The State and Fate of Linguistic Diversity in the NLP world

Pratik Joshi*, Sebastin Santy*, Amar Budhiraja*,
Kalika Bali, Monojit Choudhury

Data Sciene in India – IKDD

ACL 2020 Theme Track
“ Taking stock of where the community is and where it is going. ”

Microsoft
Microsoft Research Bangalore, India

aka.ms/statefate

The Actual State

	Dutch	Somali
#Speakers	29M	18M
#Resources (LDC+ELRA)	69	2
Other details	SOTA translation systems	Very few, inferior translation systems

The tiger moved across the grass

Dutch Translation

De tijger bewoog over het gras

English Translation

The tiger moved across the grass

Somali Translation

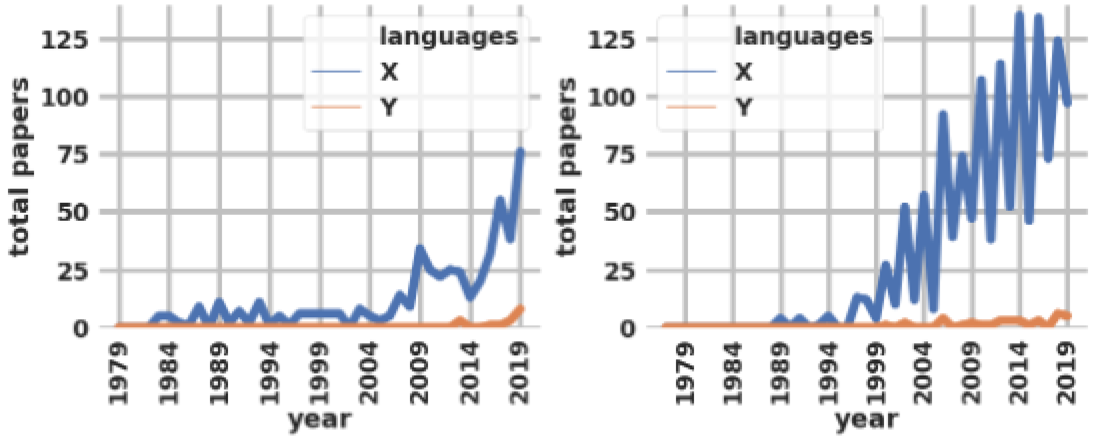
Digirta ayaa ku dul dhaqaaqday cawska

English Translation

Beans moved on grass

The Actual State

	Dutch	Somali
#Speakers	29M	18M
#Resources (LDC+ELRA)	69	2
Other details	SOTA translation systems	Very few, inferior translation systems



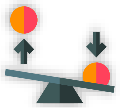
(a) ACL + NAACL + EACL + EMNLP

(b) LREC + WS

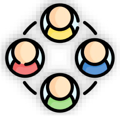
The Questions

How has the fate of different languages changed with current language technologies?

1. How many resources are available across the World's languages, and do they correlate with the number of speakers?

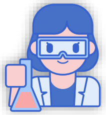


2. Which typological features have NLP systems been exposed to? Which features have been underrepresented?



3. How inclusive has ACL been in conducting and publishing research for different languages?

4. Does resource availability influence the research questions and publication venue?



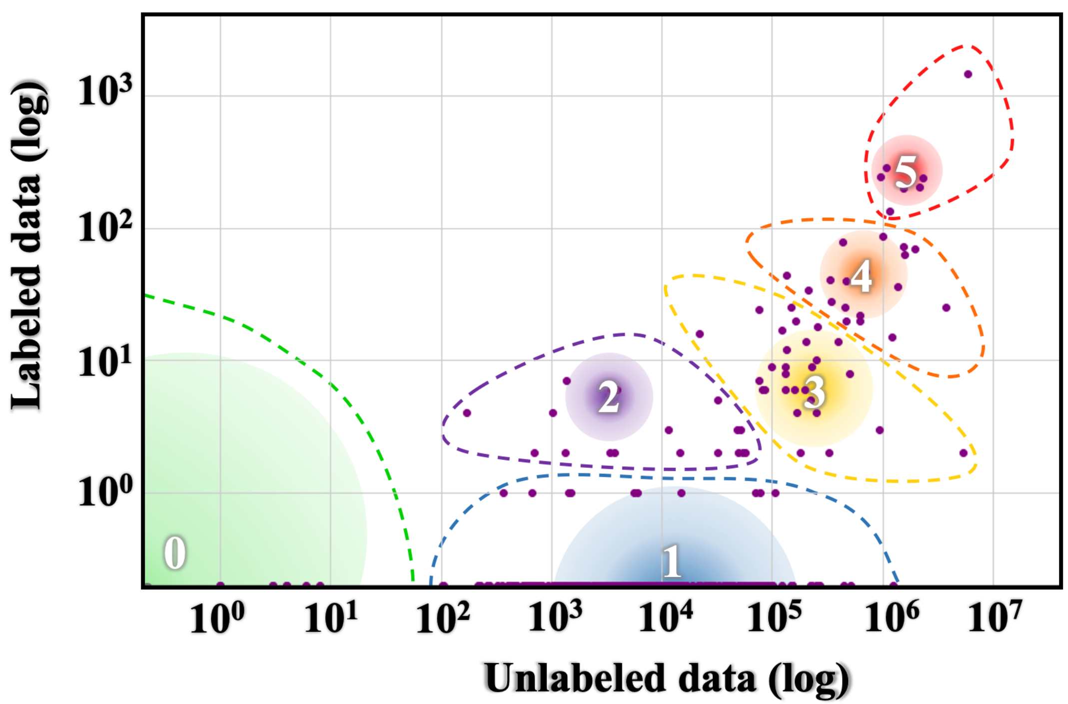
5. What role does an individual researcher or community have in bridging the resource divide?

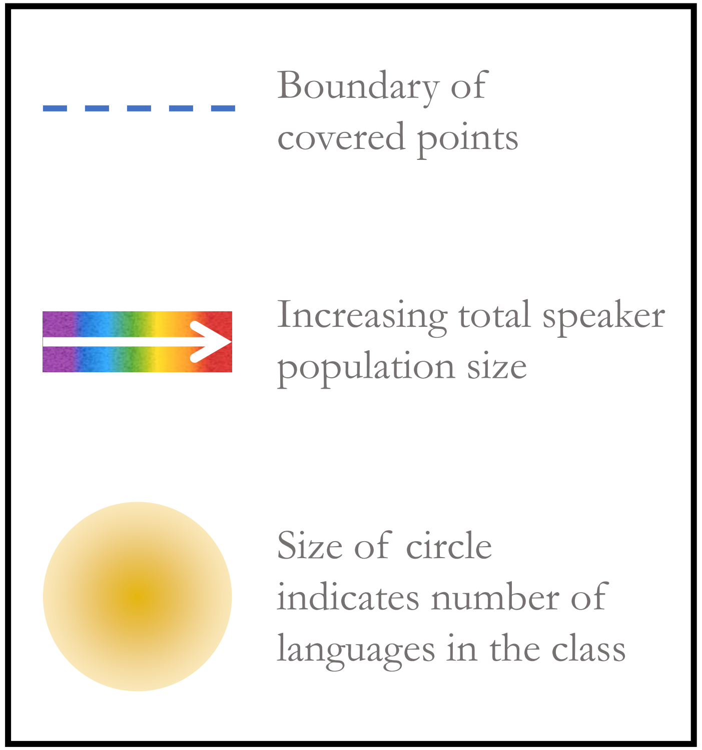
The Language Taxonomy

Setup

Labeled data
LDC catalog,
ELRA Map

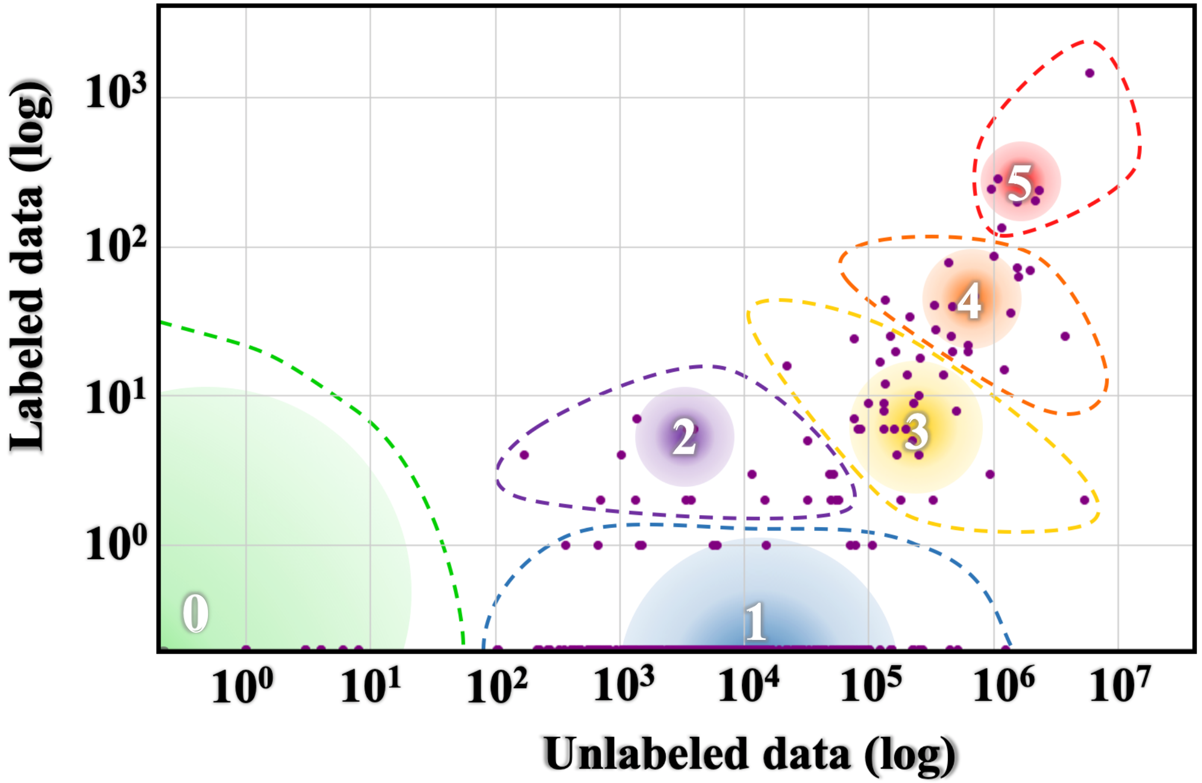
Unlabeled data
Wikipedia pages
(used in pretraining
language models)





The Language Taxonomy

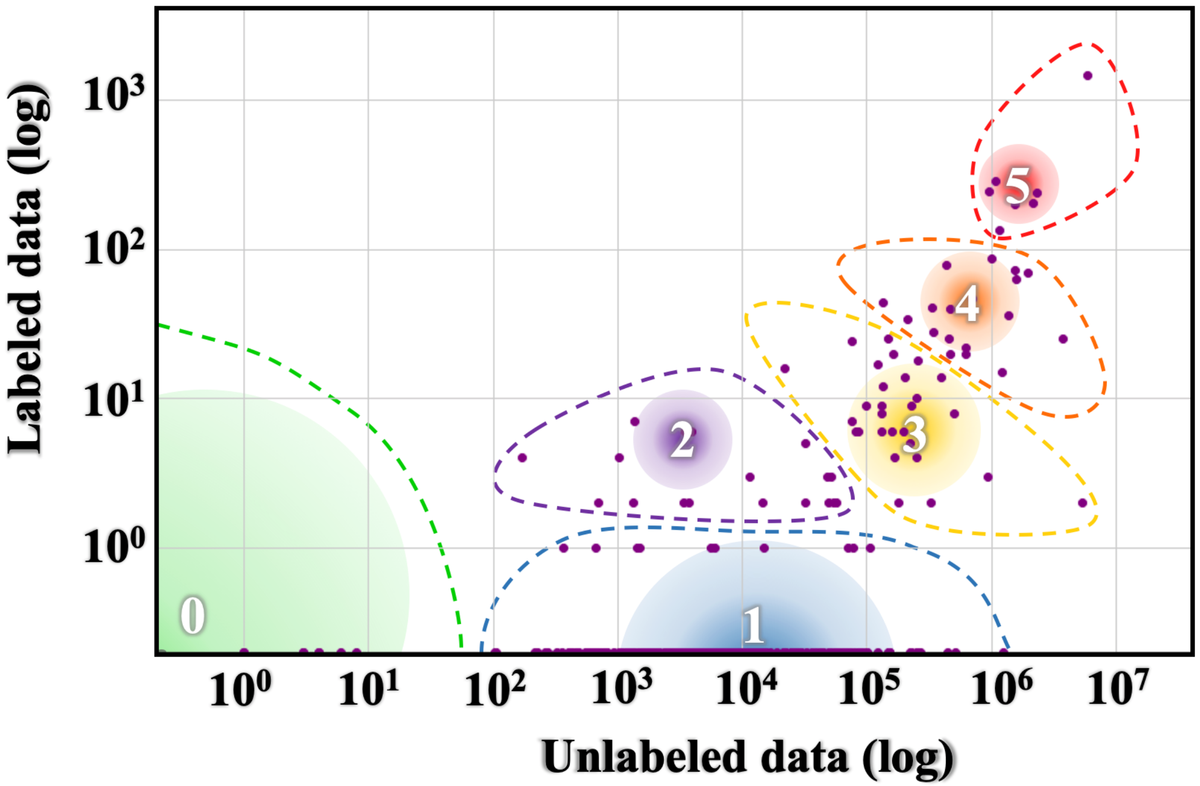
Visualization



The Language Taxonomy

Visualization

- Class 0 (The left-behinds) - Gondi, Mundari
- Class 1 (The Scraping-Bys) - Bhojpuri, Assamese
- Class 2 (The Hopefuls) - Konkani, Wolof
- Class 3 (The Rising Stars) - Tamil, Marathi
- Class 4 (The Underdogs) - Bengali, Hindi
- Class 5 (The Winners) - English, French



The Language Taxonomy

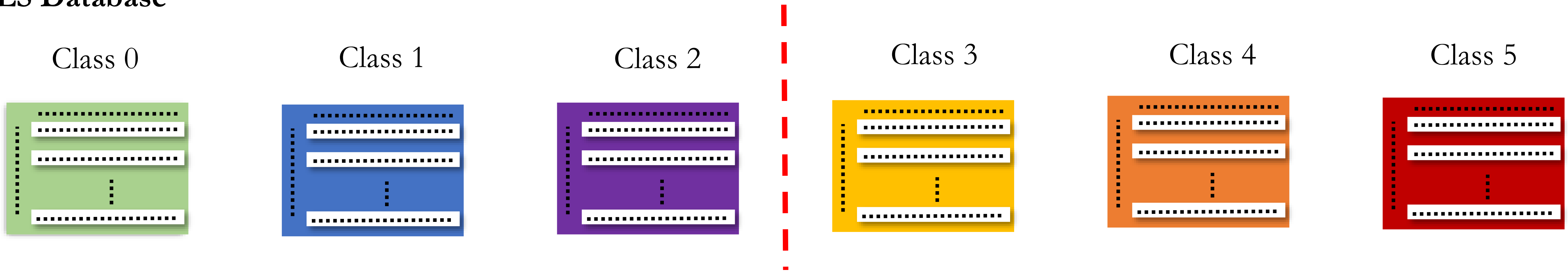
Large population left behind

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Typological Representation

Setup

WALS Database



- Typological features refer to properties/attributes of a language.

- Categories in languages of classes 0,1,2 but not 3,4,5 are 'ignored' categories.

- We then look at typological features with most 'ignored' vs. least 'ignored' categories.

Typological Representation

Far-reaching repercussions

Feature	#Cat	#Lang
144E	23	38
144M	23	45
144F	22	48
144O	21	30

Feature	#Cat	#Lang
83A	0	1321
82A	0	1302
97A	0	1146
86A	0	1083

Table 2: Most and Least 'ignored' typological features, the number of categories in each feature which have been ignored, and the number of languages which contain this feature.

Typological Representation

Transfer Learning for Family Languages

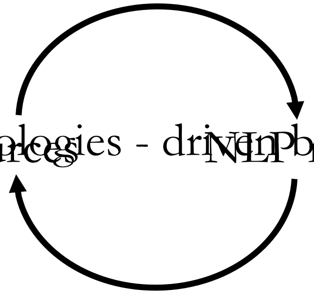
Semitic Family

Language	Class	#Speakers	'Ignored'	Error
Amharic	2	22M	9	60.71
Arabic	4	300M	0	7.8

Language and Conference

What?

Language technologies - driven by NLP Research



Conference	h-index	Remarks
ACL/NAACL/ EMNLP/EACL	106/61/ 88/36	Data-Driven lately
CL (Journal)	25	Computational Linguistics focused
COLING	41	Oldest conference
LREC	45	Multilingual Research
WS (Workshop Proceedings)	n/a	Factoring papers accepted in workshops of above conferences

Language and Conference

Year-wise Language Occurrence

- Understand how multilinguality is changing over conference iterations
- Language mentions in papers are a measure for language inclusion
- Use Entropy as a unified measure to calculate skew in language distribution for a conference iteration.
- No. of languages = $(2)^{\text{entropy}}$

Language and Conference

Year-wise Language Occurrence

- Understand how multilinguality is changing over conference iterations
- Language mentions in papers are a measure for language inclusion
- Use Entropy as a unified measure to calculate skew in language distribution for a conference iteration.
- No. of languages = $(2)^{\text{entropy}}$

Conference = ACL 2019

Langs →	af	am	ar	de	en	es	hi	it	pt	tr	vi	…….	zh
All Papers	3	4	10	9	31	11	7	9	8	4	1	…….	10

Language and Conference

Year-wise Language Occurrence

- Understand how multilinguality is changing over conference iterations
- Language mentions in papers are a measure for language inclusion
- Use Entropy as a unified measure to calculate skew in language distribution for a conference iteration.
- No. of languages = $(2)^{\text{entropy}}$

Conference = ACL 2019

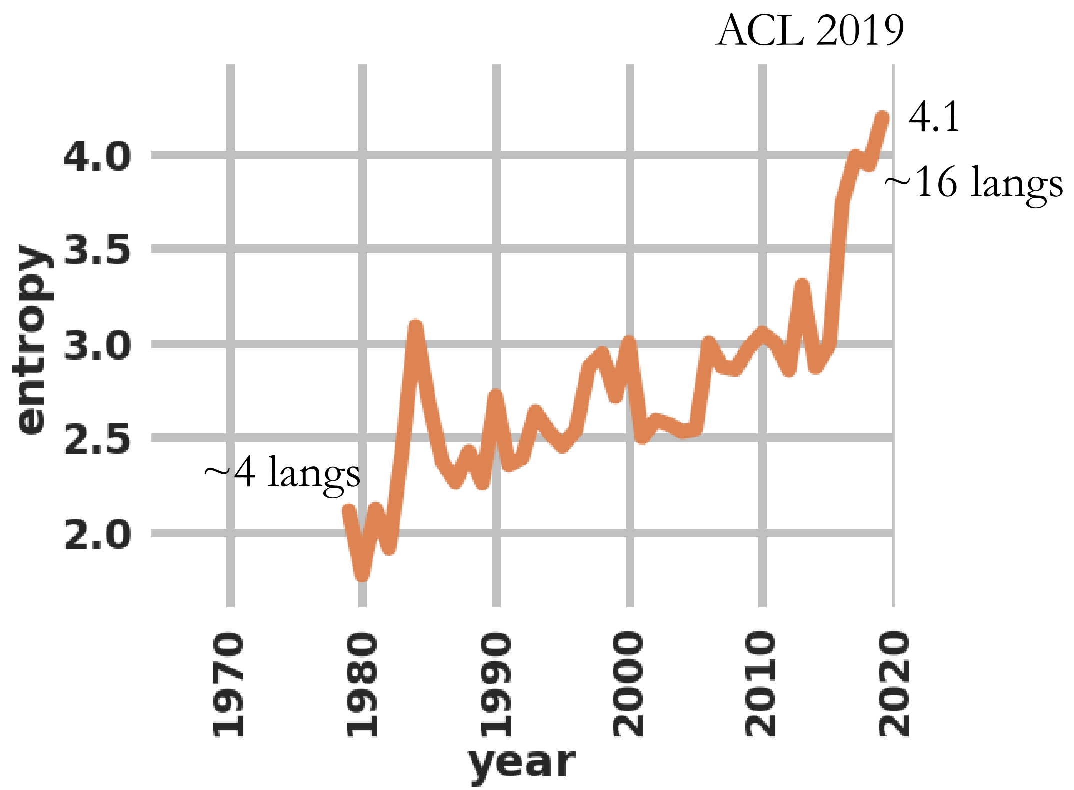
Entropy

4.1

Language and Conference

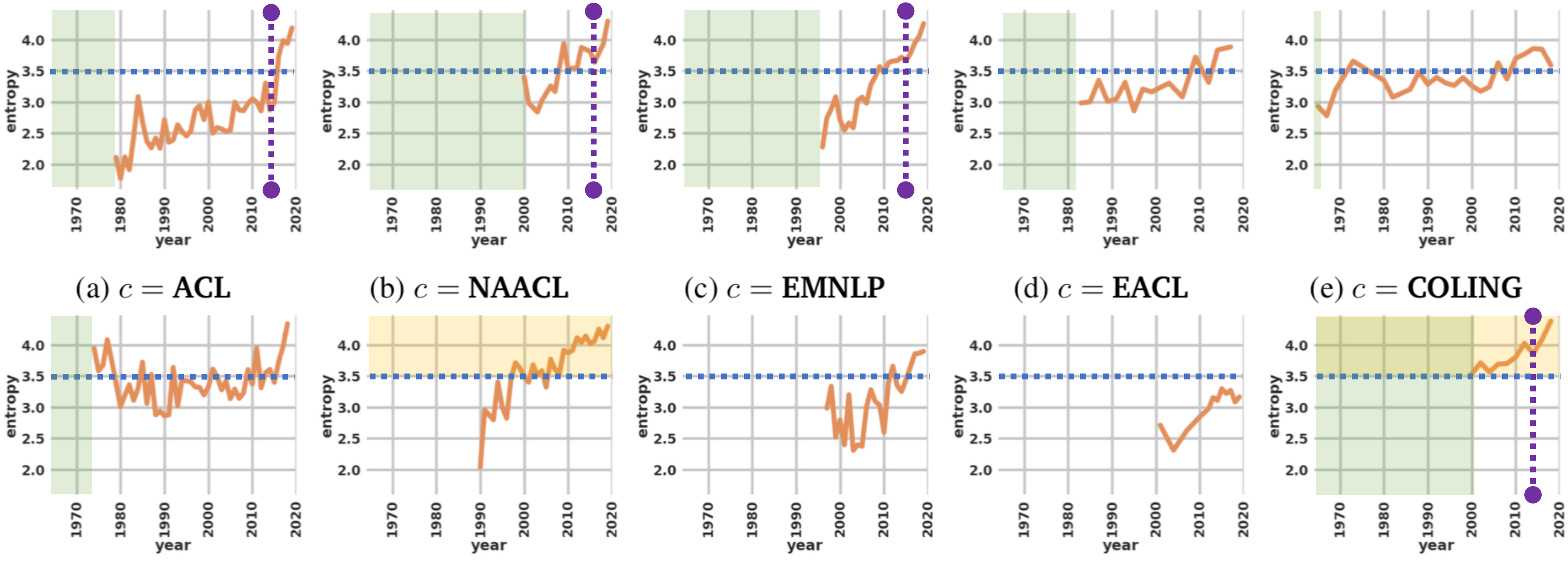
Year-wise Language Occurrence

- Understand how multilinguality is changing over conference iterations
- Language mentions in papers are a measure for language inclusion
- Use Entropy as a unified measure to calculate skew in language distribution for a conference iteration.
- No. of languages = $(2)^{\text{entropy}}$



Language and Conference

Year-wise Language Occurrence



(a) $c =$ **ACL** (b) $c =$ **NAACL** (c) $c =$ **EMNLP** (d) $c =$ **EACL** (e) $c =$ **COLING**

(f) $c =$ **CL** (g) $c =$ **WS** (h) $c =$ **CONLL** (i) $c =$ **SEMEVAL** (j) $c =$ **LREC**

Language and Conference

Class-wise Language Representation

Determine the standing of each language class in a conference.

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}$$

rank_i → language rank in a particular conference ordered by mention frequency.

Q → number of languages in each class.

Conf ↓ / Class →	0	1	2	3	4	5
ACL	725	372	157	63	20	3
CL	647	401	175	76	27	3
COLING	670	462	185	74	21	2
CONLL	836	576	224	64	16	3
EACL	839	514	195	63	15	3
EMNLP	698	367	172	67	19	3
LREC	811	261	104	45	13	2
NAACL	754	365	136	63	18	3
SEMEVAL	730	983	296	121	19	3
TACL	974	400	180	50	15	3
WS	667	293	133	59	15	3

Heterogenous Entity Embeddings

Motivation

Previous analysis indicates variance in acceptance of different languages across different NLP conferences

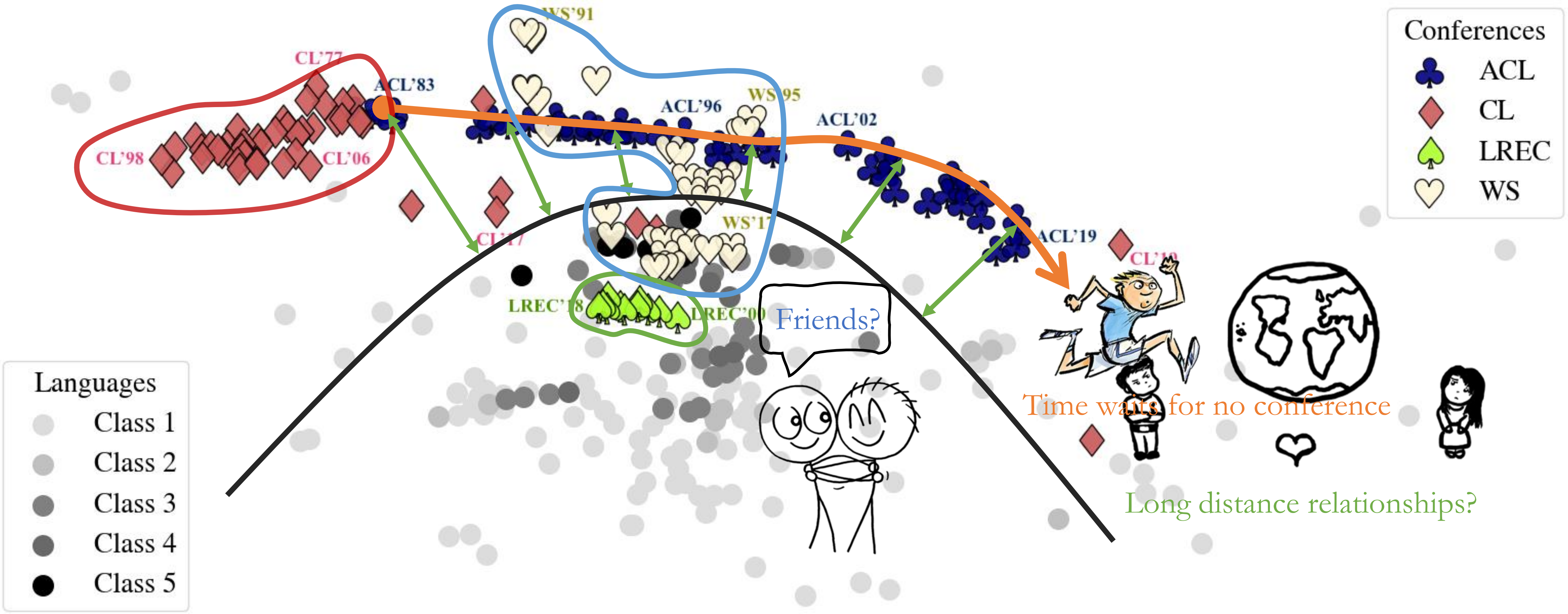
Vanilla statistics fail to capture the subtle nuances in the data that might be affecting these outcomes

Embeddings have been shown to capture complex relationships directly from the data without supervision

Proposal: Jointly learn the representations of Conferences, Authors and Languages, collectively termed as Entities.

Heterogenous Entity Embeddings

Spatial Representation



Heterogenous Entity Embeddings

Role of Community

Not all superheroes wear capes

- Mean Reciprocal Rank (MRR) of a language signifies how many authors in the research community are exclusively close to this language.
- Higher MRR indicates more focused research community

Class	MRR(10)
0	0.69146
1	0.52585
2	0.45265
3	0.52670
4	0.47795
5	0.51471

Takeaways

Recommendations

Evident Taxonomy

Typology Consideration

Inclusive Conferences

Focused Communities

A call to look at the language disparity at the conferences

Linguistic Diversity & Inclusion clauses

aka.ms/statefate