The State and Fate of Linguistic Diversity and Inclusion in the NLP World

Pratik Joshi[*]   **Sebastin Santy**[*]   **Amar Budhiraja**[*]
Kalika Bali   **Monojit Choudhury**
Microsoft Research, India
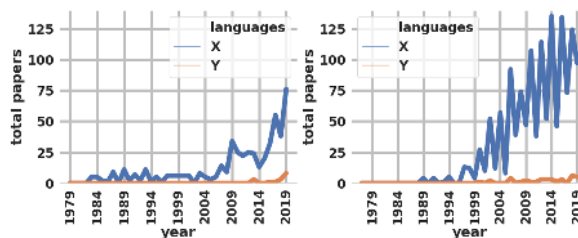{t-prjos, t-sesan, amar.budhiraja, kalikab, monojitc}@microsoft.com

Abstract

Language technologies contribute to promoting multilingualism and linguistic diversity around the world. However, only a very small number of the over 7000 languages of the world are represented in the rapidly evolving language technologies and applications. In this paper we look at the relation between the types of languages, resources, and their representation in NLP conferences to understand the trajectory that different languages have followed over time. Our quantitative investigation underlines the disparity between languages, especially in terms of their resources, and calls into question the "language agnostic" status of current models and systems. Through this paper, we attempt to convince the ACL community to prioritise the resolution of the predicaments highlighted here, so that no language is left behind.

1 The Questions

Languages **X** and **Y** are the official languages of two different countries; they have around 29M and 18M native speakers, and 2M and 5.5K Wikipedia articles, respectively. **X** is syntactically quite similar to English, though uses dimunitives and has grammatical gender. **Y**, on the other hand, has a different word order from English, and has a rare typological feature - generally it is a head-final language, but noun phrases are head-initial. It also features full and partial reduplication. 69 items on LDC and ELRA contain data in **X**, whereas for **Y** there are only 2 items. **X** boasts of some of the best online machine translation systems, whereas **Y** is supported by very few online MT systems and that too with far inferior translation quality. Figure 1 shows the number of papers in conferences (**ACL**, **NAACL**, **EACL**, **EMNLP**, **LREC**, **WS**) that

(a) ACL + NAACL + EACL + EMNLP   (b) LREC + WS

Figure 1: Number of papers with mentions of **X** and **Y** language for two sets of conferences.

mention **X** and **Y** in the paper, across the years. As you can see, while **X** has a steady and growing trend of research, our community has been mostly oblivious to **Y**, until recently when some of the zero-shot learning papers have started mentioning it. Can you guess what **X** and **Y** are?

Regardless of whether you can guess the exact answer, most NLP researchers surely know of (and might even speak) several languages which are in the same boat as **X**; languages which have a large amount of resources and therefore access to the benefits of the current NLP breakthroughs, and languages like **Y**; those which lack resources and consequently the attention of the NLP community, despite having similar speaker base sizes and typologically diverse features.

You probably have come across the issue of extremely skewed distribution of resources across the world's languages before. You might also be aware of the fact that most of our NLP systems, which are typically declared language agnostic, are not truly so (Bender, 2011). The handful of languages on which NLP systems are trained and tested are often related and from the same geography, drawn from a few dominant language families, leading to a typological echo-chamber. As a result, a vast majority of typologically diverse linguistic phenomena are never seen by our NLP systems (Ponti et al., 2019).

---

[*] Authors contributed equally to the work.
https://microsoft.github.io/linguisticdiversity

Nevertheless, it would be prudent to re-examine these issues in the light of recent advances in deep learning. Neural systems, on one hand, require a lot more data for training than rule-based or traditional ML systems, creating a bigger technological divide between the **X**s and **Y**s; yet, some of the most recent techniques on zero-shot learning of massively multilingual systems (Devlin et al., 2019; Conneau and Lample, 2019; Aharoni et al., 2019; Artetxe and Schwenk, 2019) bridge this gap by obliterating the need for large labeled datasets in all languages. Instead, they need only large unlabeled corpora across languages and labeled data in only some languages. Assuming that this approach can be taken to its promising end, how does the fate of different languages change?

We break down this complex prescient question into the following more tractable and quantifiable questions on *Linguistic Diversity and Inclusion*:

1. How many resources, labeled and unlabeled, are available across the World's languages? How does this distribution correlate to their number of native speakers? What can we expect to achieve today and in the near future for these languages?

2. Which typological features have current NLP systems been exposed to, and which typological features mostly remain unexplored by systems because we have hardly created any resources and conducted data-driven research in those languages?

3. As a community, how inclusive has **ACL** been in conducting and publishing research on various languages? In 1980s and early 90s, when large scale datasets were not the prime drivers of research, was the linguistic diversity of **ACL** higher than what it has been in 2000s and 2010s? Or has **ACL** become more inclusive and diverse over the years?

4. Does the amount of resource available in a language influence the research questions and the venue of publication? If so, how?

5. What role does an individual researcher, or a research community have to play in bridging the linguistic-resource divide?

In this paper, we take a multi-pronged quantitative approach to study and answer the aforementioned questions, presented in order, in the following five sections. One of the key findings of our study, to spill the beans a bit, is that the languages of the World can be broadly classified into 6 classes based on how much and what kind of resources they have; the languages in each class have followed a

distinct and different trajectory in the history of **ACL**, and some of the hitherto neglected classes of languages have more hope of coming to the forefront of NLP technology with the promised potential of zero-shot learning.

2 The Six Kinds of Languages

In order to summarize the digital status and 'richness' of languages in the context of data availability, we propose a taxonomy based on the number of language resources which exist for different languages. We frame the rest of our analyses based on this taxonomy and use it to emphasize the existence of such resource disparities.

2.1 Features

We design this taxonomy using two feature axes: number of unlabeled resources vs. number of labeled resources. Previous methods have mostly relied on supervised learning techniques which require labeled corpora. However, the advent of transfer learning methods have boosted the importance of unlabeled data: massively multilingual models such as mBERT use Wikipedia for pre-training, and then fine-tune on downstream NLP tasks. These features are suitable because the current NLP research is predominantly data-driven, and language inclusion depends on how much labeled or unlabeled data is available. We believe these features are sufficient for the taxonomical design as the required metadata is consistently available across all languages, whereas features such as number of hours required to collect data aren't available.

We treat each data resource as a fundamental unit, based on the assumption that the collection of one unit is proportional to a certain extent of effort being invested towards the resource improvement of that language. Moreover, this feature discretization is unambiguous and concrete. Other units such as the total number of datapoints across datasets can be misleading because different NLP tasks have different data requirements. For example, while Machine Translation (MT) models require datapoints to the order of millions (Koehn and Knowles, 2017) to perform competitively, competent models in Question Answering require around 100 thousand datapoints (Rajpurkar et al., 2016). Moreover, the unit of datapoints vary across different technologies (e.g. Speech data measured in hours, MT data measured in number of parallel sentences).
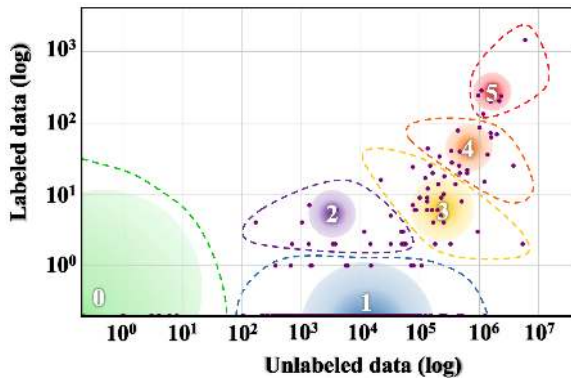
Figure 2: Language Resource Distribution: The size of the gradient circle represents the number of languages in the class. The color spectrum VIBGYOR, represents the total speaker population size from low to high. Bounding curves used to demonstrate covered points by that language class.

2.2 Repositories

We focus our attention on the LDC catalog[1] and the ELRA Map[2] for labeled datasets. Although there are other repositories of data available online, we found it practical to treat these organized collections as a representation of labeled dataset availability. This way, we look at standardized datasets that have established data quality and consistency, and which have been used in prior work. There are strong efforts such as PanLex (Kamholz et al., 2014), which is a large lexical database of a wide range of languages being used for a lexical translator, and OLAC (Simons and Bird, 2003), which contains a range of information for different languages (e.g. text collections, audio recordings, and dictionaries). However, keeping within the purview of NLP datasets used in *CL conferences, we decided to focus on popular repositories such as the above-mentioned.

We look at Wikipedia pages as a measure for unlabeled data resources. With regards to language technologies, Wikipedia pages represent a strong source of unsupervised training data which are freely and easily accessible. In the perspective of digital resource availability, they are a comprehensive source of factual information and are accessed by a large, diverse set of online users.

2.3 Language Classes

Figure 2 is a visualization of the taxonomy. We find a set of distinct partitions which can be used

to categorize languages into 6 unique positions in the language resource 'race':

0 - The Left-Behinds These languages have been and are still ignored in the aspect of language technologies. With exceptionally limited resources, it will be a monumentous, probably impossible effort to lift them up in the digital space. Unsupervised pre-training methods only make the 'poor poorer', since there is virtually no unlabeled data to use.

1 - The Scraping-Bys With some amount of unlabeled data, there is a possibility that they could be in a better position in the 'race' in a matter of years. However, this task will take a solid, organized movement that increases awareness about these languages, and also sparks a strong effort to collect labelled datasets for them, seeing as they have almost none.

2 - The Hopefuls With light at the end of the tunnel, these languages still fight on with their gasping breath. A small set of labeled datasets has been collected for these languages, meaning that there are researchers and language support communities which strive to keep them alive in the digital world. Promising NLP tools can be created for these languages a few years down the line.

3 - The Rising Stars Unsupervised pre-training has been an energy boost for these languages. With a strong web presence, there is a thriving cultural community online for them. However, they have been let down by insufficient efforts in labeled data collection. With the right steps, these languages can be very well off if they continue to ride the 'pre-training' wave.

4 - The Underdogs Powerful and capable, these languages pack serious amounts of resource 'firepower'. They have a large amount of unlabeled data, comparable to those possessed by the winners, and are only challenged by lesser amount of labeled data. With dedicated NLP communities conducting research on these languages, they have the potential to become winners and enjoy the fruits of 'digital superiority'.

5 - The Winners Running strong and fast, these languages have been in the lead for quite a while now, some longer than others. With a dominant online presence, there have been massive industrial and government investments in the development of resources and technologies for these languages. They are the quintessential rich-resource

---

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%

Table 1: Number of languages, number of speakers, and percentage of total languages for each language class.
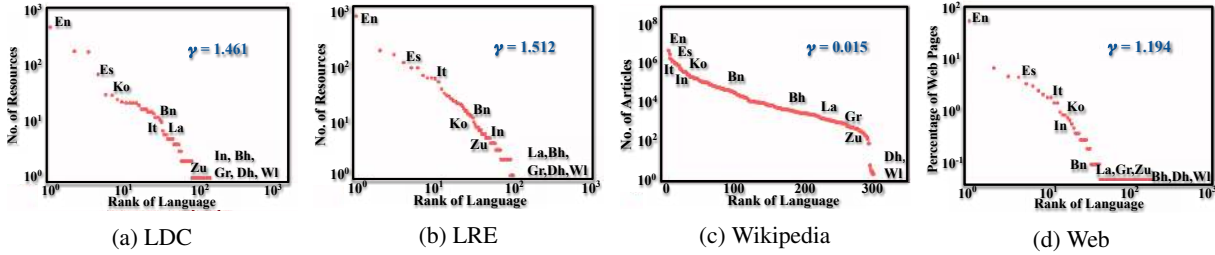

(a) LDC   (b) LRE   (c) Wikipedia   (d) Web

Figure 3: Plots of different available resources for different languages. Languages to the far right do not have a representation in the resource category. Languages annotated are: Class 0-Dahalo (Dh), Wallisian(Wl); Class 1-Bhojpuri (Bh), Greenlandic (Gr); Class 2-Lao (La), Zulu (Zu); Class 3- Bengali (Bn), Indonesian (In); Class 4- Korean (Ko), Italian (It); Class 5- English (En), Spanish (Es).

languages, reaping benefit from each state-of-the-art NLP breakthrough.

Some more information about the taxonomy is shown in Table 1. We also take 10 languages, and annotate their positions in Figure 3.

2.4 Findings

On your marks As can be seen in Figure 3, the Winners take pole position in all rankings, and Class 0 languages remain 'out of the race' with no representation in any resource. The Wikipedia distribution seems to be more fair for classes 1, 2, and 3 when compared to classes 4 and 5, whereas the Web distribution has a clear disparity.

Talk ain't cheap Looking at Table 1, we see that Class 0 contains the largest section of languages and represents 15% of all speakers across classes. Although there is a large chunk of speakers which converse with Class 5 languages, the lack of technological inclusion for different languages could draw native speakers away from Class 0 languages and towards Class 5, exacerbating the disparity.

3 Typology

Linguistic typology is a field which involves the classification of languages based on their structural and semantic properties. Large-scale efforts have led to the creation of a database of typological

features (Dryer and Haspelmath, 2013). Such documentation becomes important as there are barely any other classifications of similar scale. In the context of NLP research, there has been work indicating the effectiveness of injecting typological information to guide the design of models (Ponti et al., 2019). Also, transfer learning of resource-rich to resource-poor languages have been shown to work better if the respective languages contain similar typological features (Pires et al., 2019). We look at how skewed language resource availability leads to an under-representation of certain typological features, which may in turn cause zero-shot inference models to fail on NLP tasks for certain languages.

We look at the WALS data (Dryer and Haspelmath, 2013), which contains typological features for 2679 languages. There are a total of 192 typological features, with an average of 5.93 categories per feature. We take the languages in classes 0, 1, 2, all of which have limited or no data resources as compared to 3, 4, 5 and look at how many categories, across all features, exist in classes 0, 1, 2 but not 3, 4, 5. This comes to a total of 549 out of 1139 unique categories, with an average of 2.86 categories per feature being ignored. Typological features with the most and least 'ignored' categories are shown in Table 2.

To get an idea of what these typological 'exclu-

Feature	#Cat	#Lang		Feature	#Cat	#Lang
144E	23	38		83A	0	1321
144M	23	45		82A	0	1302
144F	22	48		97A	0	1146
144O	21	30		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.

Language	Class	#Speakers	'Ignored'	Error
Amharic	2	22M	9	60.71
Breton	1	210k	7	83.50
Swahili	2	18M	8	45.64
Kabyle	1	5.6M	8	39.10

Table 3: Relevant examples of typologically 'excluded' languages. The error rate is that of English → Language from Artetxe and Schwenk (2019).

sions' mean in the context of modern multilingual methods, we look at the specific languages which contain these 'excluded' categories in the respective features, and compare their performances in similarity search, from the results of Artetxe and Schwenk (2019). Table 3 shows some examples of how 'ignored' features have been difficult to deal with even when jointly training of all languages.

3.1 Findings

Far-reaching repercussions The most 'ignored' feature in Table 2, 144E (Multiple Negative Constructions in SVO Languages), is a rare feature, existing in only 38 languages over the world. These languages, however, are from various regions (e.g. `Wolof`, `Icelandic`, and `Kilivila`). Language tools in all these areas can be adversely affected without sufficient typological representation. On the other hand, common features such as 83A (Order of Object and Verb) are well represented with definite feature values for 1321 languages, ranging from `English` to `Mundari`.

Does it run in the family? `Amharic`, in Table 3, which among the Semitic family of languages, is the second most spoken language after `Arabic` (which has ⁓300M speakers). However, it has 9 'ignored' typological features, whereas `Arabic` has none. This reflects in the error rate of `English` to `Amharic` (60.71), which is significantly worse compared to 7.8 for `English` to `Arabic`.

4 Conference-Language Inclusion

NLP conferences have a huge impact on how language resources and technologies are constructed. Exciting research in venues such as **ACL**, **EMNLP**, **LREC** have the ability to turn heads in both industry and government and have the potential to attract funds to a particular technology. Has the usage of a small set of resource-rich languages in such conferences led to a disparity, pushing the less represented to the bottom of the ladder in terms of research? We analyze the involvement of various languages in NLP research conferences over the years.

4.1 Dataset

The ACL Anthology Corpus (ACL-ARC) (Bird et al., 2008) is the most extensively used dataset for analyzing trends in NLP research. This dataset contains PDFs, and parsed XMLs of Anthology papers. However, the latest versioned copy of ACL-ARC is till 2015 which makes it insufficient for analyzing trends in the most recent years. Moreover, paper data for non-ACL conferences such as **LREC**, **COLING** are absent from this dataset. In order to create a consistent data model, we augment this dataset by using Semantic Scholar's API and scraping ACL Anthology itself. Thus, we gather a consolidated dataset for 11 conferences which are relevant in judging global trends in NLP research. These include **ACL**, **NAACL**, **EMNLP**, **EACL**, **COLING**, **LREC**, **CONLL**, Workshops (**WS**) (all since 1990), **SEMEVAL**, **TACL** and **CL** Journals. We have attached the statistics of the dataset in Appendix A.

4.2 Analysis

4.2.1 Language Occurrence Entropy

The primary step of measuring the language diversity and inclusion of a conference and their progress is to measure the usage of language in that conference over multiple iterations. One of the ways to do it is by using frequency-based techniques where we can measure the occurrence of languages in that iteration. However, it is not a unified measure which represents the nature of language distribution with a single number. To this end, we use entropy as our metric to measure language inclusivity of each conference. It efficiently captures the skew in the distribution of languages,
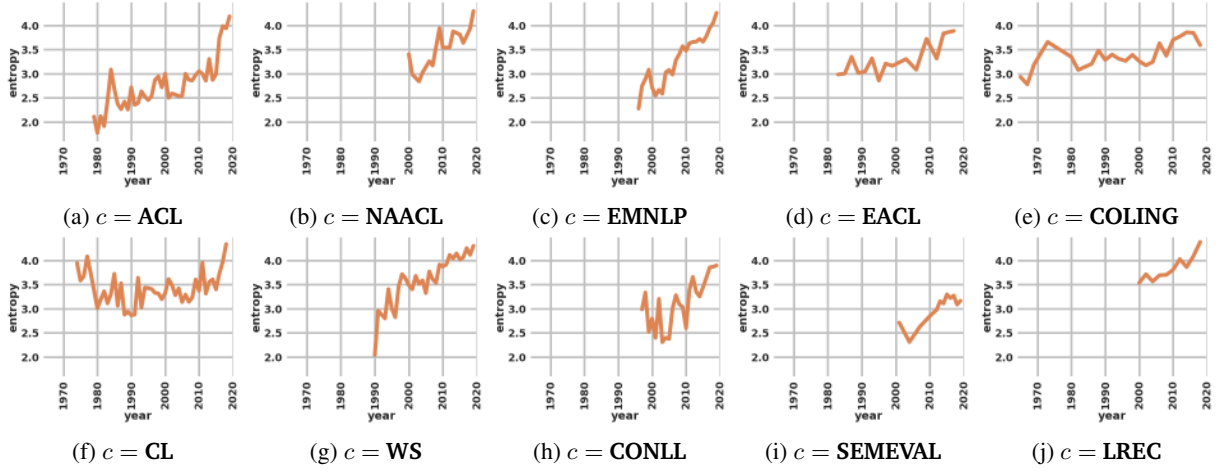
(a) c = **ACL** (b) c = **NAACL** (c) c = **EMNLP** (d) c = **EACL** (e) c = **COLING**

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

Figure 4: Language occurrence entropy over the years for different conferences ($\{S\}^{c,y}$).

thereby making the disparity in language usage more clearer. The language occurrence entropy is calculated as follows:

For a conference c held in year y having P papers, there exists a binary matrix $\{M_{P \times L}\}^{c,y}$ where M_{ij} is 1 if i^{th} paper ($\in P$) mentions the jth language ($\in L$). Then the entropy $\{S\}^{c,y}$ is:

$$\{S_j\}^{c,y} = \frac{1}{P} \sum_{i=1}^{P} \{M_{ij}\}^{c,y}$$

$$\{S'_j\}^{c,y} = \frac{\{S_j\}^{c,y}}{\sum_{j=1}^{L} \{S_j\}^{c,y}} \quad (1)$$

$$\{S\}^{c,y} = -\sum_{j=1}^{L} \{S'_j\}^{c,y} log_e \{S'_j\}^{c,y}$$

where $\{S_j\}^{c,y}$ is a array of length L accounting for number of papers in a specific language, $\{S'_j\}^{c,y}$ is normalization done in order to get probability distribution for calculating entropy. In short, the higher the entropy, the more spread out is the distribution over the languages. The more peaked or skewed the distribution is, the lower is the entropy. In Figure 4, we can observe the entropy S plotted for each c as a function of y.

4.2.2 Class-wise Mean Reciprocal Rank

To quantify the extent of inclusion of language classes from our taxonomy in different conferences, we employ class-wise Mean Reciprocal Rank (MRR) as a metric. This helps in determining the standing of each class in a conference. If the rank of the language (rank_i) is ordered by the frequency of being mentioned in papers of a particular conference, and Q is the total number of queries

aka number of languages in each class, then:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \quad (2)$$

Table 4 shows inverse mean reciprocal ranks of each category for a conference. The smaller the inverse MRR value, the more inclusive that conference is to that language 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

Table 4: Class-wise (1/MRR) for each conference.

4.3 Findings

All-Inclusive Looking at the combined trends, both the entropy plots and the MRR figures suggest that **LREC** and **WS** have been the most inclusive across all categories and have been continuing to do so over the years.

A ray of hope With regards to the proceedings of **ACL**, **EMNLP**, **NAACL**, **LREC**, we note a marked spike in entropy in the 2010s, which is absent in

other conferences. This might be due to the increased buzz surrounding cross-lingual techniques.

The later the merrier An interesting point to note is that conferences which started later have taken lessons from past in matters of language inclusion. While the earlier established conferences have continued to maintain interest in a particular underlying theme of research which may or may not favour multilingual systems. This can be observed in : **COLING**, **ACL**, **EACL**, **EMNLP** (order of their start dates).

Falling off the radar The taxonomical hierarchy is fairly evident when looking at the MRR table (Table 4) with class 5 coming within rank 2/3 and class 0 being 'left-behind' with average ranks ranging from 600 to 1000. While the dip in ranks is more forgiving for conferences such as **LREC**, **WS**, it is more stark in **CONLL**, **TACL**, **SEMEVAL**.

5 Entity Embedding Analysis

The measures discussed in the previous section signal at variance in acceptance of different languages at different NLP venues across time. However, there are usually multiple subtle factors which vanilla statistics fail to capture. Embeddings, on the other hand, have been found extensively useful in NLP tasks as they are able to learn relevant signals directly from the data and uncover these rather complex nuances. To this end, we propose a novel approach to jointly learn the representations of conferences, authors and languages, which we collectively term as entities. The proposed embedding method allows us to project these entities in the same space enabling us to effectively reveal patterns revolving around them.

5.1 Model

We define the following model to jointly learn the embeddings of entities such that entities which have similar contextual distributions should co-occur together. For example, for an author A, who works more extensively on language L_i than L_j and publishes more at conference C_m than at conference C_n, the embeddings of A would be closer L_i than L_j and C_m than C_n.

Given an entity and a paper associated with the entity, the learning task of the model is to predict K randomly sampled words from the title and the abstract of the paper. We only select the title and abstract as compared to the entire paper text as
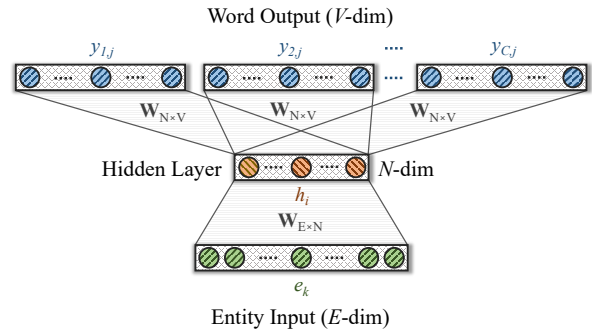

Figure 5: Model architecture to learn entity embeddings. $W_{E \times N}$ is the weight matrix from input layer (entity layer) to the hidden layer, and $W_{N \times V}$ is the weight matrix for the hidden layer to output layer computation. At the end of training, $W_{E \times N}$ is the matrix containing embeddings of entities and $W_{N \times V}$ is the matrix containing the embeddings of words.

they provide a concise signal with reduced noise. This model draws parallels to the Skipgram model of Word2Vec (Mikolov et al., 2013), where given an input word in Skipgram model, the task is to predict the context around the word. The input entity and K randomly sampled words in our case correspond to the input word and context in the Skipgram model. The goal of the model is to maximize probability of predicting the random K words, given the entity id as the input:

$$\frac{1}{M} \frac{1}{K} \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{i=1}^{I} p(w_k | E_{<i, P_j>}) \qquad (3)$$

where $E_{<i,P_j>}$ is the entity E_i which is associated with the $P_j{}^{th}$ paper and p is the probability of predicting the word w_i out of the K words sampled from the paper and M is the total number of papers in the dataset. To optimize for the above distribution, we define the typical SGD based learning strategy similar to Word2Vec(Mikolov et al., 2013).

Figure 5 shows an outline of the model. The entity input layer has dimension equal to the total number of entities in the dataset (E). Hidden layer size is set to the desired embedding dimension (N). The output layer predicts words for the input entity and is of the same size as the vocabulary (V). The entities we learn are: (1) authors of the paper, (2) languages mentioned in the paper, (3) conference where the paper was accepted (e.g. **ACL**), and (4) the conference iteration (e.g. ACL'19). We describe the model detail and hyperparameter tuning in Appendix A.

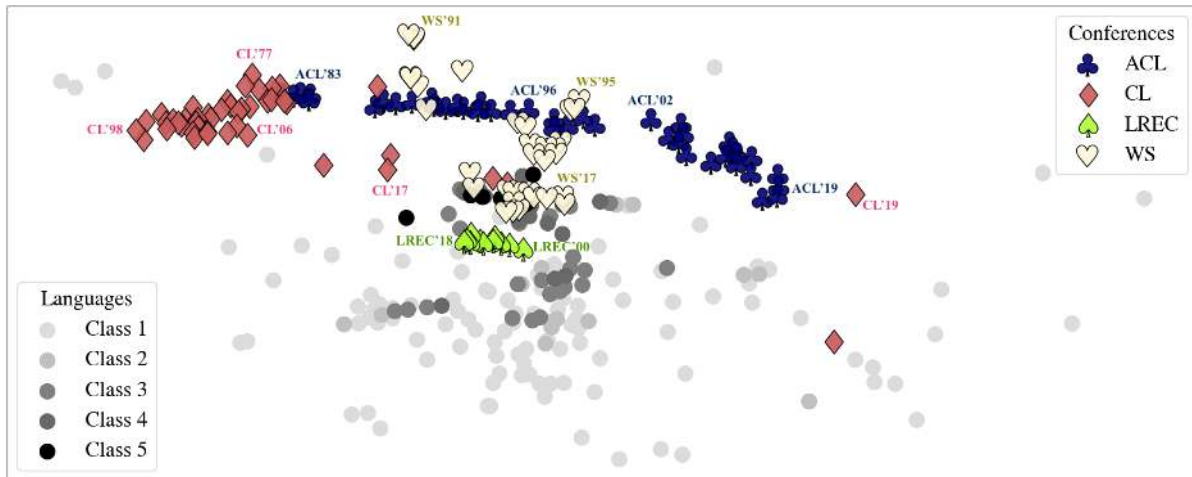Figure 6: t-SNE visualization of the learnt conference and language embeddings.

Class	MRR(5)	MRR(10)	MRR(15)	MRR(20)
0	0.72281	0.69146	0.63852	0.57441
1	0.57210	0.52585	0.45354	0.40904
2	0.47039	0.45265	0.41521	0.38157
3	0.59838	0.52670	0.45131	0.42899
4	0.56016	0.47795	0.51199	0.50681
5	0.56548	0.51471	0.54326	0.47619

Table 5: Language-Author-Language MRR on Taxonomy Classes. MRR(K) considers the closest K authors.

5.2 Analysis

In order to better understand how languages are represented at different venues, we visualize the distribution of entity embeddings by projecting the generated embeddings into 2 dimensions using t-SNE (Maaten and Hinton, 2008) (as shown in Figure 6). For clarity, we only plot **ACL**, **LREC**, **WS** and **CL** among the conferences, and all languages from the taxonomy, except those in Class 0. We omit plotting Class 0 languages as their projections are noisy and scattered due to their infrequent occurrence in papers.

To understand the research contributions of individual authors or communities towards research in respective language classes, we leverage the distribution between author and language entities by computing a variation of the Mean Reciprocal Rank (MRR). We consider a language L, and take the K closest authors to L using cosine distance, and then take the closest M languages to each author. If L is present in the closest languages of an author, then we take the rank of L in that list, inverse it, and average it for the K authors. To compute this

metric for a class of languages from the taxonomy, we take the mean of the MRR for all languages in that class. We fix M to be 20, so as to understand the impact of the community when the number of languages remains unchanged. Table 5 shows the MRR of various class of languages. A higher value of this measure indicates a more focused community working on that particular language, rather than a diverse range of authors.

5.3 Findings

Time waits for no conference We can see a left to right trend in Figure 6 with **ACL** in 1983 in the left, and subsequent iterations laid out as we go right. We observe the same trend for **EACL**, **NAACL**, **EMNLP**, **CONLL**, **TACL**, and **COLING**. We can say that the axis represents the progression of time to a certain extent. Alternatively, it may even represent a shift in the focus of NLP research, moving from theoretical research focused on grammar and formalisms on the left to a data-driven, more ML-oriented approach on the right. This can be observed as most of the **CL** embeddings are positioned on the left given their theoretical research focus.

Long distance relationships? From Figure 6, we can note that the less-resourced language classes are farther away from the trend-line of **ACL** than the more resourced ones, with class 5 being closest, and class 1 being farthest. The visualization illustrates that languages are spreading out radially downwards from the **ACL** trendline with popular classes of taxonomy like class 5 and class 4 being closer while others spreading out farther. Again, as

previous analyses have shown us, **LREC** and **WS** embeddings are closer to the language embeddings as compared to the other conferences as shown in Figure 6. In fact, **LREC** cluster is right in the middle of language clusters and so is the major part of the **WS** cluster, especially in recent iterations.

Not all heroes wear capes Table 5 shows the MRR for each class of languages in the taxonomy. From Table 5, it can be seen that class 0 has the highest MRR across different **K** values. This shows that perhaps low resource languages have some research groups solely focused on the challenges related to them. There is a decreasing trend of MRR from class 0 to class 5, except for class 2, thereby indicating that more popular languages are addressed by more authors. We also observe that even though `Japanese`, `Mandarin`, `Turkish` and `Hindi` (MRR(10) > 0.75) are part of class 5 and class 4, their MRR is higher even compared to low resource languages in another classes, indicating that these languages have focused research communities working on them. On the other end of the spectrum, we observe a lot of low resource languages like `Burmese` (MRR(10) = 0.02), `Javanese` (MRR(10) = 0.23) and `Igbo` (MRR(10) = 0.13) which have millions of speakers but significantly low MRR values, potentially indicating that not a lot of attention is being given to them in the research community.

6 Conclusion

We set out to answer some critical questions about the state of language resource availability and research. We do so by conducting a series of quantitative analyses through the lens of a defined taxonomy. As a result, we uncover a set of interesting insights and also yield consistent findings about language disparity:

— The taxonomical hierarchy is repeatedly evident from individual resource availabilities (LDC, LRE, Wikipedia, Web), entropy calculations for conferences, and the embeddings analysis.

— **LREC** and Workshops(**WS**) have been more inclusive across different classes of languages, seen through the inverse MRR statistics, entropy plots and the embeddings projection.

— There are typological features (such as 144E), existing in languages over spread out regions, represented in many resource-poor languages but not sufficiently in resource-rich languages. This could potentially reduce the performance of language tools relying on transfer learning.

— Newer conferences have been more language-inclusive, whereas older ones have maintained interests in certain themes of research which don't necessarily favour multilingual systems.

— There is a possible indication of a time progression or even a technological shift in NLP, which can be visualized in the embeddings projection.

— There is hope for low-resource languages, with MRR figures indicating that there are focused communities working on these languages and publishing works on them, but there are still plenty of languages, such as `Javanese` and `Igbo`, which do not have any such support.

We believe these findings will play a strong role in making the community aware of the gap that needs to be filled before we can truly claim state-of-the-art technologies to be language agnostic. Pertinent questions should be posed to authors of future publications about whether their proposed language technologies extend to other languages.

There are ways to improve the inclusivity of ACL conferences. Special tracks could be initiated for low-resource, language-specific tasks, although we believe that in doing so, we risk further marginalization of those languages. Instead, a way to promote change could be the addition of D&I (Diversity and Inclusion) clauses involving language-related questions in the submission and reviewer forms: *Do your methods and experiments apply (or scale) to a range of languages? Are your findings and contributions contributing to the inclusivity of various languages?*

Finally, in case you're still itching to know, Language **X** is `Dutch`, and **Y** is `Somali`.

7 Acknowledgements

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technolo-*

gies, Volume 1 (Long and Short Papers), pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Emily M. Bender. 2011. On achieving and evaluating language-independence in NLP. In *Vol 6: Interaction of Linguistics and Computational Linguistics*.

Steven Bird, Robert Dale, Bonnie J Dorr, Bryan Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R Radev, and Yee Fan Tan. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. *LREC'08*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for pan-lingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Gary Simons and Steven Bird. 2003. The open language archives community: An infrastructure for distributed archiving of language resources. *Computing Research Repository - CORR*, 18:117–128.

A Appendix

A.1 Embedding Visualization

We have compiled a visualization of the embedding space of conferences and languages which can be run on a browser. This is available as an interactive visualization on https://microsoft.github.io/linguisticdiversity, and can be used to play around with different combinations to see how NLP research has progressed over the years in terms of language inclusion. The legends are self-explanatory and are clickable to add or remove those points. The numbers in the legend represent the respective classes.

A.2 ACL Anthology Dataset Statistics

We have accounted for all the papers which have appeared in the main track proceedings of the conference. This includes all the long and short papers and excludes System Demonstrations, Tutorial Abstracts, Student Research Workshops, Special Issues, and other such tracks out of the scope of measuring language usage trends in general NLP research. We are in the process of releasing the dataset along with the documentation.

Conf / Class	#Papers	#Body	#NoProc	% Missing
LREC	5835	15	6	0.1%
WS	17844	337	332	1.86%
CONLL	1035	0	0	0.0%
EACL	1165	4	1	0.09%
ACL	5776	46	29	0.5%
TACL	280	7	0	0.0%
CL	2025	88	0	0.0%
NAACL	2188	2	1	0.05%
COLING	4233	5	2	0.05%
EMNLP	3865	16	16	0.41%

Table 6: Dataset Statistics.

A.3 Hyperparameter Tuning

Our model has same hyperparameters as that of Word2Vec. To determine the optimal hyperparameters for the model, we take the entire dataset and split it into a 80-20 ratio, and given the embedding of a paper, the task is to predict the year in which the paper is published. Given this vector for a paper, we use a linear regression model such that given this vector, the model is supposed to predict the year in which the paper was published. We measured both R^2 measure of variance in regression and mean absolute error (MAE). R^2 is usually in the range of 0 to 1.00 (or 0 to 100%) where 1.00 is considered to be the best. MAE has no upper bound but the smaller it is the better, and 0 is its ideal value. We observed that our model does not show significant difference across any hyperparaeters except for the size of embeddings. The best dimension size for our embeddings is 75, and, we observed the corresponding R^2 value of 0.6 and an MAE value of 4.04.

A.4 Cosine distance between conferences and languages

From Figure 6, we can see that languages are somewhat below the conferences are closer to some conferences while distant from others. To quantify this analysis, we compute the cosine distance between the conference vector and the mean of the vector each category of the taxonomy. Table 7 shows the cosine distance between the conferences and the each category of languages and we see a very similar trend that while **ACL** is an at average distance of 0.291 from category 5 languages, its almost more than double far away from category 2. There is also a very steep rise in distance of the **ACL** vector from category 4 to category 3. In fact, similar trends are visible for other **ACL** related conferences including **EACL**, **NAACL**, **EMNLP** and **TACL**. We can also see that in Table 7, **WS** and **LREC** are closest from category 2 to category 5 whereas almost all conferences are somewhat at the same distance from category, except the CL journal. The trend for category 0 languages seems somewhat different than the usual trend is this table, probably because of the large number of languages in this category as well as the sparsity in papers.

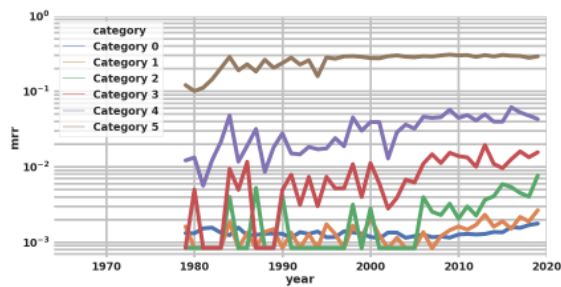| Conf / Class	0	1	2	3	4	5
LREC	0.51	0.51	0.52	0.42	0.36	0.32
WS	0.50	0.55	0.53	0.40	0.28	0.21
CONLL	0.54	0.60	0.63	0.49	0.40	0.46
EACL	0.53	0.55	0.59	0.45	0.34	0.32
ACL	0.48	0.51	0.60	0.42	0.34	0.29
TACL	0.52	0.56	0.66	0.48	0.38	0.47
CL	0.67	0.78	0.80	0.75	0.65	0.59
NAACL	0.48	0.52	0.59	0.47	0.39	0.33
COLING	0.48	0.53	0.55	0.46	0.37	0.30
EMNLP	0.57	0.59	0.66	0.51	0.46	0.45

Table 7: Cosine Distance between conference vectors and mean class vectors of languages.
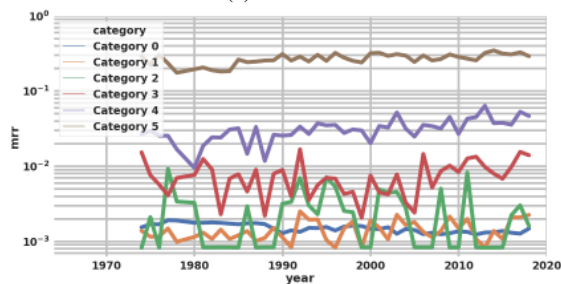
A.5 Taxonomy classification

We release our full language taxonomy classification on the website: https://microsoft.github.io/linguisticdiversity.

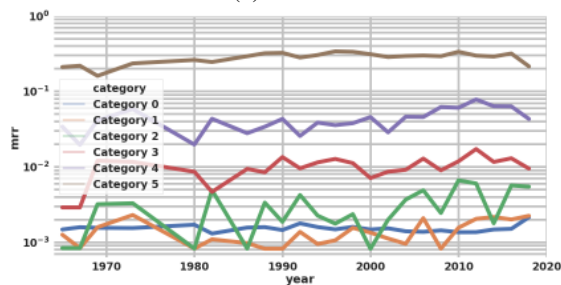## A.6 Class-wise log(MRR) over the years per conference

We plot MRR on a log scale for each conference to measure the progress of inclusion of the defined taxonomy classes over the years. It is very interesting to note how **LREC** has very smooth forward progression.
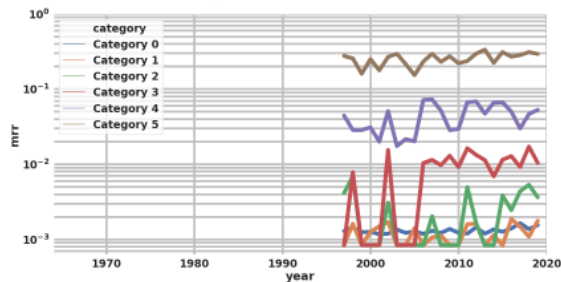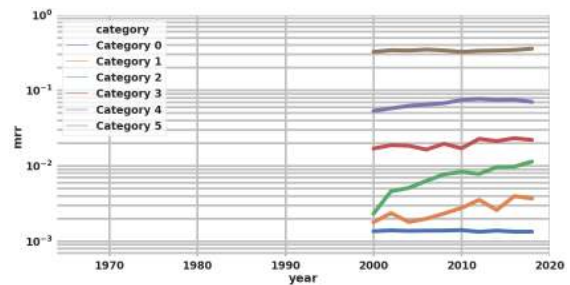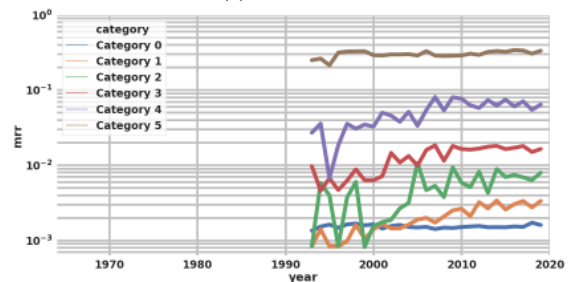

(a) $c = $ **LREC**

(a) $c = $ **ACL**

(b) $c = $ **WS**

(b) $c = $ **CL**

(c) $c = $ **EMNLP**

(c) $c = $ **COLING**

(d) $c = $ **NAACL**

(d) $c = $ **CONLL**

(e) $c = $ **SEMEVAL**

6293