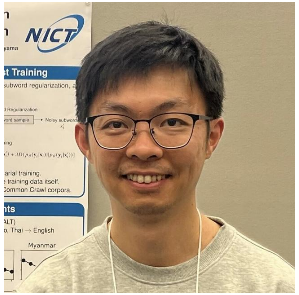
Connecting Ideas in ‘Lower-Resource’ Scenarios: NLP for National Varieties, Creoles and Other Low-resource Scenarios | COLING 2025 Tutorial | 20th January 2025 | Abu Dhabi, United Arab Emirates

Connecting Ideas in ‘Lower-Resource’ Scenarios: NLP for National Varieties, Creoles and Other Low-resource Scenarios

Aditya Joshi | Diptesh Kanojia | Heather Lent | Hour Kaing | Haiyue Song




UNSW SYDNEY

People-Centred AI
UNIVERSITY OF SURREY

AALBORG UNIVERSITY

NICT

NICT

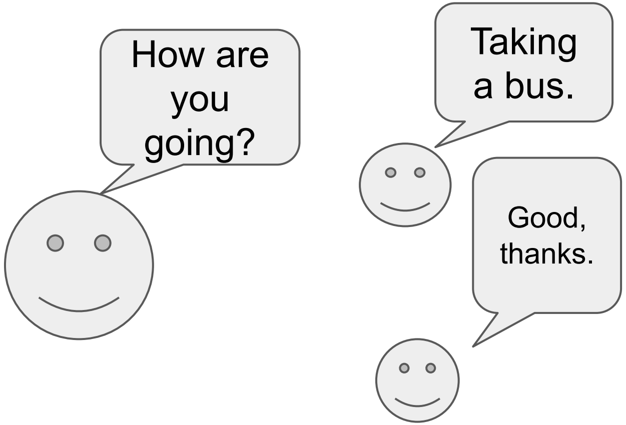
COLING 2025 • Abu Dhabi
The 31st International Conference on Computational Linguistics

Tutorial | 20th January 2025 | Abu Dhabi, United Arab Emirates

https://github.com/surrey-nlp/COLING-Tutorial-LowResScene-2025

Ice-breaker

tomay-to or tomah-to?



What are ‘these’ called in your country?
apartment/unit/block/ flat/condo…..

Has something similar happened to you?

Tell the person(s) sitting next to you…

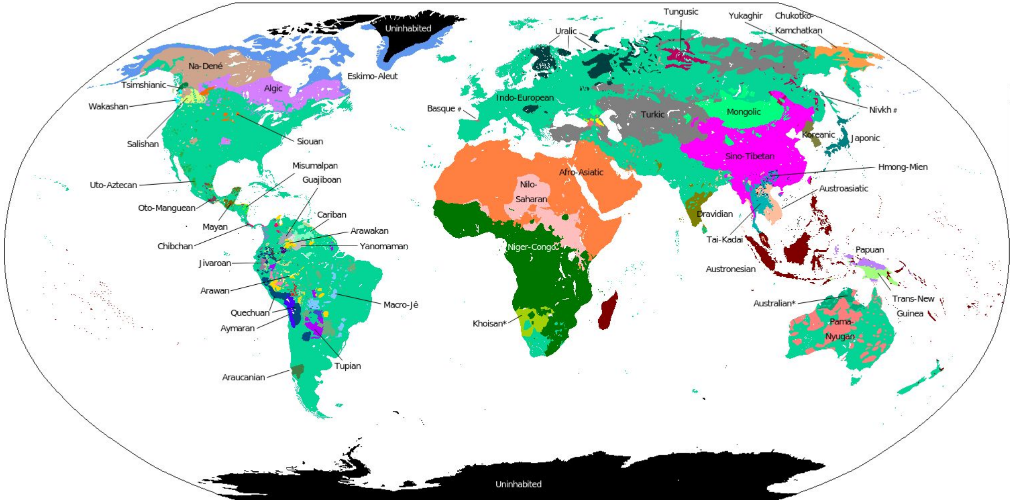
What was the first language you learned?

In which countries is your first language spoken?

What language(s) do you understand to some or little extent, although you never formally learned them?

“Languages”

Language	Number of speakers, both native and second-language (mln)	Number of native speakers (mln)	Percentage of the share of web content featured in this language (%)
English	1520	380	52.1
Chinese (Mandarin)	1140	941	1.3
Hindi	609	345	less than 0.5
Spanish	560	486	5.5
Arabic	422	313	0.6
French	321	189	4.3
Bengali	273	230	less than 0.5
Portuguese	264	236	3.1
Russian	255	148	4.5
Urdu	232	70	less than 0.5



https://www.icls.edu/blog/most-spoken-languages-in-the-world

Source: Wikimedia Commons

Module 1: Introduction

This module discusses recent developments in NLP, and establishes the motivation and structure for this tutorial. (30 minutes)

– Transformer & Language Models

– Dialects, Creoles, and other lower-resourced languages

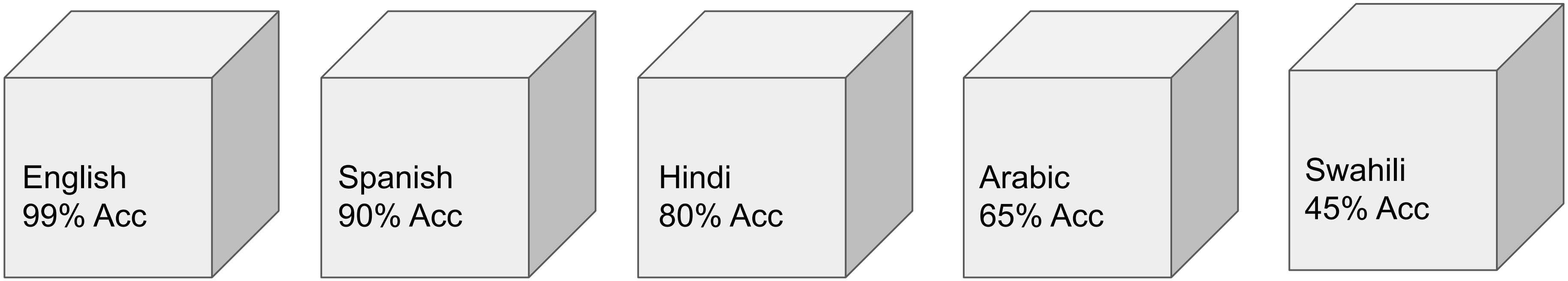
– Motivation & Computational tasks

– Objectives & Structure of the tutorial.

Connecting Ideas in ‘Lower-Resource’ Scenarios: NLP for National Varieties, Creoles and Other Low-resource Scenarios

Connecting Ideas in ‘Lower-Resource’ Scenarios: NLP for **National Varieties, Creoles and Other Low-resource Scenarios**

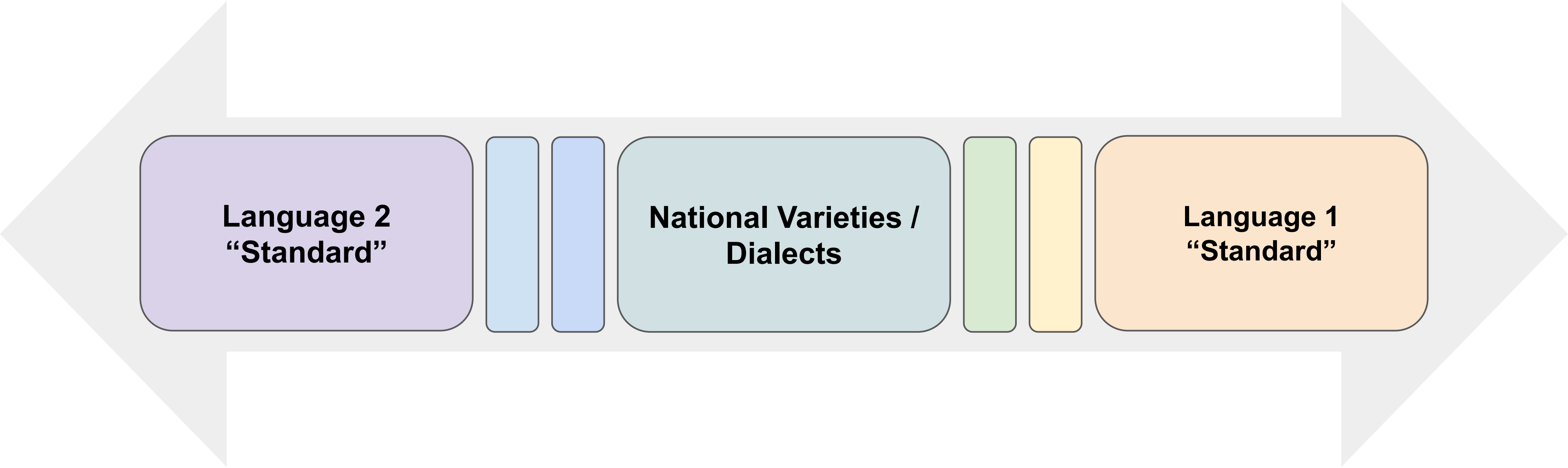
“Language” in NLP (Historically)



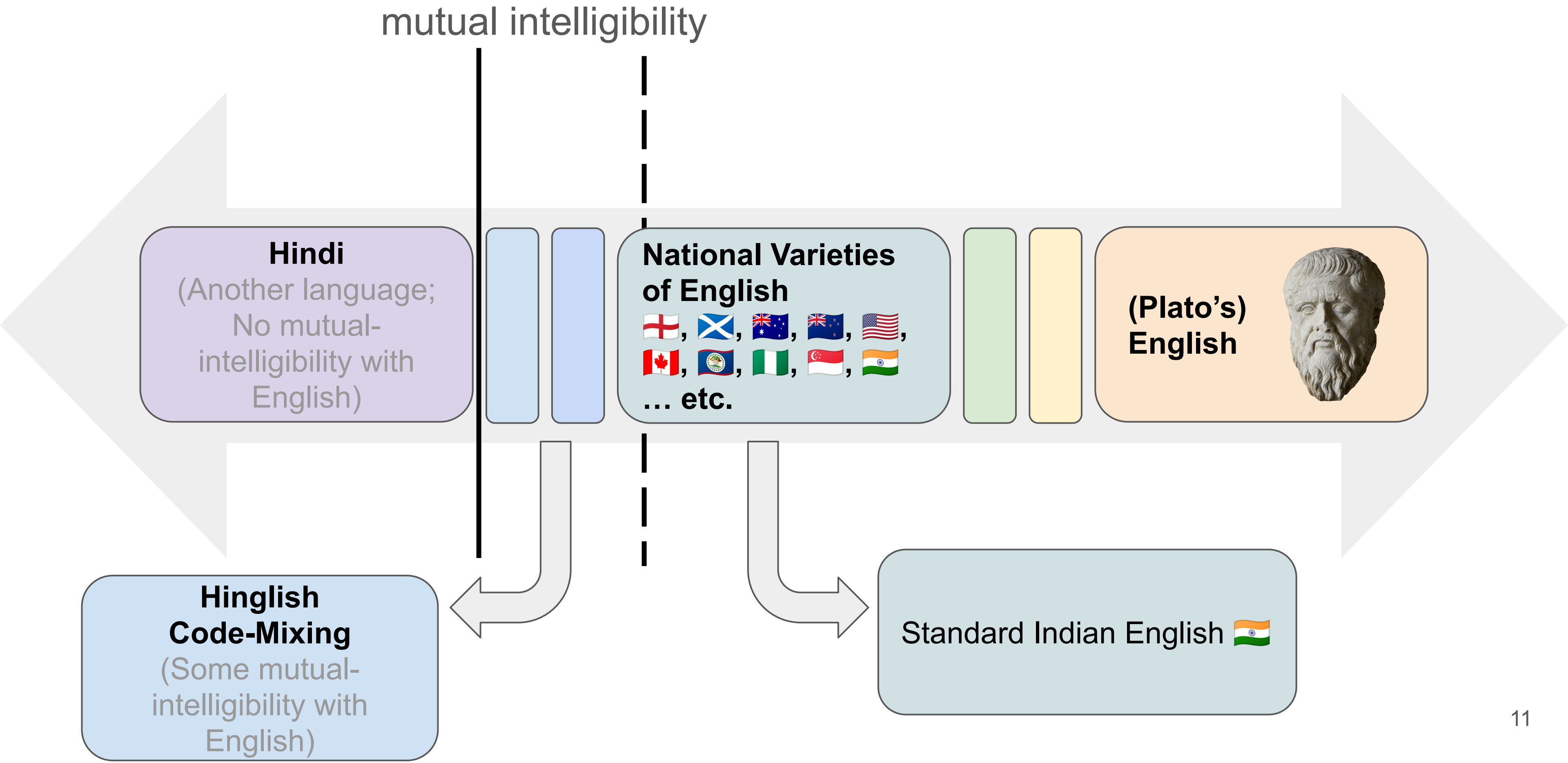
But this model of language doesn’t fit all varieties!

Singaporean English vs **American English** ?
English-Hindi code-mixing ?
Nigerian Pidgin, with vocab from English, Portuguese, & Yoruba ?

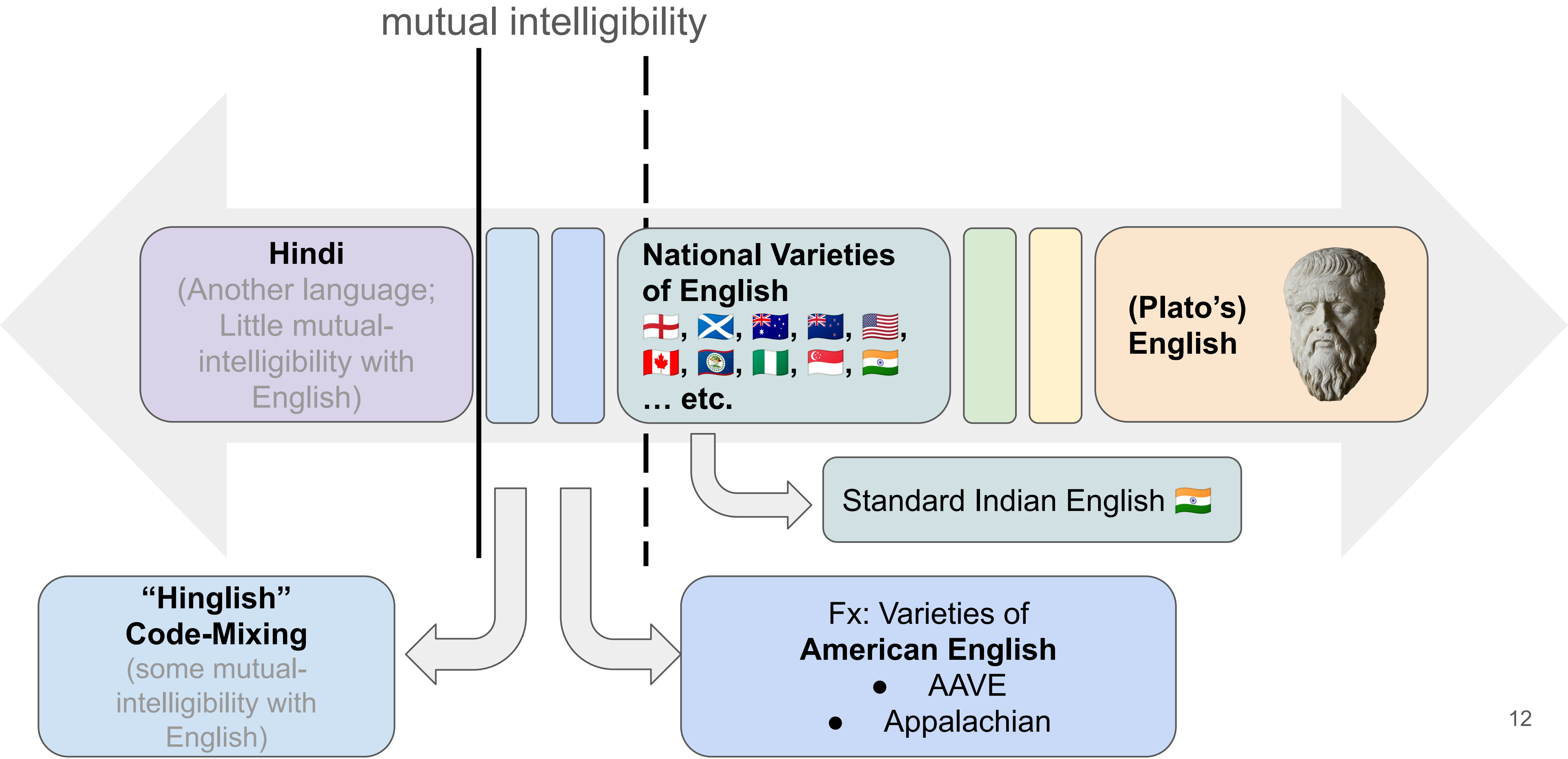
Language as a Continuum



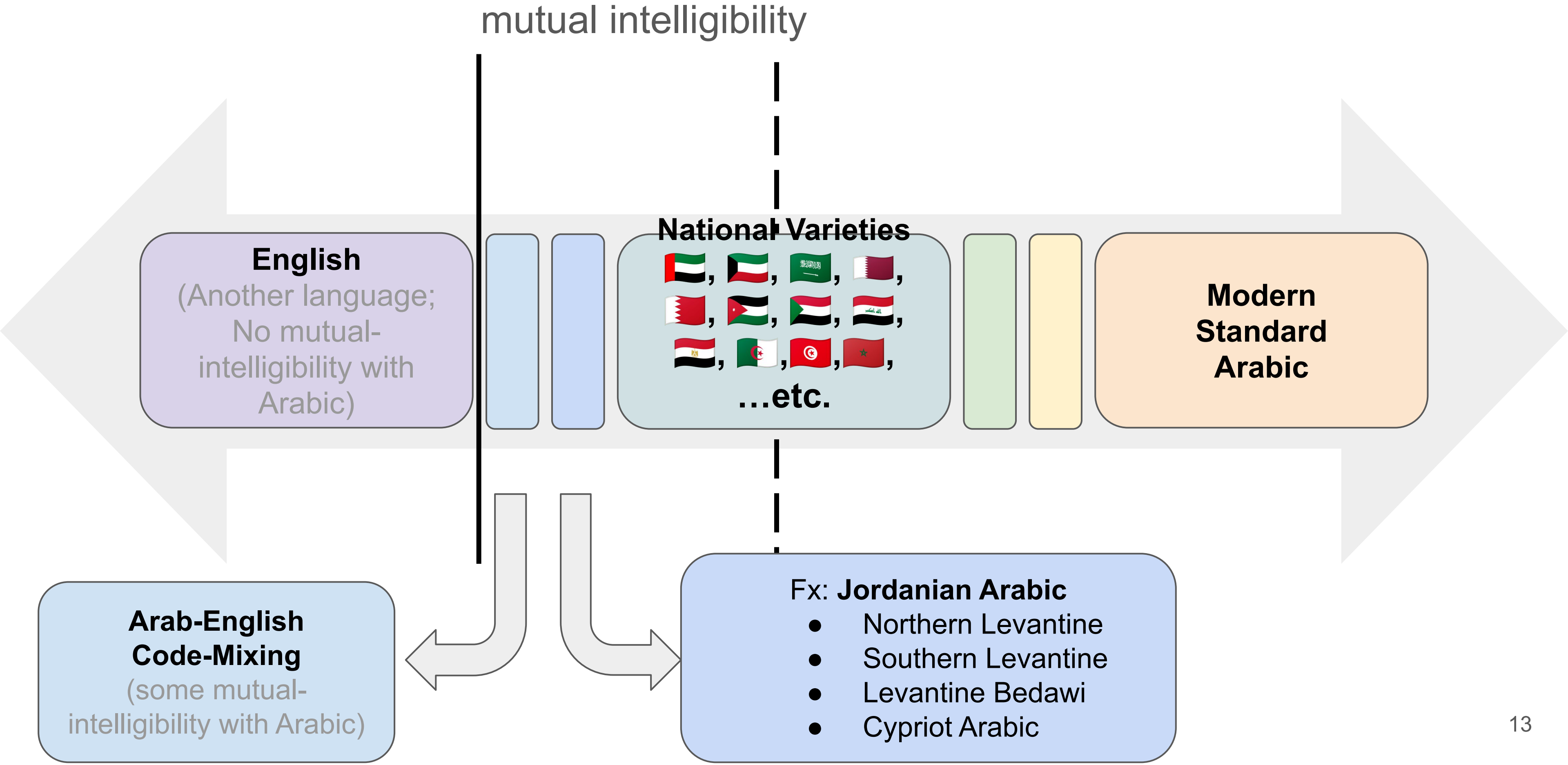
Language as a Continuum



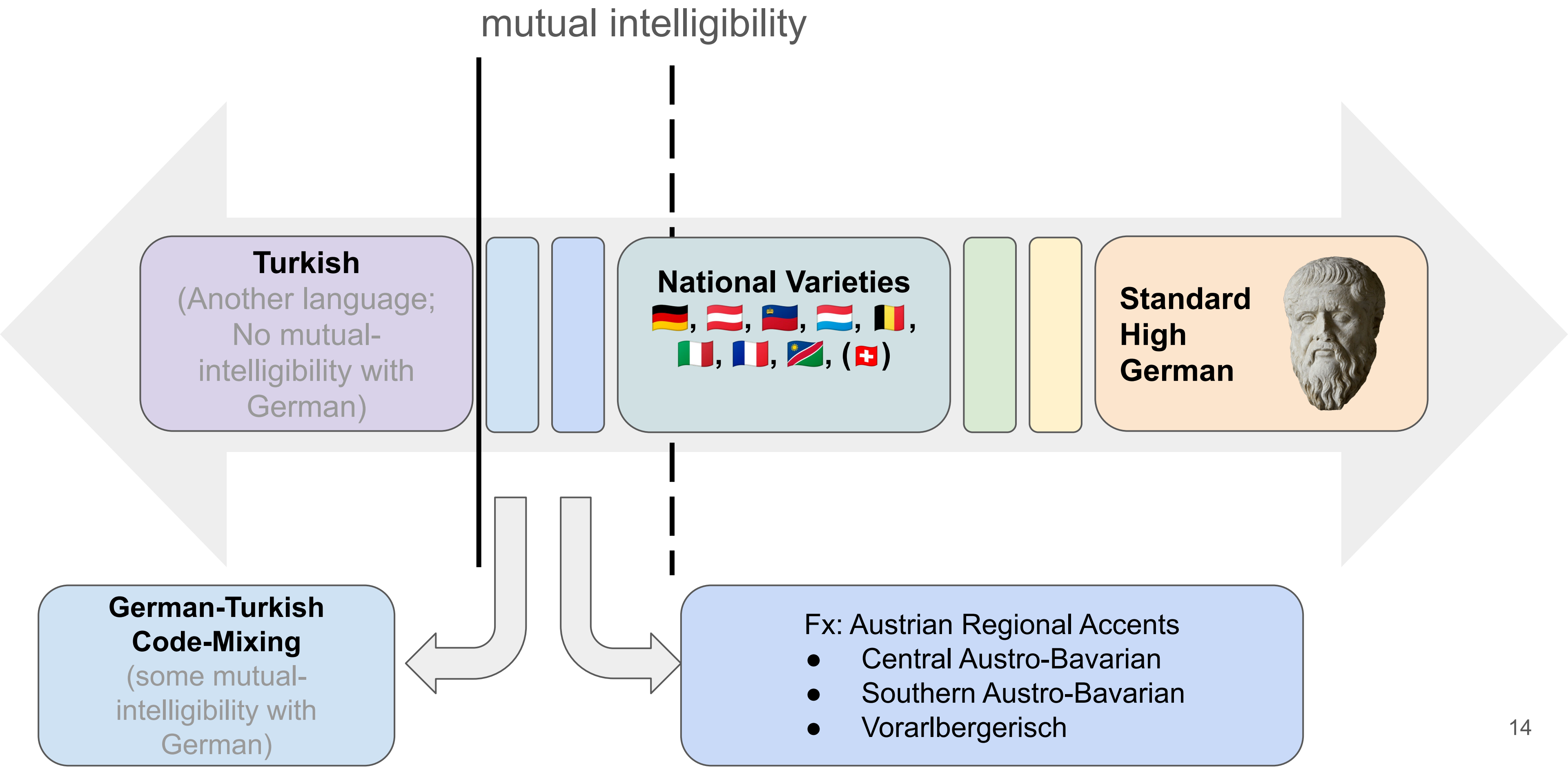
Language as a Continuum



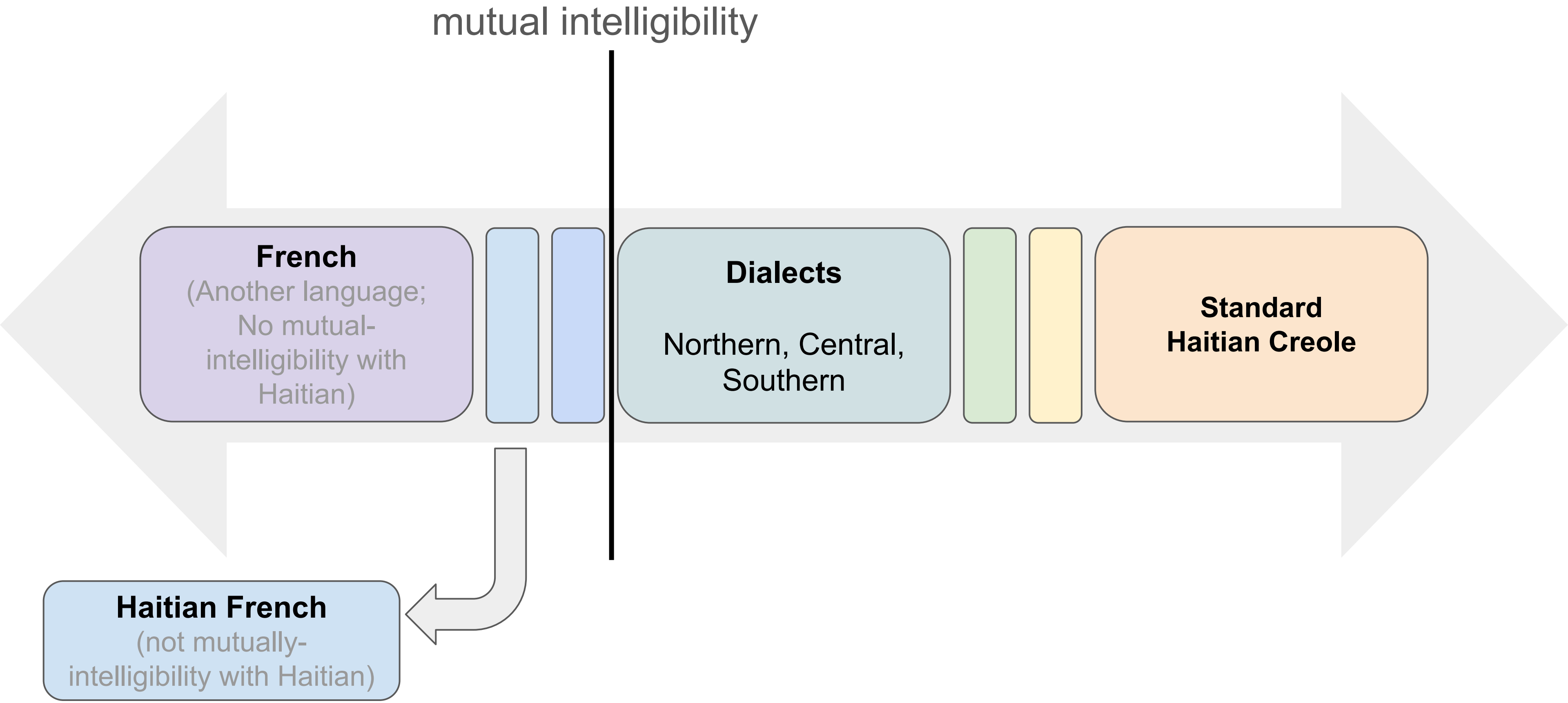
Language as a Continuum



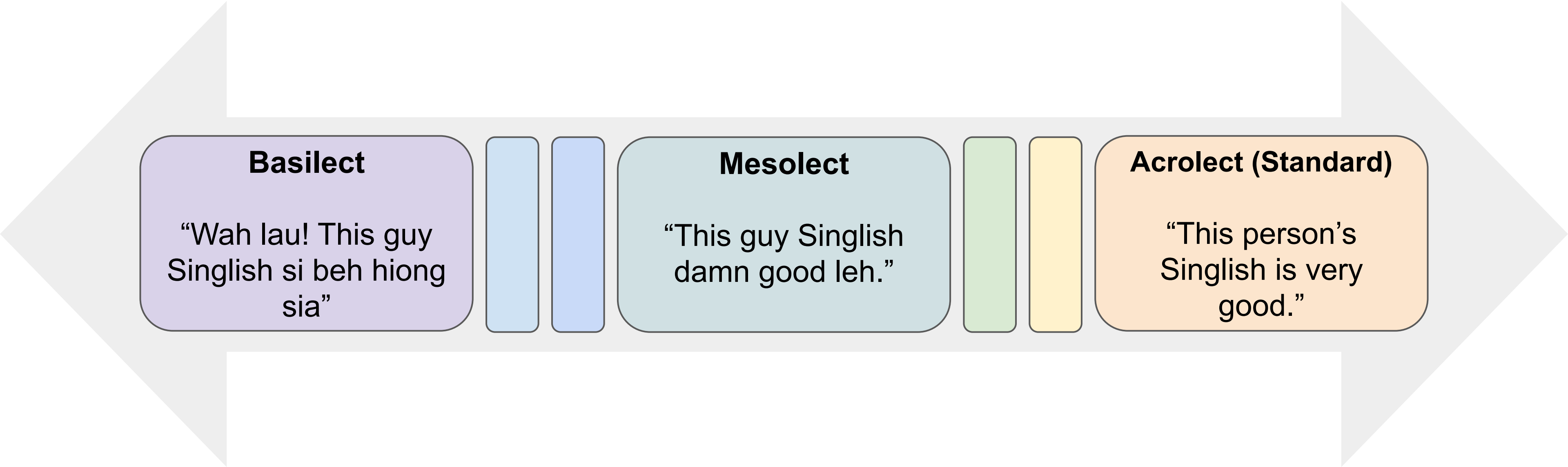
Language as a Continuum



Language as a Continuum



Creole Continuum (Singlish)



Stigmatisation of languages & language varieties

Languages, languages everywhere

National varieties: language varieties characterised by certain national backgrounds

Languages, languages everywhere

National varieties: language varieties characterised by certain national backgrounds

Dialects: a variety of a language that is a characteristic of a particular group of the language’s speakers.

Languages, languages everywhere

National varieties: language varieties characterised by certain national backgrounds

Dialects: a variety of a language that is a characteristic of a particular group of the language’s speakers.

Creoles: languages that develop from linguistic contact between different languages, resulting into a new full-fledged language

Languages, languages everywhere

National varieties: language varieties characterised by certain national backgrounds

Dialects: a variety of a language that is a characteristic of a particular group of the language’s speakers.

Creoles: languages that develop from linguistic contact between different languages, resulting into a new full-fledged language

Low-resource languages: Languages with insufficient amount of datasets or techniques

“Lower-resource scenarios”

National varieties: language varieties characterised by certain national backgrounds

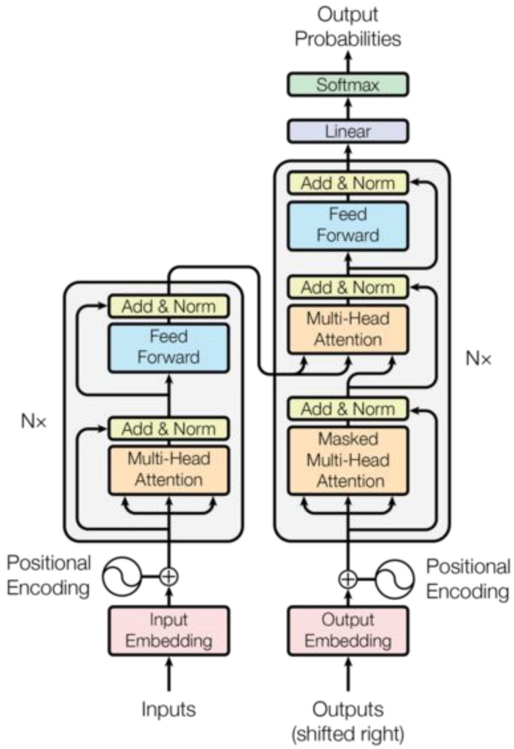
Dialects: a variety of a language that is a characteristic of a particular group of the language’s speakers.

Creoles: languages that develop from linguistic contact between different languages, resulting into a new full-fledged language

Low-resource languages: Languages with insufficient amount of datasets or techniques

Connecting Ideas in ‘Lower-Resource’ Scenarios: NLP for National Varieties, Creoles and Other Low-resource Scenarios

Transformer



Seq2Seq architecture w/ self- and multi-headed attention

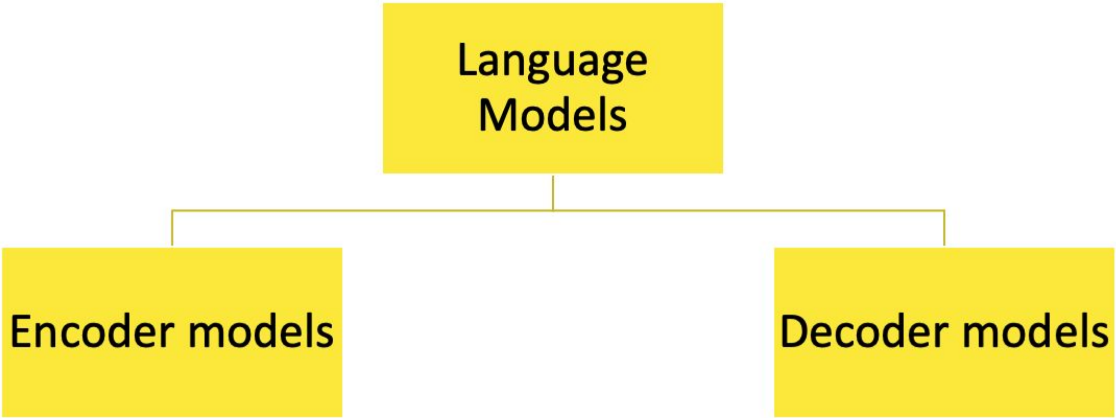
Derivatives: Encoder & decoder models

Large language models

Versatile models that can be customized to varying degrees

Vaswani, A. et al. "Attention is all you need." *Advances in Neural Information Processing Systems* (2017).

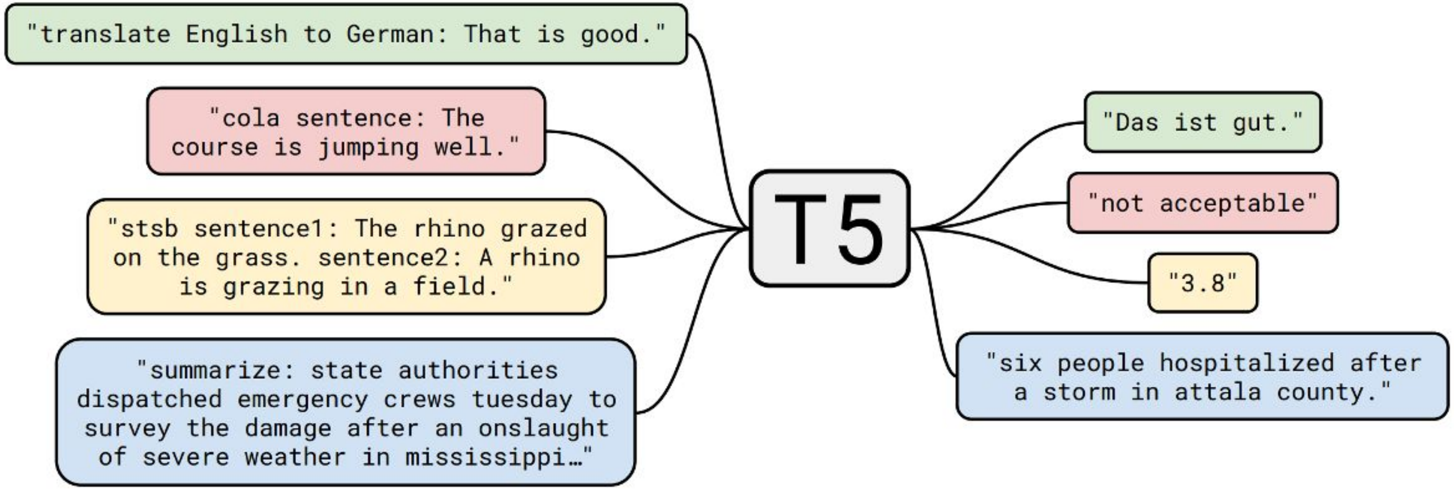
Encoder & Decoder Models -> Large Language Models



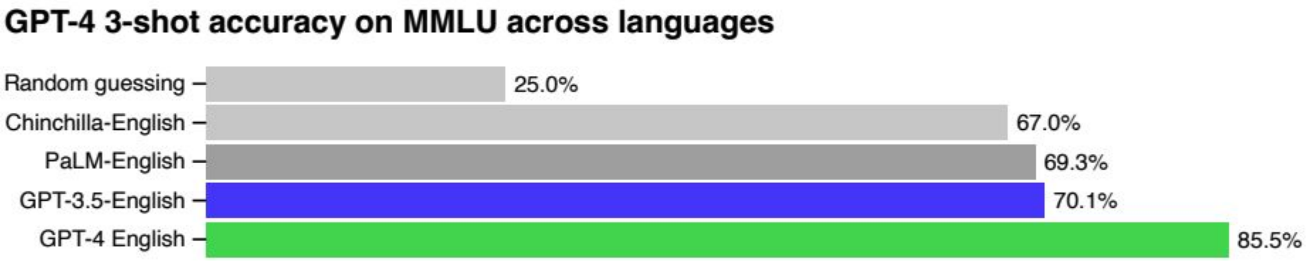
Use the encoder of the Transformer
Current word is estimated from neighbouring words.

Use the decoder of the Transformer
Current word is estimated from previous words.

LLMs are “versatile”



LLMs perform “phenomenally” well.. for English



Gemini: A Family of Highly Capable Multimodal Models

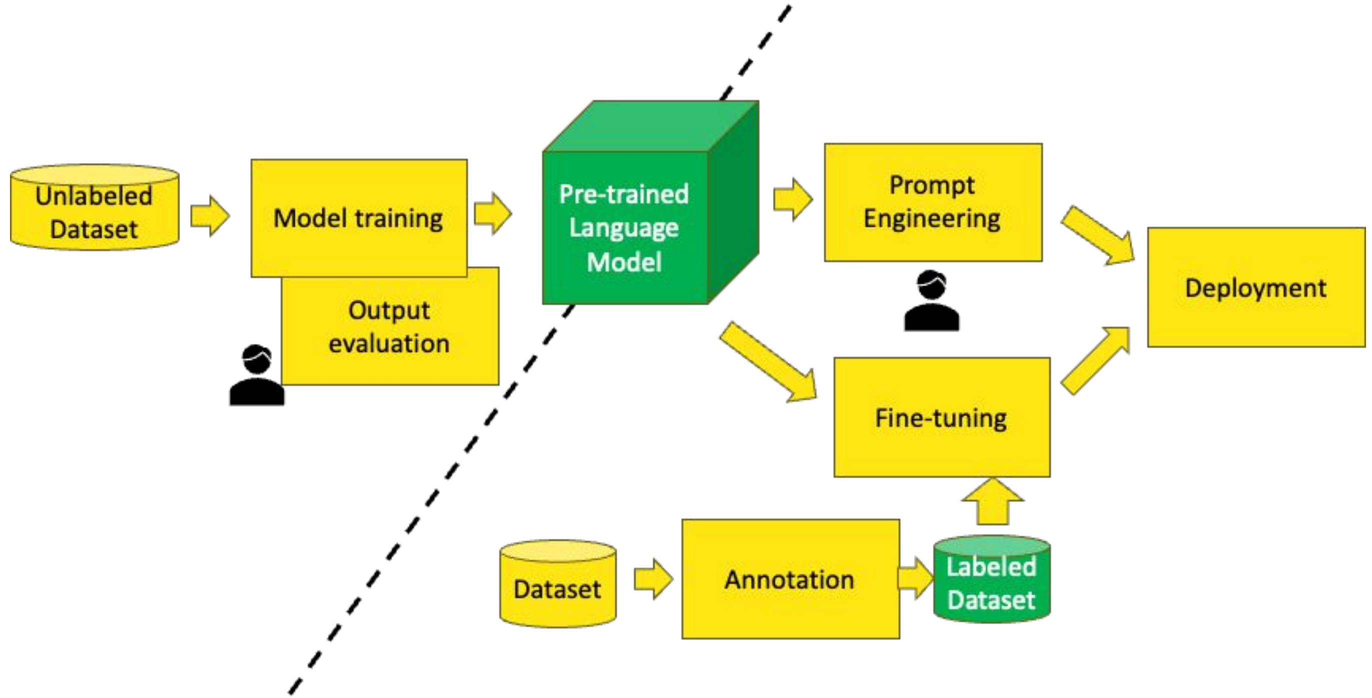
	Gemini Ultra	Gemini Pro	GPT-4	GPT-3.5	PaLM 2-L	Claude 2	Inflect-ion-2	Grok 1	LLAMA-2
MMLU Multiple-choice questions in 57 subjects (professional & academic) (Hendrycks et al., 2021a)	**90.04%** CoT@32*	79.13% CoT@8*	87.29% CoT@32 (via API**)	70% 5-shot	78.4% 5-shot	78.5% 5-shot CoT	79.6% 5-shot	73.0% 5-shot	68.0%***
	83.7% 5-shot	71.8% 5-shot	86.4% 5-shot (reported)						

https://arxiv.org/pdf/2312.11805

LLMs can be adapted to specialised tasks and domains

https://aclanthology.org/2024.findings-acl.445.pdf

Modern NLP workflow



Aren’t SoTA results already fairly high?

Aren’t SoTA results already fairly high?

NLP techniques are NOT equally effective for majority of languages.

Language Resource Distribution

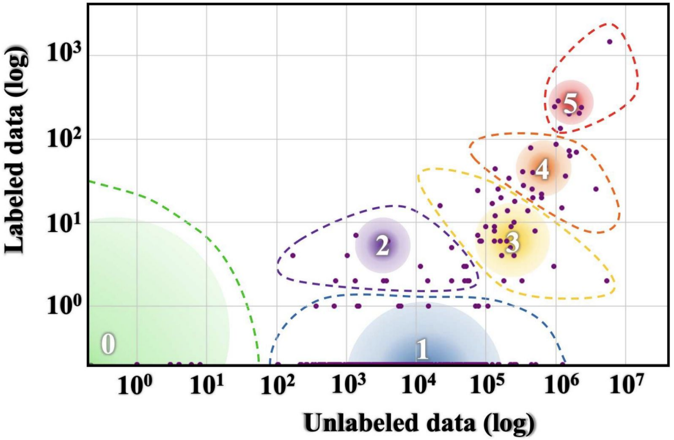


Figure 2: Language Resource Distribution: The size of the gradient circle represents the number of languages in the class. The color spectrum VIBGYOR, represents the total speaker population size from low to high. Bounding curves used to demonstrate covered points by that language class.

Joshi, Pratik, et al. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.

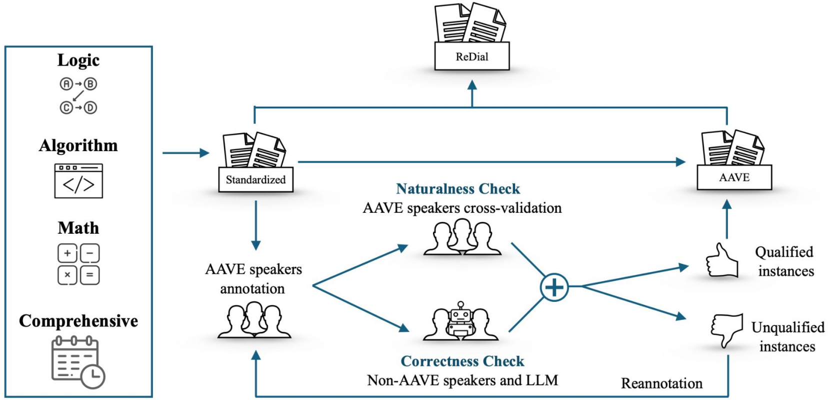
NLP for low-resource languages

Method	Requirements	Outcome	For low-resource languages	For low-resource domains
Data Augmentation (§ 4.1)	labeled data, heuristics*	additional labeled data	✓	✓
Distant Supervision (§ 4.2)	unlabeled data, heuristics*	additional labeled data	✓	✓
Cross-lingual projections (§ 4.3)	unlabeled data, high-resource labeled data, cross-lingual alignment	additional labeled data	✓	✗
Embeddings & Pre-trained LMs (§ 5.1)	unlabeled data	better language representation	✓	✓
LM domain adaptation (§ 5.2)	existing LM, unlabeled domain data	domain-specific language representation	✗	✓
Multilingual LMs (§ 5.3)	multilingual unlabeled data	multilingual feature representation	✓	✗
Adversarial Discriminator (§ 6)	additional datasets	independent representations	✓	✓
Meta-Learning (§ 6)	multiple auxiliary tasks	better target task performance	✓	✓

Table 1: Overview of low-resource methods surveyed in this paper. * Heuristics are typically gathered manually.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of the 2021 NAACL*.

What about varieties of languages? African-American English?



		Algorithm	Math	Logic	Comprehensive	Average
Zero-shot	Original	0.602	0.733	0.578	0.191	0.546
	AAVE	**0.517**$_{\Delta=0.085}$	**0.665**$_{\Delta=0.068}$	**0.522**$_{\Delta=0.056}$	**0.101**$_{\Delta=0.090}$	**0.473**$_{\Delta=0.073}$
CoT	Original	0.597	0.811	0.580	0.240	0.574
	AAVE	**0.495**$_{\Delta=0.102}$	**0.742**$_{\Delta=0.068}$	**0.530**$_{\Delta=0.050}$	**0.177**$_{\Delta=0.063}$	**0.504**$_{\Delta=0.070}$

Lin, Fangru, et al. "One Language, Many Gaps: Evaluating Dialect Fairness and Robustness of Large Language Models in Reasoning Tasks." *arXiv preprint arXiv:2410.11005* (2024).

.. and Indian English?

Model	Subset	TWP						TWS					
		Similarity			Accuracy			Similarity			Accuracy		
		PT	FT	Δ	PT	FT	Δ	PT	FT	Δ	PT	FT	Δ
GPT-4	en-US	77.4	–	–	67.8	–	–	85.7	–	–	78.8	–	–
	en-IN	63.0	–	–	45.6	–	–	79.0	–	–	72.5	–	–
	en-MV	75.6	–	–	60.0	–	–	83.6	–	–	74.4	–	–
	en-TR	62.8	–	–	45.8	–	–	83.4	–	–	77.1	–	–
	δ	-14.4	–	–	-22.0	–	–	-6.7	–	–	-6.3	–	–
GPT-3.5	en-US	66.3	72.2	5.9	52.7	59.1	6.4	66.4	80.8	14.4	50.8	71.3	20.5
	en-IN	53.2	59.1	5.9	34.4	40.0	5.6	61.9	70.7	8.8	47.5	60.6	13.1
	en-MV	57.6	71.3	13.7	40.0	54.4	14.4	52.4	71.5	19.1	31.6	57.6	26.0
	en-TR	59.4	61.0	1.6	39.7	41.2	1.5	70.7	73.0	2.3	57.3	60.3	3.0
	δ	-13.1	-13.1	–	-18.3	-19.1	–	-4.5	-10.1	–	-21.0	-16.2	–
LLAMA-3	en-US	70.8	78.0	7.2	60.5	65.3	4.8	78.0	81.8	3.8	67.5	74.6	7.1
	en-IN	59.8	66.3	6.5	43.8	54.4	10.6	68.8	80.8	12.0	56.9	74.4	17.5
	en-MV	68.6	73.8	5.2	54.0	61.6	7.6	72.3	77.6	5.3	58.8	67.2	8.4
	en-TR	60.7	57.5	-3.2	45.8	42.7	-3.1	70.8	79.5	8.7	60.3	72.5	12.2
	δ	-11.0	-11.7	–	-16.7	-10.9	–	-9.2	-1.8	–	-10.6	-0.2	–

Table 3: Performance on the two tasks: TWP and TWS. PT/FT: Pre-trained/Fine-tuned. δ is the difference in performance between en-IN and en-US (en-IN minus en-US). Δ is the difference in performance between FT and PT.

Srirag, Dipankar et al. "Evaluating Dialect Robustness of Language Models via Conversation Understanding". SumEval workshop at COLING 2025.

NLP for dialects of a language

NLP Task	Paper	Impact
Language classification	[Blodgett et al. 2016]	Language detection shows lower performance for African-American English.
Sentiment classification	[Okpala et al. 2022]	Text in African-American English may be predicted more commonly as hate speech.
Natural Language Understanding	[Ziems et al. 2022]	Popular models perform worse on GLUE tasks for African-American English text.
Summarisation	[Keswani and Celis 2021]	Generated multi-document summaries may be biased towards majority dialect.
Machine translation	[Kantharuban et al. 2023]	Significant drop in MT from and to dialects of Portuguese/Bengali/etc. to and from English.
Parsing	[Scannell 2020]	Lower performance of parsers on Manx Gaelic as compared to Irish/Scottish Gaelic.

Table 1. Examples of adverse impact on NLP task performance due to dialectal variations.

Joshi, Aditya, et al. "Natural language processing for dialects of a language: A survey." *arXiv preprint arXiv:2401.05632* (2024). [To Appear: ACM CSUR 2025]

.. and for Creoles

	mBERT	XLM-R
Haitian-direct	51.60%	39.16%
Haitian-localized	50.83%	43.33%
Mauritian	49.10%	43.33%
English	63.33%	45.00%

Table 2: Accuracy results for MCTest160 development data when trained on the English MC160 training data.

Lent, Heather, et al. "CreoleVal: Multilingual multitask benchmarks for creoles." *Transactions of the Association for Computational Linguistics* 12 (2024): 950-978.

Prompting is biased towards dominant dialect

Models default to “standard” varieties for ten dialects of English

Tested on GPT-3.5 Turbo and GPT-4

Stereotyping, demeaning content, lack of comprehension and condescending responses

Variety of English	# Features: Inputs	# Features: Outputs	% Retention ↑
SAE	295	230	78%
SBE	291	210	72%
Indian	73	12	16%
Nigerian	44	5.5	13%
Kenyan	90	9	10%
Irish	26	1	4%
AAE	63	2	3%
Scottish	37	1	3%
Singaporean	40	1	3%
Jamaican	51	1	2%

Table 1: Overview of language varieties and features represented in inputs and GPT-3.5 outputs.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024.Linguistic Bias in ChatGPT: Language Models Reinforce Dialect Discrimination. In EMNLP.

Rise in datasets of language varieties

(Rise? Twelfth VarDial 2025 workshop happened yesterday! Do check out their papers!)

Rise in datasets of language varieties

(Rise? Twelfth VarDial 2025 workshop happened yesterday! Do check out their papers!)

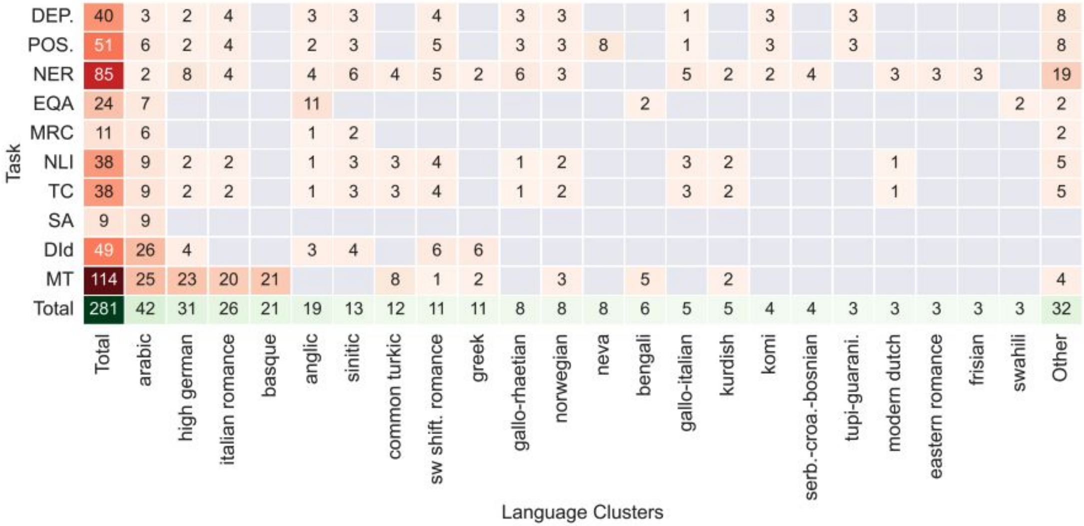
Task	Total	arabic	high german	italian romance	basque	anglic	sinitic	common turkic	sw shift. romance	greek	gallo-rhaetian	norwegian	neva	bengali	gallo-italian	kurdish	komi	serb.-croa.-bosnian	tupi-guarani.	modern dutch	eastern romance	frisian	swahili	Other
DEP.	40	3	2	4		3	3		4		3	3			1		3		3					8
POS.	51	6	2	4		2	3		5		3	3	8		1		3		3					8
NER	85	2	8	4		4	6	4	5	2	6	3			5	2	2	4		3	3	3		19
EQA	24	7				11								2									2	2
MRC	11	6				1	2																	2
NLI	38	9	2	2		1	3	3	4		1	2			3	2				1				5
TC	38	9	2	2		1	3	3	4		1	2			3	2				1				5
SA	9	9																						
DId	49	26	4			3	4		6	6														
MT	114	25	23	20	21			8	1	2		3		5		2								4
Total	281	42	31	26	21	19	13	12	11	11	8	8	8	6	5	5	4	4	3	3	3	3	3	32

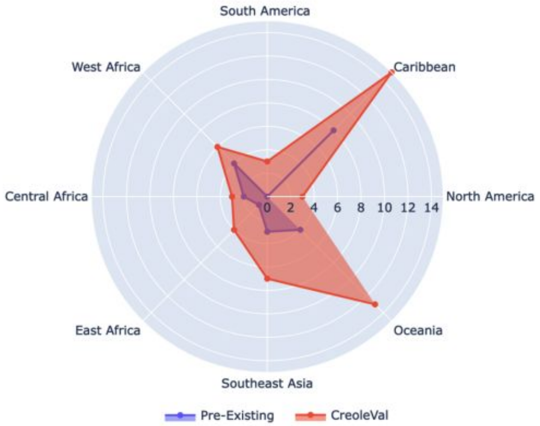
Language Clusters

Faisal, F., Ahia, O., Srivastava, A., Ahuja, K., Chiang, D., Tsvetkov, Y., & Anastasopoulos, A. (2024). DIALECTBENCH: A NLP Benchmark for Dialects, Varieties, and Closely-Related Languages. *ACL 2024.*

Rise in datasets of language varieties

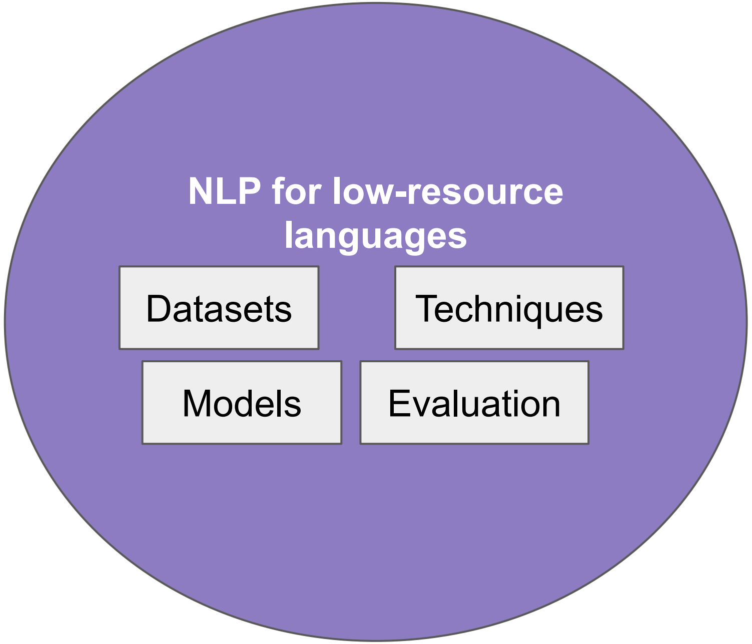
(Rise? Twelfth VarDial 2025 workshop happened yesterday! Do check out their papers!)



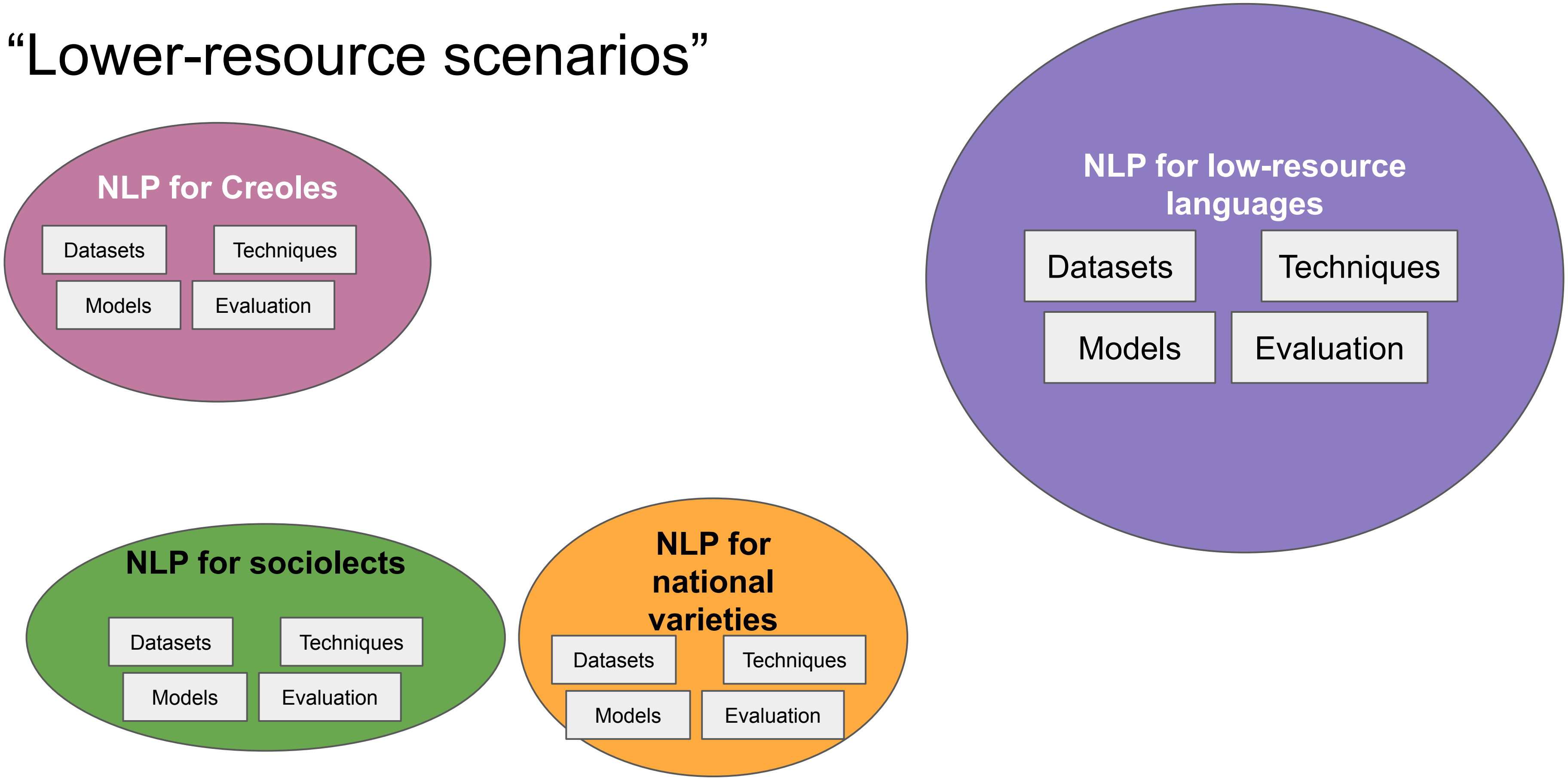


Faisal, F., Ahia, O., Srivastava, A., Ahuja, K., Chiang, D., Tsvetkov, Y., & Anastasopoulos, A. (2024). DIALECTBENCH: A NLP Benchmark for Dialects, Varieties, and Closely-Related Languages. *ACL 2024.*

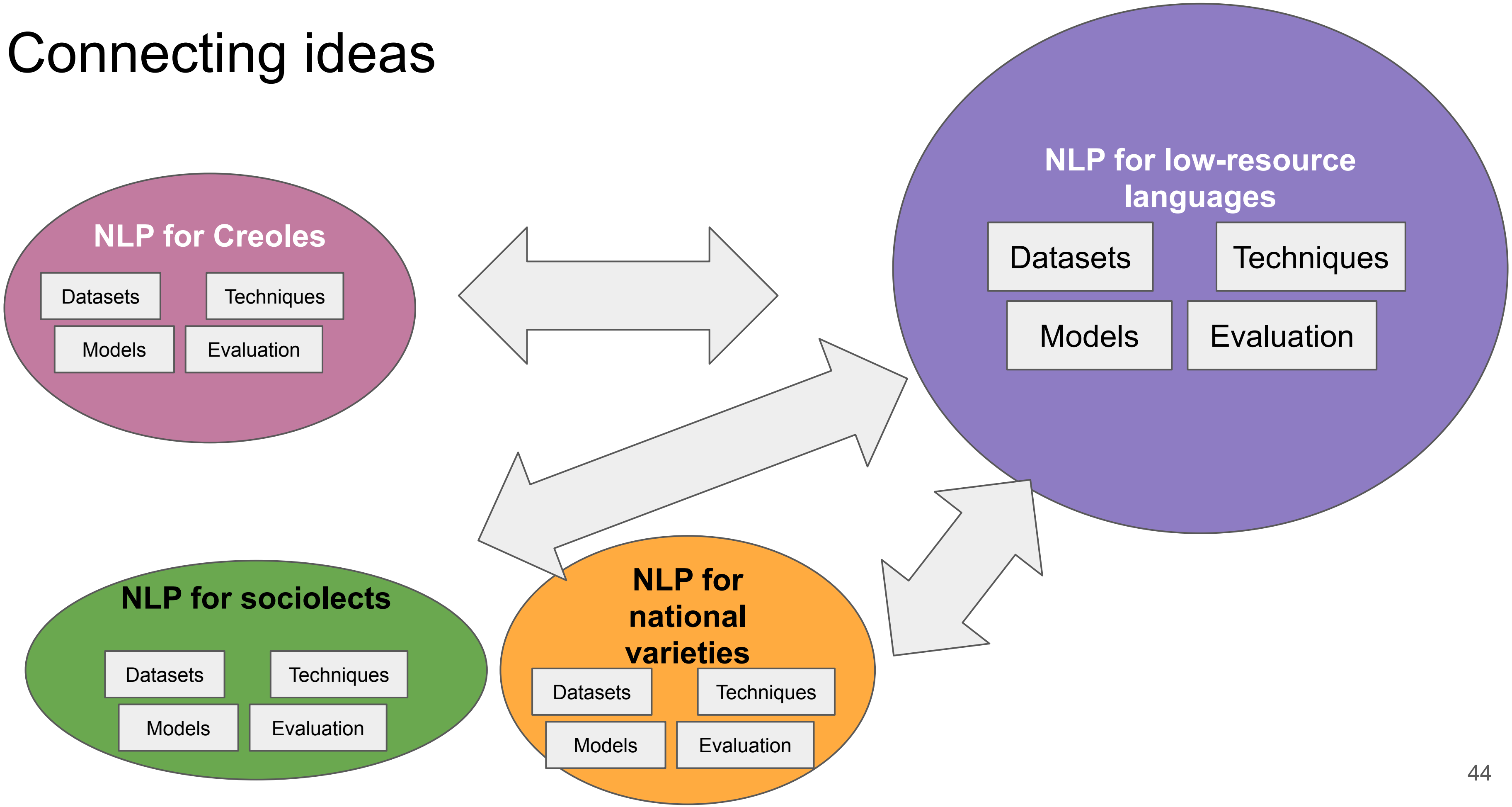
Lent, Heather, et al. "CreoleVal: Multilingual multitask benchmarks for creoles." *Transactions of the Association for Computational Linguistics* 12 (2024): 950-978.

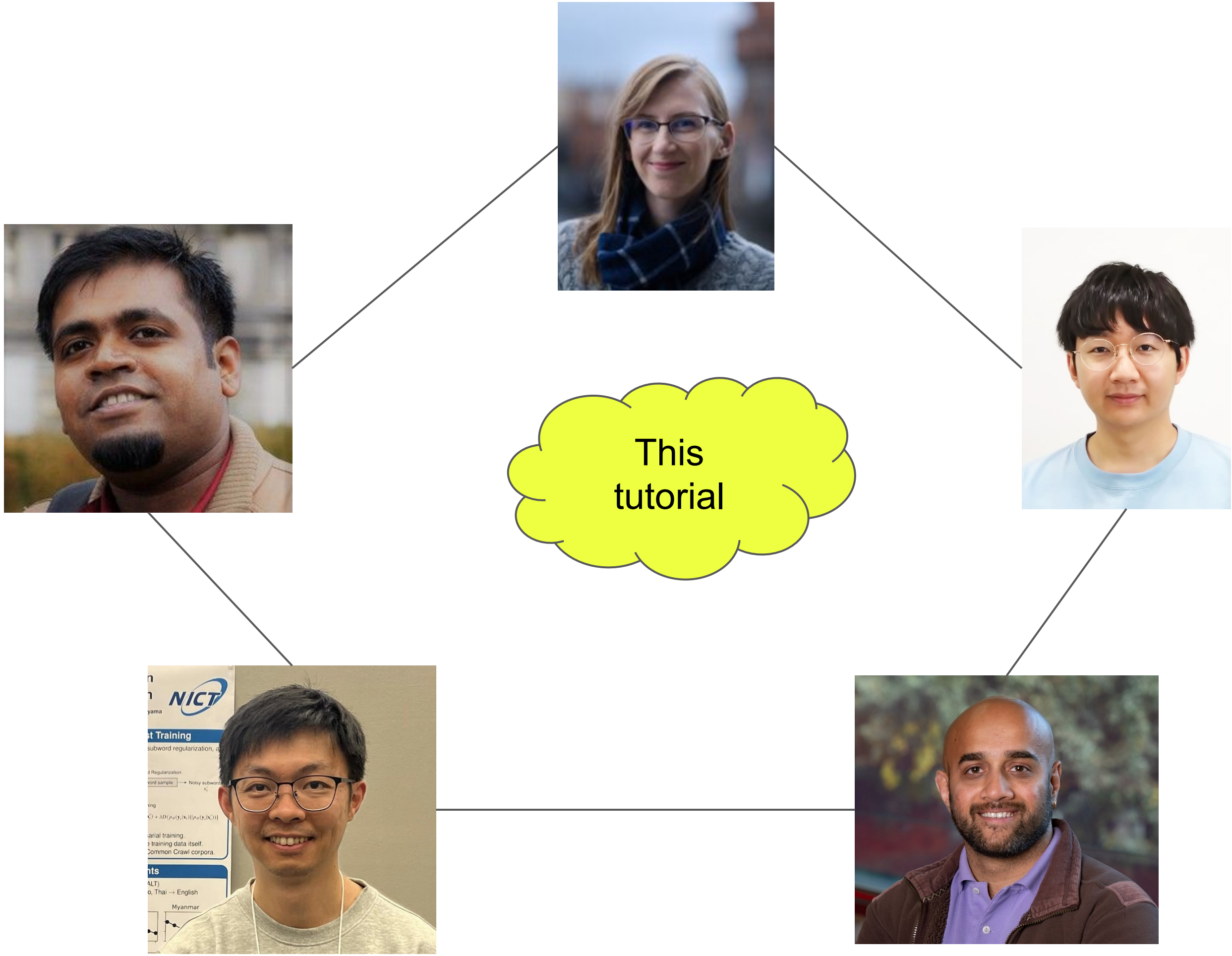

NLP for low-resource languages
Datasets
Techniques
Models
Evaluation

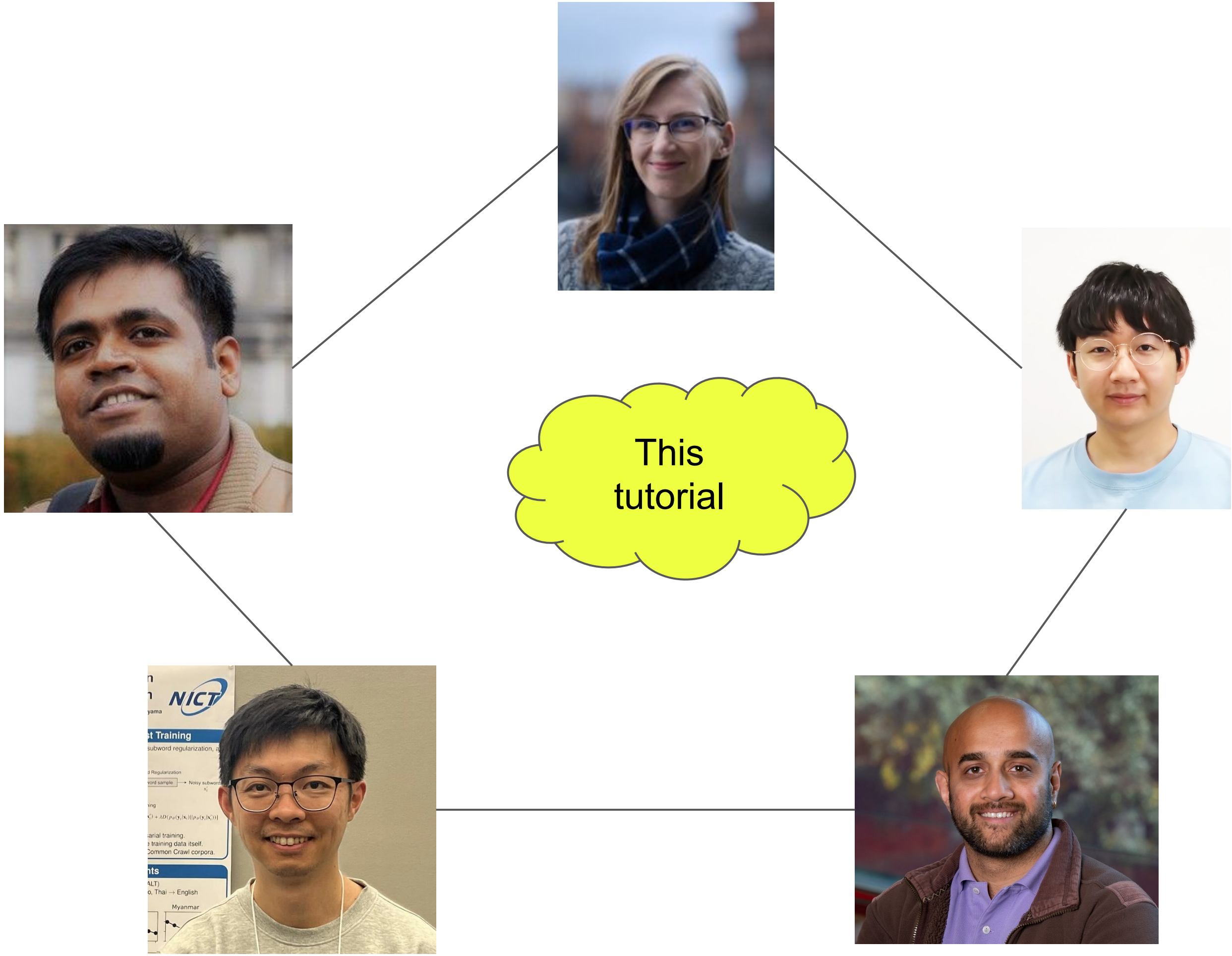
“Lower-resource scenarios”



Connecting ideas




This
tutorial



Common concerns and challenges.

Shared lessons.

Techniques that can be adapted between scenarios.

Why should we be interested?

Social Implications of Dialectal NLP

Performance of LLMs and per-capita GDP

Positive correlation between GDP per capita and performance of dialectal machine translation (Kantharuban et al., 2023)

Racial biases in LLM performance

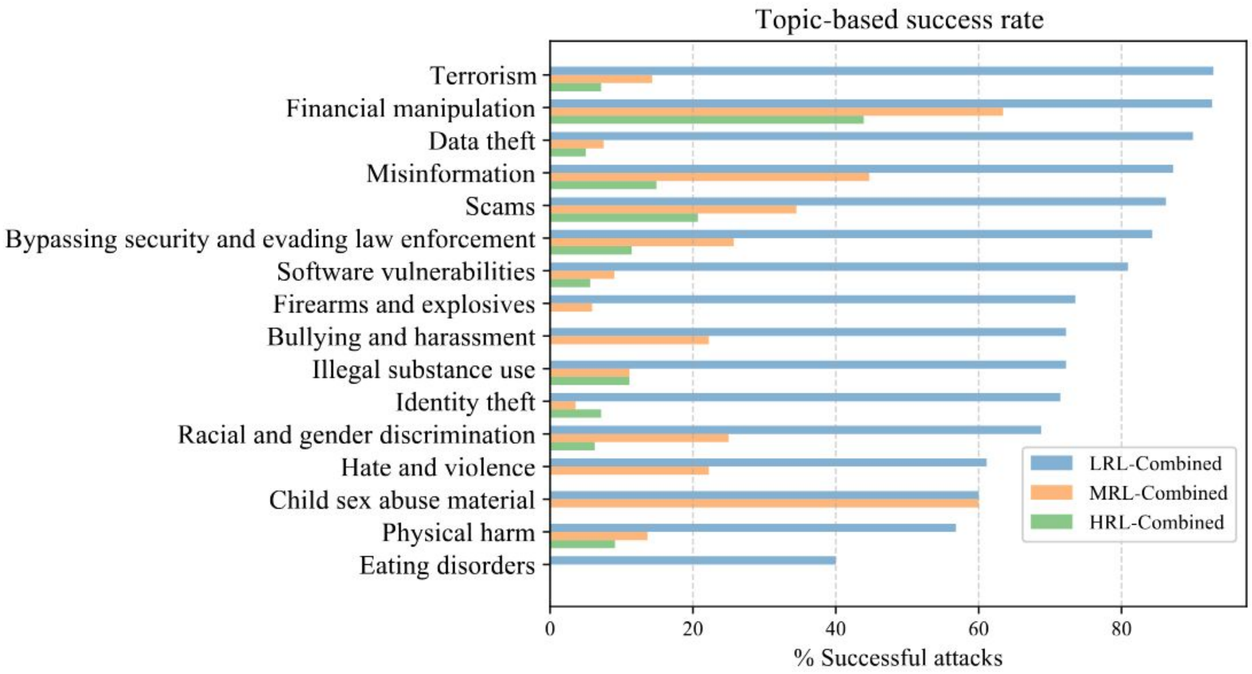
Hate speech classifiers may be biased against certain language varieties, i.e., African-American English (Okpala et al., 2022)

Prejudice in employability

Employability or criminality prediction using LLMs influenced by their choice of language variety (Hofmann et al, 2024)

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the Dialect Gap and its Correlates Across Languages. In Findings of EMNLP.
Ebuka Okpala, Long Cheng, Nicodemus Mbwambo, and Feng Luo. 2022. AAEBERT: Debiasing BERT-based Hate Speech Detection Models via Adversarial Learning. In ICMLA.
Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Dialect prejudice predicts AI decisions about people's character, employability, and criminality. arXiv preprint arXiv:2403.00742 (2024).

Security Implications of Low-Resource NLP



Low-resource scenarios can be **weaponized** against LLMs, with a threat of safety to all.

Yong, Z., Menghini, C., & Bach, S.H. (2023). Low-Resource Languages Jailbreak GPT-4. *NeurIPS Workshop on Socially Responsible Language Modelling Research (SoLaR) 2023. Best Paper Award.*

“Dubious assumptions” of ‘Language Technology for all’

More than this even, the agenda of *Language Technology for All* rests on dubious assumptions:

1. that language technologies must be capable of simulating human communication;
2. that the Eurocentric practice of delimiting languages should be applied globally;
3. that all languages should be standardised;
4. that all languages have a standard orthography or would benefit from one;
5. that vernacular language literacy is universal, or universally desirable;
6. that all people are monolingual and use a single language for all communicative functions;
7. that all people use pure language, not routinely mixing vernaculars, or mixing the vernacular with the vehicular;
8. that human communication is adequately represented by the noisy-channel model;
9. that language technology scalability requires one-size-fits-all solutions; and
10. that sufficient manipulation of linguistic forms will ultimately arrive at meaning.

Steven Bird. 2024. Must NLP be Extractive?. In *Proceedings of ACL*.

Lower-resource NLP is a balancing act

Communities,
their language, & their
wants and needs from HLT

NLP community norms,
What's publishable (papers),
What's fundable (grants)

Lower-resource NLP is a balancing act

Communities,
their language, & their
wants and needs from HLT

“We (🇭🇹) need a spell-checker”
versus “We (🇳🇬) want speech tech”

NLP community norms,
What’s publishable (papers),
What’s fundable (grants)

“Massively multilingual”

“Scalable across languages”

Lower-resource NLP is a balancing act

Communities,
their language, & their
wants and needs from HLT

“We (🇭🇹) need a spell-checker”
versus “We (🇳🇬) want speech tech”

Reality for end-users:
Garbage in, garbage out.

Data is not
culturally-appropriate

NLP community norms,
What’s publishable (papers),
What’s fundable (grants)

“Massively multilingual”

“Scalable across languages”

More coverage at
expense of quality

Performance increases! …
But over poor data quality

Upcoming conference themes…

NAACL 2025 Theme Track: NLP in a Multicultural World

Current NLP tools and models, especially LLMs, require vast amounts of data to train. However, the data used often favors only a handful of over-represented languages, and even for these majoritarian languages only some of th geographical or cultural varieties are considered, leaving a large tail o represented languages, varieties, and cultures that have had consider attention from the NLP community. In this year's theme track we woul focus on work providing support to the vibrant multicultural world we welcome papers in the following non-exhaustive list of topics:

- Cultural localization of language models.
- New NLP applications to support people from diverse cultures.
- Revitalization or refunctionalization of endangered or sleeping la
- Analysis of cultural biases in language models.
- Historical considerations and diachronic analysis.

ACL 2025 Theme Track: Generalization of NLP Mo

Following the success of the ACL 2020-2024 Theme tracks, we are happy to announce that ACL 2025 will have a new theme with the g reflecting and stimulating discussion about the current state of development of the field of NLP.

Generalization is crucial for ensuring that models behave robustly, reliably, and fairly when making predictions on data different from their training data. Achieving good generalization is critically important for models used in real-world applications, as they should emulate human-like behavior. Humans are known for their ability to generalize well, and models should aspire to this standard.

The theme track invites empirical and theoretical research and position and survey papers reflecting on the Generalization of NLP Models. The possible topics of discussion include (but are not limited to) the following:

- How can we enhance the generalization of NLP models across various dimensions—compositional, structural, cross-task, cross-lingual, cross-domain, and robustness?
- What factors affect the generalization of NLP models?
- What are the most effective methods for evaluating the generalization capabilities of NLP models?
- While Large Language Models (LLMs) significantly enhance the generalization of NLP models, what are the key limitations of LLMs in this regard?

Interspeech 2025 will delve into **four specific strands**, each addressing critical aspects of speech scienc

1. **Factors Arising from the Individual in Human Speech Processing**
   - Exploration of individual differences in speech processing.
   - Understanding how personal factors influence speech perception and production.
   - Development of personalized speech technology applications.
2. **Under-Researched Languages, Dialects, and Accents**
   - Focus on linguistic diversity and the inclusion of under-researched languages and dialects.
   - Efforts to develop speech technologies that accommodate a wide range of accents.
   - Promotion of research that highlights the richness of global linguistic diversity.

https://2025.naacl.org/calls/papers/#naacl-2025-theme-track-nlp-in-a-multicultural-world
https://2025.aclweb.org/calls/main_conference_papers/#acl-2025-theme-track-generalization-of-nlp-models
https://www.interspeech2025.org/home

Connecting Ideas in ‘Lower-Resource’ Scenarios: NLP for National Varieties, Creoles and Other Low-resource Scenarios

Tutorial Objectives

LO1: Comprehend techniques across natural language understanding and generation for lower-resource scenarios.

LO2: Experiment with popular techniques using datasets and sample code provided.

LO3: Apply techniques for their research.

LO4: Appreciate the relationship between the scenarios from a computational perspective.

Tutorial Agenda

Introduction

Dataset Creation

NLG

Emerging Connections

NLU

Conclusion

Tutorial Agenda

Introduction ✓	Dataset Creation	NLG
Emerging Connections	NLU	Conclusion

Discussion Time

“Always name the language you’re working with”

-Emily Bender

https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html

Discussion Time

"Always name the language you're working with"

-Emily Bender

What is the extended notion of a "language" in the 'lower resource scenario'? What exactly should we name?

https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html

Discussion Time

“Always name the language you’re working with”

-Emily Bender

What is the extended notion of a “language” in the ‘lower resource scenario’? What exactly should we name?

Potentially contentious question: *Aren’t language varieties merely “domains”?*

https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html