en – hi: 0.92

de – hi: 0.90

Typological



One Representations, but Three Modalities

Ng et al. (2025) demonstrates that high-dimensional representations limit their effectiveness.

Generally, we claim:

Language representations should match the structure of the modality.

Modality-Matched Representations

(gives us modality-matched distances)

Modality	Problem	Representation	Distance
Geographic	Single point per language	Speaker distributions	Earth Mover's
Genetic	Flat one-hot encoding	Hyperbolic embeddings	Hyperbolic
Typological	High-dimensional, correlated features	Latent tree modelling	Angular

Speaker Distributions

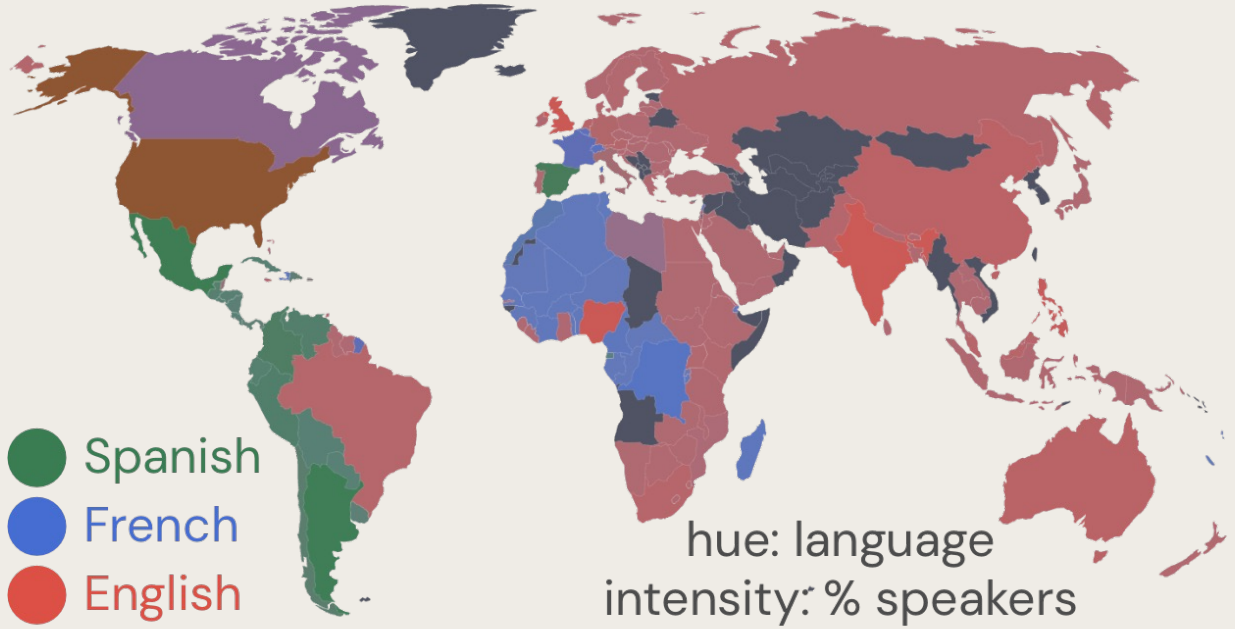
URIEL+:
Language as a single point

Speaker Distributions

URIEL+:
Language as a single point

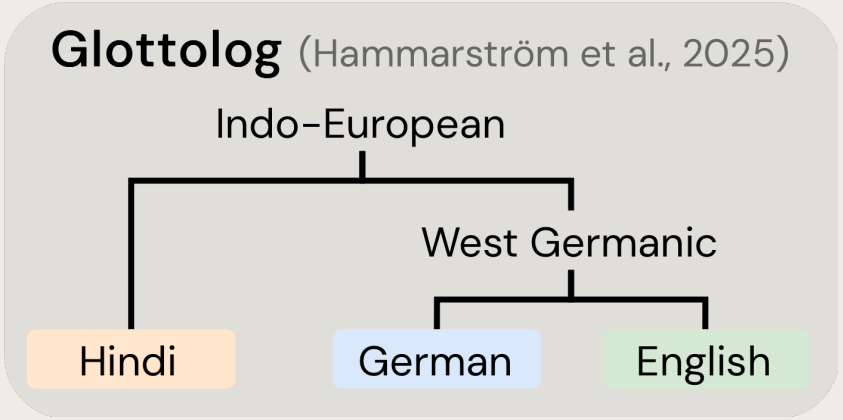
⬇

Language as a distribution over speaker locations, weighted by population



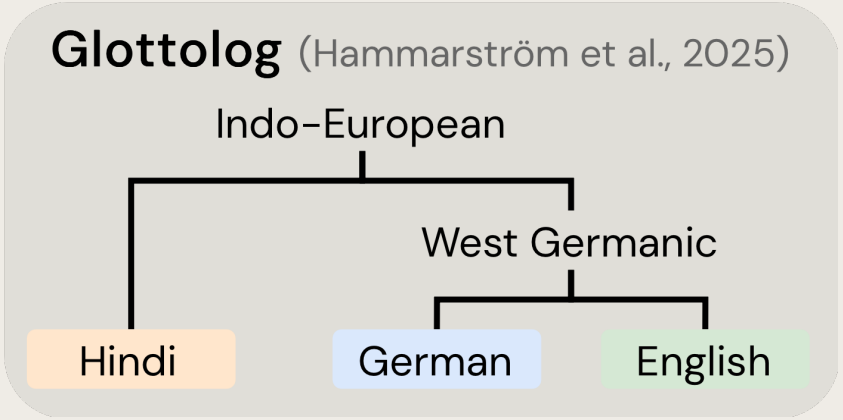
en – fr: 0.23 fr – es: 0.35 en – es: 0.39

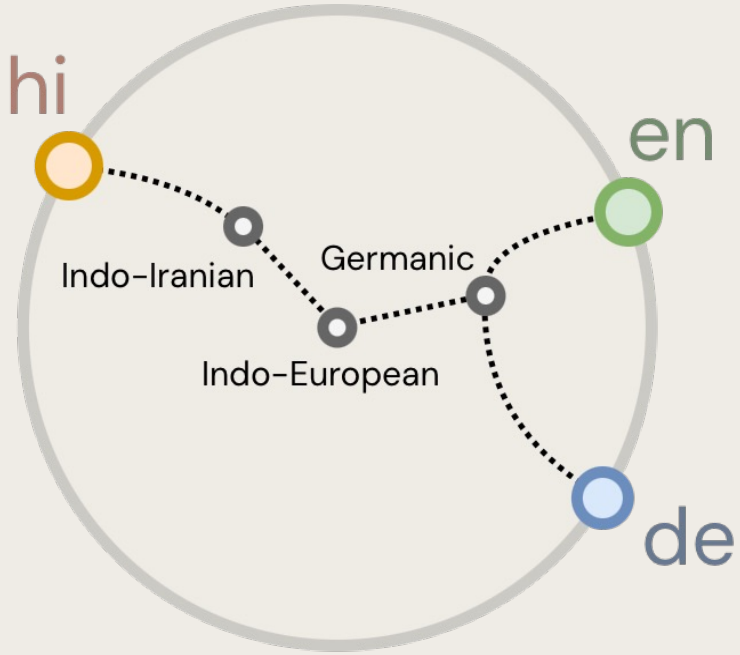
Hyperbolic Embeddings



URIEL+:
One-hot encoding of Glottolog

Hyperbolic Embeddings

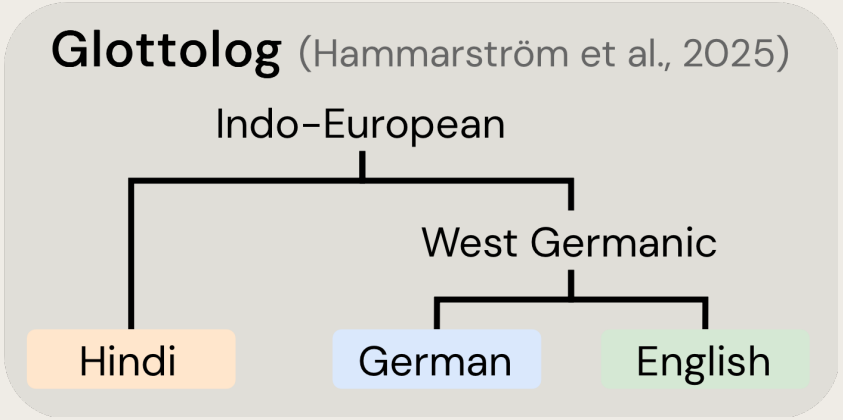




URIEL+:
One-hot encoding of Glottolog

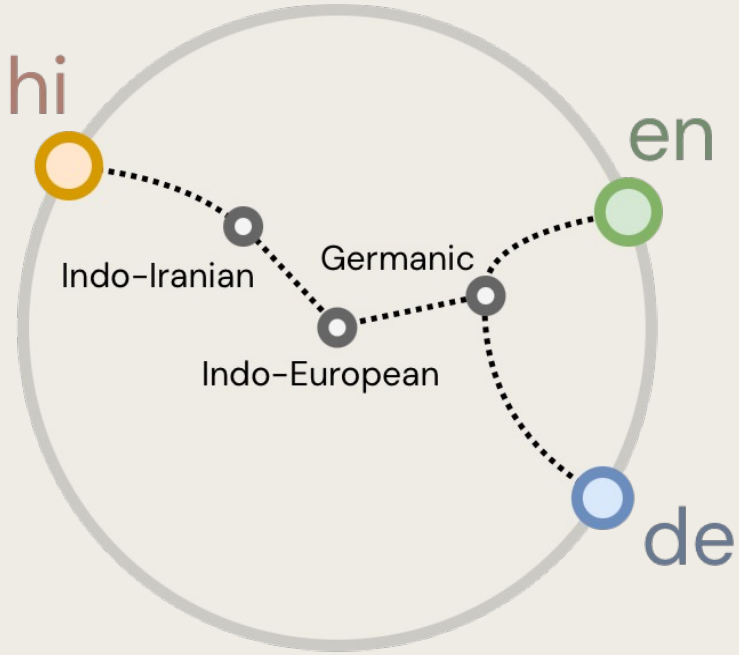
Embed Glottolog tree in hyperbolic space.

Hyperbolic Embeddings



URIEL+:
One-hot encoding of Glottolog

Embed Glottolog tree in hyperbolic space.



Reconstruction loss.
Euclidean: 0.72 MAP
Hyperboloid (d=10): 0.96 MAP

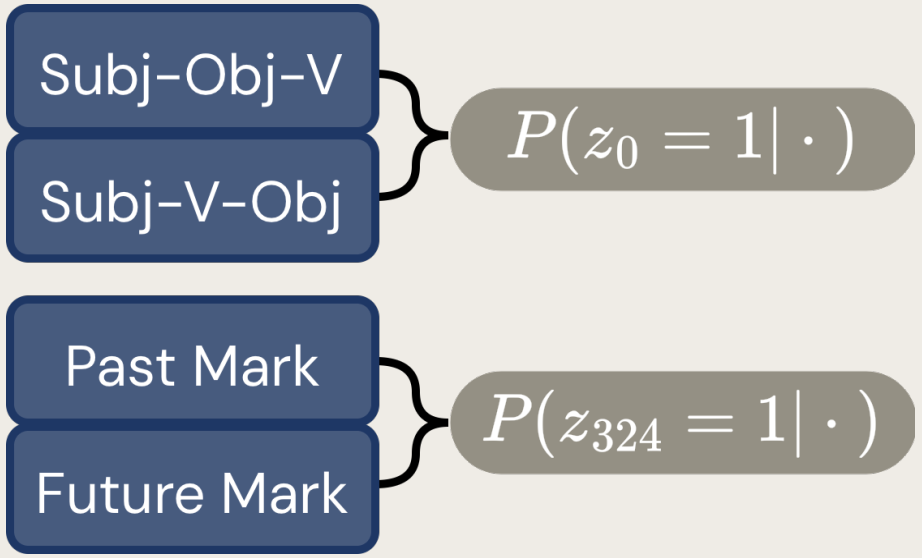
Latent Tree Modelling

URIEL+:
High-dimensional, correlated features

Latent Tree Modelling

URIEL+:
High-dimensional, correlated features

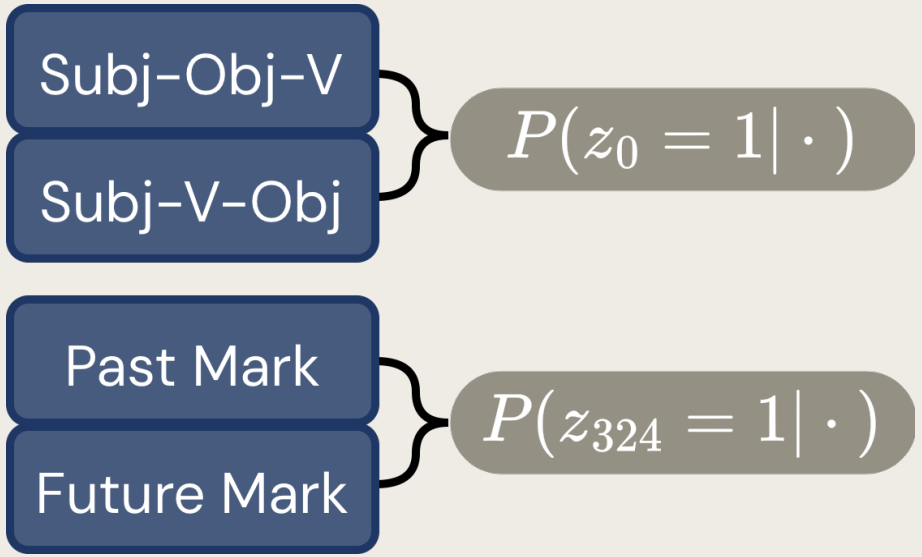
Cluster correlated features into latent binary variables ("islands")



Latent Tree Modelling

URIEL+:
High-dimensional, correlated features

Cluster correlated features into latent binary variables ("islands")



Intra-cluster correlation = 0.62

Impact on Cross-Lingual Transfer

Task: transfer language selection.

Model: LangRank (Lin et al., 2019)

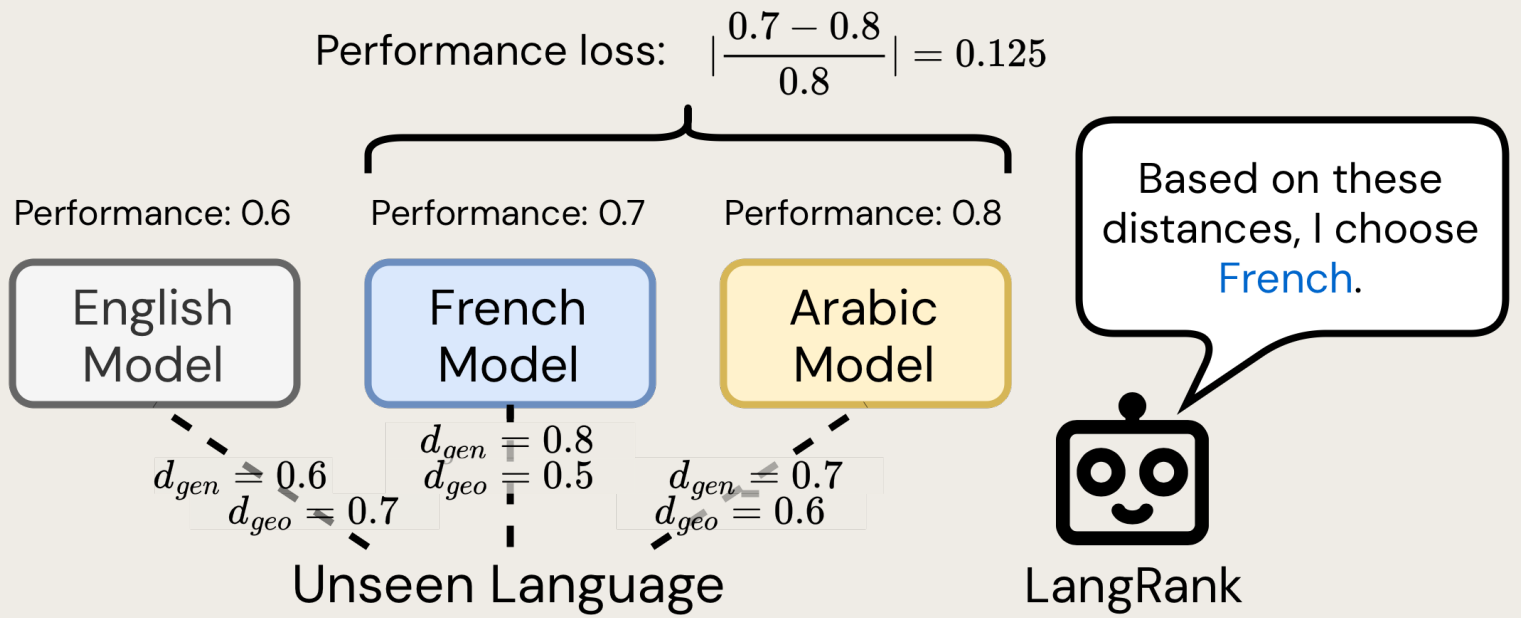
Metric: performance loss.

Impact on Cross-Lingual Transfer

Task: transfer language selection.

Model: LangRank (Lin et al., 2019)

Metric: performance loss.



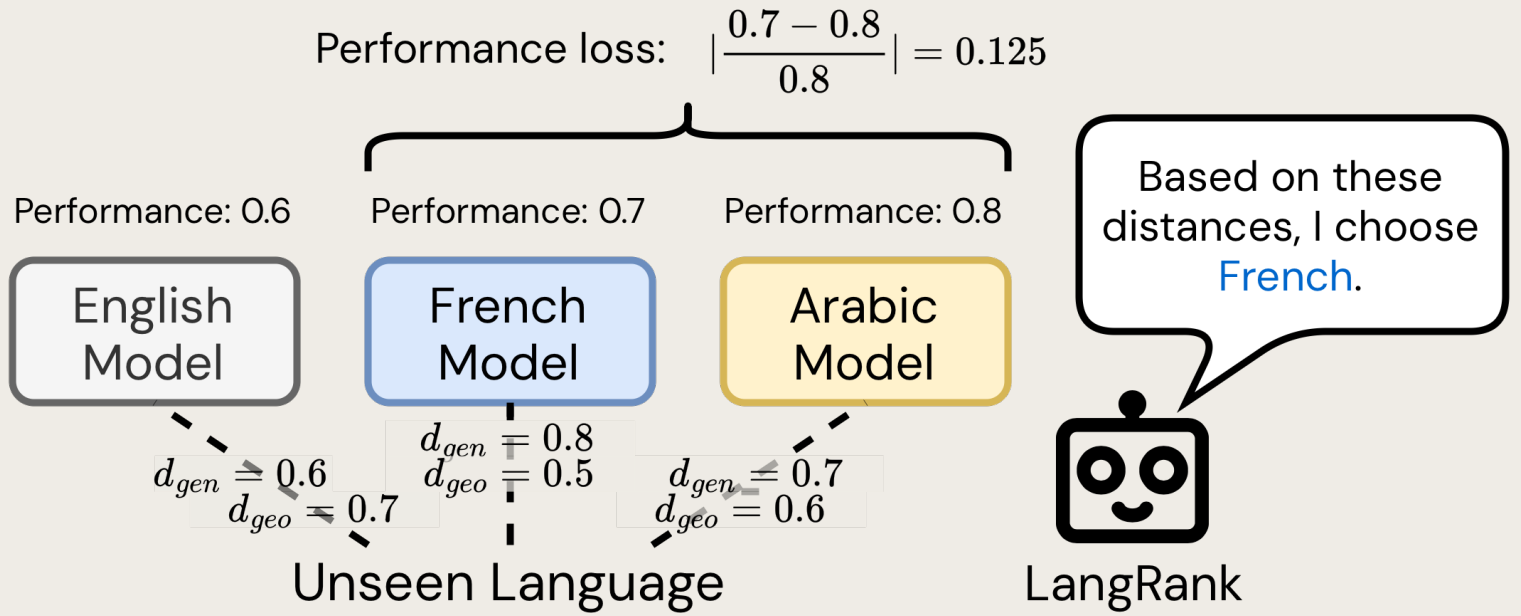
Impact on Cross-Lingual Transfer

Task: transfer language selection.

Model: LangRank (Lin et al., 2019)

Metric: performance loss.

Ablate on URIEL+ vs.
our representations
→ regression analysis.



Impact on Cross-Lingual Transfer

Task Type	Dataset	Related Work	Model	Metric	Target	Source
Machine Trans.	TED	Lin et al. (2019)	RNN+Attn	BLEU	54	54
Dep. Parsing	UD v2.2	Lin et al. (2019)	Biaffine	Accuracy	30	30
	UD v2.14	Blaschke et al. (2025)	UDPipe 2	LAS	152	70
POS Tagging	UD v2.2	Lin et al. (2019)	BiLSTM	Accuracy	60	26
	UD v2.14	Blaschke et al. (2025)	UDPipe 2	UPOS	152	70
Entity Linking	Wikipedia	Lin et al. (2019)	BiLSTM	Accuracy	54	9
Topic Class.	Taxi1500	–	mBERT	Macro F1	799	33
	SIB200	Blaschke et al. (2025)	XLM-R	Macro F1	197	160
NLI	XNLI	Philippy et al. (2023)	mBERT	Accuracy	15	15

Downstream tasks used for transfer language selection.

Impact on Cross-Lingual Transfer

Modality	Representation	DEP	EL	MT	POS
Baseline:		11.4 ± 2.9	30.0 ± 6.2	12.5 ± 1.8	27.9 ± 4.4
Typ	Laplacian	+0.8 ± 1.0	−3.8 ± 2.8	+0.7 ± 0.9	−2.1 ± 1.9
Typ	Islands	+0.5 ± 1.0	−1.2 ± 2.8	−1.0 ± 0.9	−0.4 ± 1.9
Geo	Speaker	+0.6 ± 0.7	**−7.4 ± 2.0**	−1.0 ± 0.6	−0.3 ± 1.3
Gen	Hyperbolic	−0.9 ± 0.7	+3.6 ± 2.0	**−4.5 ± 0.6**	−1.0 ± 1.3

Modality	Representation	Taxi1500	SIB200	XNLI	UD2.14 POS	UD2.14 DEP
Baseline:		38.1 ± 0.5	16.9 ± 1.1	6.2 ± 1.2	27.4 ± 1.5	35.6 ± 1.9
Typ	Laplacian	+0.4 ± 0.3	−0.2 ± 0.5	+0.4 ± 0.6	**+1.8 ± 0.8**	+1.5 ± 0.9
Typ	Islands	**−0.9 ± 0.3**	**−1.4 ± 0.5**	**−2.4 ± 0.6**	−0.6 ± 0.8	**−1.8 ± 0.9**
Geo	Speaker	**−2.1 ± 0.2**	−0.6 ± 0.3	+0.1 ± 0.4	**−1.6 ± 0.6**	+0.7 ± 0.6
Gen	Hyperbolic	**+2.7 ± 0.2**	**+1.0 ± 0.3**	−0.1 ± 0.4	**−2.6 ± 0.6**	**−3.9 ± 0.6**

Regression coefficients (p.p.). Lower is better. Bold indicates p<0.05.
Baseline: loss when using URIEL+ distances.
Subsequent: change in loss when switching to new distances.

Takeaways

Match structure to modality.

Structure-aware
language representations
consistently outperform URIEL+ in
cross-lingual transfer.

Takeaways

Match structure to modality.

Structure-aware language representations consistently outperform URIEL+ in cross-lingual transfer.

No one-size-fits-all solution.

Each modality's utility for cross-lingual transfer is task-dependent. Language representations not a drop-in replacement.

Thank you!

Paper

Code

{yorkng, adityakhan}@cs.toronto.edu