Some Languages are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World

Anonymous ACL submission

Abstract

Linguistic disparity in the NLP world is a problem that has been widely acknowledged recently. However, different facets of this problem, or the reasons behind this disparity are seldom discussed within the NLP community. This paper provides a comprehensive analysis of the disparity that exists within the languages of the world. Using an existing language categorisation based on speaker population and vitality, we analyse the distribution of language data resources, amount of NLP/CL research, inclusion in multilingual web-based platforms, and the inclusion in pre-trained multilingual models. We show that many languages do not get covered in these resources or platforms, and even within the languages belonging to the same language group, there is wide disparity. We analyse the impact of family, geographical location, and the speaker population of languages, provide possible reasons for this disparity, and argue that a solution to this problem should be orchestrated by a wide alliance of stakeholders, of which ACL, as an association should be a key partner.

1 Introduction

Even after more than fifty years of the inception of the fields of Computational Linguistics (CL) and Natural Language Processing (NLP), and ACL turning 60 in 2021, we still observe a significant bias favouring the so-called *high-resource* languages in the field. Conversely, this means that the majority of the 6500+ languages in the world, which have been classified as *low-resource*, have received limited to no attention from the CL and NLP community. This resource poverty is not merely an academic or theoretical issue. It impacts the lives and the well-being of people concerned in a very present and practical manner, and deprives the populations that use the low-resource languages from reaping the benefits that NLP brings in areas such as healthcare (Perez-Rosas et al., 2020), disaster response (Ray Chowdhury et al., 2019), and education (Taghipour and Ng, 2016).

There is newfound hope for emergence from obscurity, as this digital divide between high-resource and low-resource languages (LRLs)[1] has been brought into the spotlight by many scholars in the field (Bender, 2019; Cains, 2019; Joshi et al., 2020; Anastasopoulos et al., 2020). Consequently, there have been efforts to build data sets covering low-resource languages (Conneau et al., 2020; Ebrahimi et al., 2021), benchmarks (Hu et al., 2020), and techniques that favor low-resource languages (Schwartz et al., 2019); all of which, are very promising developments. However, everyone would agree, that there is much more to be done. In doing so, having a clear idea of the disparity that exists between the languages in the world with respect to resource availability and other socio-economic conditions is helpful.

The '*resourcefulness*' of a language can be analysed with respect to different socio-linguistic aspects. Besacier et al. (2014) identify these factors as: 1) The existence of a unique writing system, 2) The amount of presence on the World Wide Web, 3) The availability of linguistic expertise, and/or 4) The availability of electronic resources such as corpora (monolingual and parallel), and vocabulary lists. Singh (2008), on the other-hand, identifies these factors as: 1) The amount of linguistic study, 2) The availability of language resources, 3) The level of computerisation, 4) The availability of language processing tools, and 5) other privileges such as finance and human resource.

The paper title is inspired by the quote "*All animals are equal, but some animals are more equal than others*" by Orwell (1945) which satirically alludes to disparities that exist in places which, ostensibly are supposed to be homogeneous. In this paper, we discuss how the same phenomenon is observed in the broadly used language categorisation systems.

[1] An LRL is also known as under resourced, low-density, resource-poor, low data, or less-resourced language (Besacier et al., 2014)

As a general practice, NLP researchers have mainly considered the availability of electronic data resources as the main descriptor of '*resourcefulness*' of languages. For example, Joshi et al. (2020) considered the availability of annotated and raw corpora, while the later study, Hedderich et al. (2021), considered the availability of auxiliary resources such as lexicons as an additional criterion. Joshi et al. (2020) used their criterion to categorise 2485 languages into six groups, based on the availability of unannotated data (number of wikipedia pages), and the number of annotated data sets available in the LDC[2] and ELRA[3] data repositories. Figure 5a shows a recreation of these language categories[4].

According to this categorisation, an astounding 2191 languages fall into *Category 0*- those that have exceptionally low amount of resources. This paints a very grim picture of the linguistic diversity and inclusion in the NLP world. This is not surprising though; this categorisation is based on wikipedia data as the source of monolingual data, and wikipedia has articles only in 325 languages including 7 constructed languages such as *Esperanto*[5]. Therefore, inherently, all the other languages automatically get labeled as extremely low resourced.

However, Joshi et al. (2020)'s analysis focused only on data availability as well as the amount of language-related research in ACL Anthology. They did not consider other aspects of resourcefulness, such as the inclusion of a language in multilingual web-based platforms such as Facebook, or the inclusion in pre-trained multilingual neural models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2019). Moreover, this language categorisation does not shed light on how this language disparity could be explained with respect to other socio-economic-linguistic factors such as language family, geographical location or the speaker population.

This paper intends to take Joshi et al. (2020)'s analysis a step further, and provides a deeper analysis into the less-known facts of the well-known problem of linguistic disparity in the world. We start with an existing language categorisation based on speaker population and vitality (Ethnologue[6]) (Eberhard and Fennig, 2021), and analyse the distribution of language data resources, amount of NLP/CL research, inclusion in multilingual web-based platforms, and the inclusion in pre-trained multilingual models. We show that many languages are neglected with respect to all these criteria, and even within the languages belonging to the same language group, there is wide disparity. We analyse this disparity with respect to the family, geographical location, as well as the speaker population of languages. We also provide possible reasons for this disparity, and argue that most these reasons are beyond the control of ACL, as an organization. Based on this argument, we provide a preliminary set of recommendations that may be implemented by various stakeholders, in reducing this disparity across languages.

2 The 12 Kinds of Languages

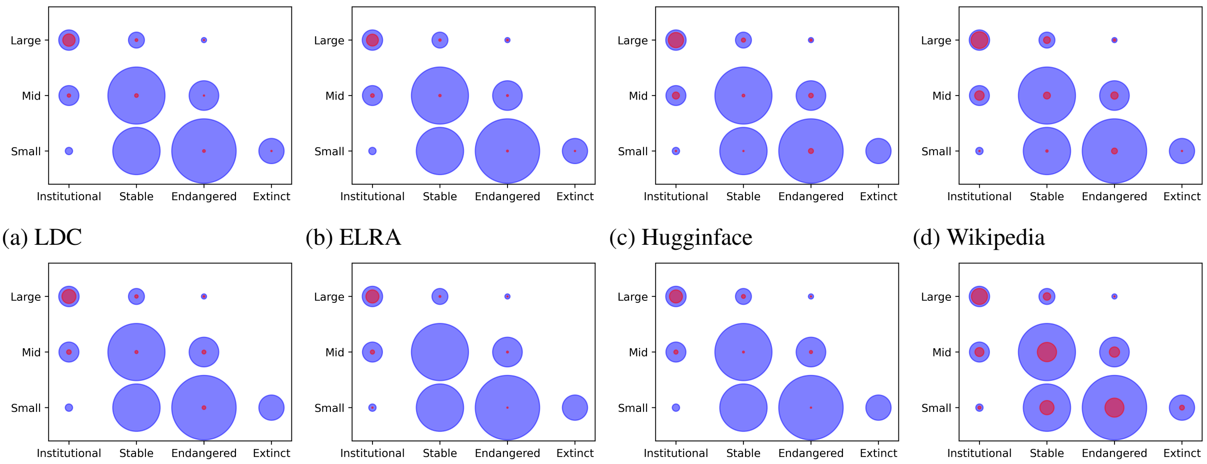
Ethnologue is an annual publication that provides statistics and other information of the living languages in the world. It has 7139 language entries, including dialects. We could identify 6420 unique languages by considering alternate names, dialects, and minor schisms to map to their most prominent entry. Languages in Ethnologue are categorised into 12 classes, considering two variables: *Population* and *Vitality*. Firstly, *Population* is "the estimated number of all users (including both first and second language speakers) in terms of three levels", the aforementioned three levels being: *large*, *Mid-sized*, and *small* (Eberhard and Fennig, 2021). On the other hand, *Vitality* is categorised into four distinct classes: *institutional*, *stable*, *endangered*, and *extinct*, according to the Expanded Graded Intergenerational Disruption Scale (EGIDS) grid (Lewis and Simons, 2010).

Figure 1 shows the languages categorised in a 12-point grid, according to vitality and number of speaker population. The size of the blue circles correspond to the number of languages in one category. According to this figure, a large number of languages are endangered with small speaker populations, or stable but with mid or small number of speaker populations.

3 Resource & Tool Support Distribution

We analyse how languages in the different Ethnologue categories are being treated with respect to data (annotated and un-annotated), inclusion in

[2] https://catalog.ldc.upenn.edu/
[3] http://catalog.elra.info/en-us/
[4] Refer Appendix A for class descriptions.
[5] https://bit.ly/WikiList
[6] https://bit.ly/3kJircB



(a) LDC (b) ELRA (c) Hugginface (d) Wikipedia

(e) Facebook (f) Google Keyboard (g) XLMR+mBERT (h) ACL Anthology

Figure 1: The 12 Ethnologue language classes where the size of each blue circle corresponds to the number of languages in that category and the size of each red circle corresponds to the coverage of that class in the relevant resource.

multilingual web-based platforms, and inclusion in pre-trained multilingual models. Ideally, this analysis should have been carried out for the availability of language technologies as well, as done by META-NET (2020). However, this would be a daunting task, and is out of the scope of this research. With that restriction, we discuss the available resources and tools in Sections 3.1 through 3.4, which is then followed by an aggregated analysis in Section 3.5.

3.1 Un-annotated Data Availability

There are two possible sources to be used here: wikipedia data and common crawl. However, the latter covers only 160 languages[7], compared to the 318 languages in wikipedia (excluding the 7 constructed languages). Thus, we focus our analysis on wikipedia data as the main source of un-annotated data. The common crawl data analysis has been briefly reported in Appendix B.

3.2 Annotated Data Availability

In addition to *LDC* and *ELRA*, we included the *Huggingface* data sets[8] as well. Despite being relatively new and with less standardization, this repository has data in comparable amounts to the other repositories. Another possible repository is the Kaggle data sets. However, it does not have a proper way of filtering out data sets with respect to the language.

[7] https://bit.ly/3F9iK87

[8] https://huggingface.co/docs/datasets/

3.3 Multilingual Web-based Platforms

Facebook, Google, and Twitter are examples for widely used multilingual web-based platforms. The availability of a platform interface in the native language of a user encourages them to use that platform to express themselves in the same, which of course results in more web content. Conversely, the languages that are not supported will be less and less used (Bird, 2020a). For our analysis, we considered the languages covered by Google type (Google keyboard) and the languages supported by Facebook, as these have the widest language coverage.

3.4 Pre-trained Multilingual Model Coverage

Out of the many competing models, the ones with the widest coverage and popularity are *mBERT* and *XLM-R*. These models have been quite effective in zero-shot and few-shot NLP tasks (Hu et al., 2020; Lauscher et al., 2020). They perform better for languages that are included in the pre-training stage, compared to those that are not (Ebrahimi and Kann, 2021). These models have also shown to outperform their monolingual counterparts for low resource languages (Wu and Dredze, 2020). Considering the above facts, and the fact that it is computationally expensive to train such multilingual models, languages that are already included in such multilingual models would have an edge over those that are not.

3.5 Aggregated Analysis

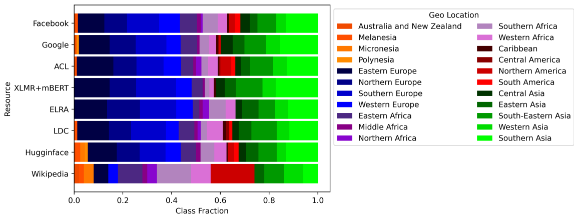
Figure 1 as well as Tables 1 and 2 show how the languages from different categories have been included in different types of resources and web-based platforms. It is evident that language resource creation and technology availability has been mostly centred around institutional languages with high speaker populations, while small and endangered languages have mostly been ignored.

Interestingly, Table 1 shows that, wikipedia does have some coverage for all the categories, including extinct languages, which we believe may be partly due to research efforts[9] (Paranjape et al., 2016). However, LDC, ELRA, and Hugginface have comparatively less coverage. This is to be expected, as annotated data creation takes a different level of expertise and more time (and money) compared to writing wikipedia articles.

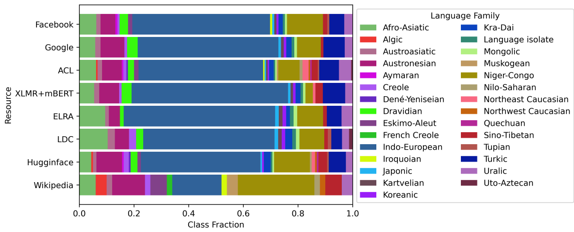
According to Table 2, we observe that Facebook and Google platforms mainly cover institutional languages, with a negligible representation of other languages, which would have been motivated by the speaker population. The same is observed for the coverage in the multilingual pre-trained models *mBERT* and *XLM-R*, released by Google and Facebook, respectively. Given that such multilingual models suffer from 'curse of multilinguality' (Lauscher et al., 2020), the selection of languages to be included in the models would have had similar motivations.

Figure 2a and 2b visualize the coverage of these different platforms and resources with respect to the geographical location and family of a language. We can see that all these criteria are biased towards a certain set of language families and geographical locations, namely the *Indo-European* family and the *Europe* region. This is not surprising, given the emphasis placed on language resource development by the European region (META-NET, 2020). This also explains observation made by Hu et al. (2020), where multilingual pre-trained models perform better for Indo-European languages. Interestingly, wikipedia has been more democratic compared to other resources[10]. LDC and ELRA data sources are more concentrated in the Europe area. In contrast, Hugginface is more distributed.

However, Figure 1 only can be misleading, as the amount of data varies across different languages even within the same category. In order to get a better view of the amount of data resources, we derived the box plots shown in Figure 3 which uncovered a noticeable disparity between different language categories. Aside from the inter-class disparities, 3d especially shows a noticeable variance in wikipedia data availability within the *Large-Institutional* class. In order to understand this variance, we plotted the graph shown in Figure 4. As can be seen, the number of wikipedia articles available has a *strong correlation* (0.518789) to the population that speaks the language[11]. A surprising observation is that about 70 languages belonging to *Large-Institutional* class do not have a presence in wikipedia. We looked at these languages more closely - a vast majority of these languages are in the African region.



(a) By Geological Location



(b) By Language Families

Figure 2: The distribution of Resources

4 Revisiting Data Availability-based Language Categorisation

As mentioned earlier, NLP researchers have considered the availability of language data as the criterion to categorise languages. In order to analyse the robustness of this categorisation, we recreated Joshi et al. (2020)'s language category plot. In Figure 5, we plot the availability of annotated data in LDC and ELRA against the unannotated wiki data in 5a[12]. In 5b we plot the same graph

[9] https://stanford.io/3mXQK0Z

[10] More analysis in Appendix D

[11] The coordinates are derived from the L1 and L2 speaker population reported in Wikipedia and the colour of each data point is taken according to the class in Ethnologue. Therefore, data points that violate the colour boundaries along the X-axis are instances where Wikipedia and Ethnologue do not agree.

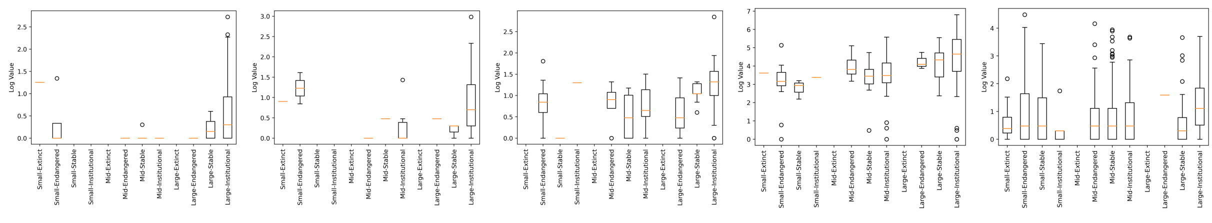
[12] Different to (Joshi et al., 2020), we considered the number of *wikipedia articles*, as considering *pages* could be mislead-

Class	LDC		ELRA		Hugginface		Wikipedia		ACL	
	Count	%	Count	%	Count	%	Count	%	Count	%
Small-Extinct	1	0.30	1	0.30	0	0.00	1	0.30	12	3.61
Small-Endangered	4	0.19	2	0.09	13	0.60	18	0.83	188	8.70
Small-Stable	0	0.00	0	0.00	1	0.09	3	0.26	105	8.99
Small-Institutional	0	0.00	0	0.00	1	3.57	1	3.57	5	17.86
Mid-Endangered	1	0.22	2	0.44	11	2.40	28	6.11	55	12.01
Mid-Stable	7	0.41	3	0.18	4	0.24	25	1.47	193	11.35
Mid-Institutional	4	1.92	5	2.40	26	12.50	46	22.12	42	20.19
Large-Endangered	0	0.00	2	14.29	3	21.43	3	21.43	1	7.14
Large-Stable	4	3.01	3	2.26	9	6.77	24	18.05	29	21.80
Large-Institutional	69	31.80	64	29.49	121	55.76	145	66.82	134	61.75

Table 1: The *Coverage* of the 12 Ethnologue language classes in the listed resources. Under each resource, the *Count* column shows the number of languages in the relevant class included in the resource and the % column shows that number as a percentage of the total number of languages in the class.

	Class	Contribution			Coverage			Language Count
		Facebook	Google	X+mB	Facebook	Google	X+mB	
Ethnologue	Small-Extinct	0.00	0.00	0.00	0	0	0	332
	Small-Endangered	4.96	0.95	0.88	0.32	0.05	0.05	2162
	Small-Stable	0.00	0.00	0.00	0	0	0	1168
	Small-Institutional	0.00	0.95	0.00	0	3.57	0	28
	Mid-Extinct	0.00	0.00	0.00	N/A	N/A	N/A	0
	Mid-Endangered	5.67	1.90	4.39	1.75	0.44	1.09	458
	Mid-Stable	3.55	0.00	1.75	0.29	0	0.12	1700
	Mid-Institutional	7.80	8.57	7.89	5.29	4.33	4.33	208
	Large-Extinct	0.00	0.00	0.00	N/A	N/A	N/A	0
	Large-Endangered	1.42	0.95	0.88	14.29	7.14	7.14	14
	Large-Stable	4.26	1.90	7.02	4.51	1.5	6.02	133
	Large-Institutional	72.34	84.76	77.19	47	41.01	40.55	217
Joshi et al. (2020)	0	7.80	0.00	1.75	0.18	0	0.03	6134
	1	11.35	3.81	9.65	12.31	3.08	8.46	130
	2	41.13	41.90	37.72	59.79	45.36	44.33	97
	3	19.86	27.62	26.32	93.33	96.67	100	30
	4	14.89	20.00	18.42	95.45	95.45	95.45	22
	5	4.96	6.67	6.14	100	100	100	7
	Total	141	105	114				6420

Table 2: *Contribution* and *Coverage* of the 12 Ethnologue language classes and Joshi et al. (2020) classes in the listed resources where *X+mB* refers to the union of *XLMR* and *mBERT*. If for Class C_i of total n_i members and a resource R_j of total m_j members, the number of members in C_i present in R_j is given by $u_{i,j}$ then, the contribution is $100(u_{i,j}/m_j)$ and the coverage is $100(u_{i,j}/n_j)$



(a) LDC (b) ELRA (c) Hugginface (d) Wikipedia (e) ACL Anthology

Figure 3: Boxplots showing the resources where the amounts corresponding to the Ethnologue language classes are countable. (As opposed to Boolean)

including the HugginFace data sets as well.

While both graphs have the same trends, as shown in Figures 5, some languages have changed the classes when Hugginface data is considered. Also the boundary between some classes is very much blurred. This cautions us not to rely on a hard categorisation based on data availability. On the other hand, we note a clear relationship between the language categories provided by Joshi et al. (2020), and the Ethnologue classes. As shown in Tables 3 and 4 , all the *Extinct* languages as well a vast ma-

ing due to admin-pages such as user pages and talk pages.

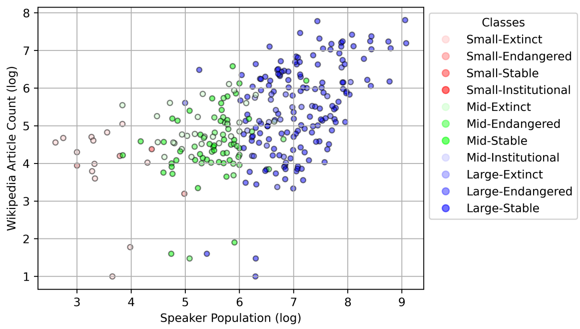


Figure 4: Speaker Population (log) vs Wikipedia Article Count (log).

jority of *Endangered* languages are in *class 0* of Joshi et al. (2020)'s categorization. On the other hand,*class 5* languages are all *Large-Institutional*.

5 Amount of Research Conducted for Different Languages

Now it is time we address the elephant in the room. What is the perspective and situation of ACL in the question we have discussed so far? Figure 1h shows that ACL Anthology has much less coverage for languages other than those belonging to *Large-Institutional* category. This observation aligns with what Joshi et al. (2020) reported in their conference-language inclusion analysis. However, interestingly, our results[13] show that ACL anthology covers more languages than what has been covered in data sources shown in Fig 1. This observation is affirmed by Fig 3e. Conversely, this also hint that those published research has not bothered to submit the associated data to public repositories.

In order to carry out further analysis on where low-resource language related papers are published, we tried to identify recently published language-specific survey papers. Surprisingly, language-specific survey papers on NLP technologies were extremely rare. We identified three survey papers: Sinhala (de Silva, 2021), Sindhi (Jamro, 2017), and Hausa (Zakari et al., 2021). We noted down the publishing venues of the research papers cited in these surveys. These results are plotted in Figure 7. In this, apart from the ACL statistics, we identified some prominent external categories: IEEE conferences, other conferences (not IEEE or ACL anthology), other journals (not in ACL anthology), pre-prints/thesis/white papers/reports. While different languages show different patterns (e.g. Sinhala mostly gets published in IEEE conferences, while Sindhi gets published in other journals) there is one common observation - there is extremely low number of papers in anthology, even for LREC and workshops published in ACL Anthology. Further look confirms that most of the other conferences and journals are either local or regional.

Further, we carried out the Google scholar queries shown in Table 5. We wanted to identify the amount of research reported for each language, with respect to NLP in general, as well as for some low-level and high-level NLP tasks. While it is obvious that Google scholar results may have false positives, the difference between ACL numbers and scholar numbers is significant.

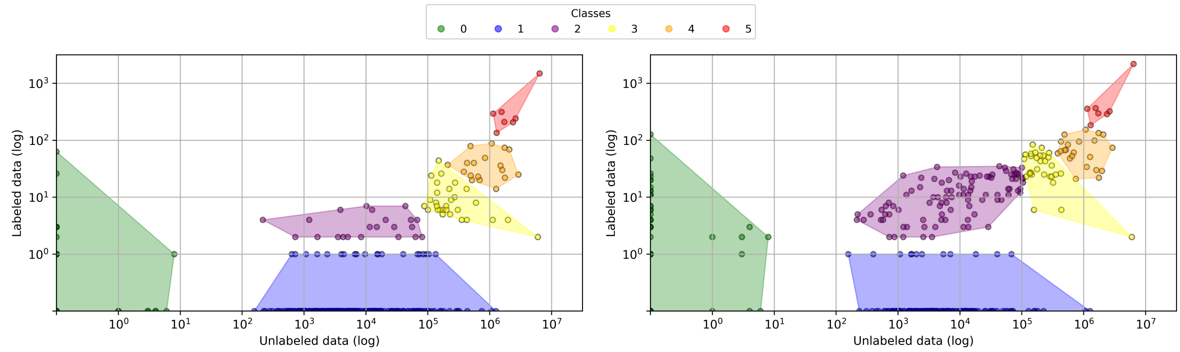
This observation could be due to several reasons: (1) the papers that are focusing on specific languages were not upto the standards of ACL main conferences or workshops, (2) some authors did not know about the ACL venues, or (3) some authors could not afford the registration and travel costs to ACL conferences. Considering the fact that most of the papers appeared in local/regional conferences and journals, the most possible reason for lack of papers in anthology could be the third.

6 Why do some languages remain low-resourced? *Case Study: Sinhala*

Out of the survey papers identified, de Silva (2021)'s paper was the most up-to-date. Thus, we went through all the Sinhala NLP papers cited in this survey paper to get an idea about the data sets presented in each of the papers, whether the code and data are publicly available and whether any tool has been released. Figure 6 visualizes this information. Only 11.43% of papers has data set publicly released, and only 9.71% of papers have code publicly released. Only 5.71% has any tool to be publicly used.

Working behind closed doors has shown its negative consequences - within a small time span, two research groups started working on Sinhala WordNets (Welgama et al., 2011; Wijesiri et al., 2014), but none has been successfully completed. Interestingly, none is available to be accessed now. This is common with some other tools that are claimed to be publicly released - they are not accessible. This suggests the lack of infrastructure support to maintain such tools. The author graph in de Silva (2021) highlights another side of the problem - the researchers seem to be working in silos, with almost zero interaction between research groups.

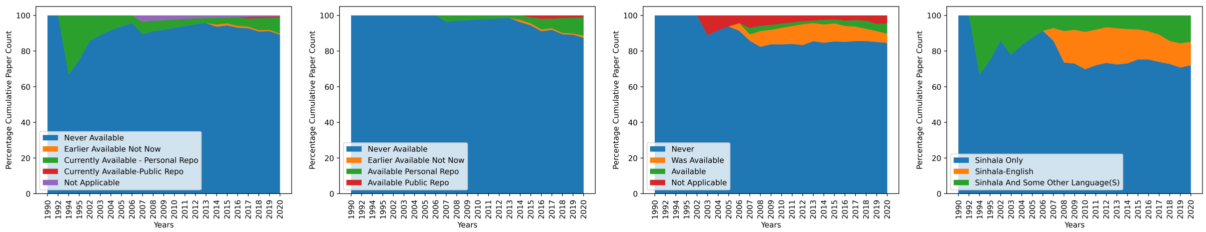
[13]More analysis in Appendix C



(a) *LDC* and *ELRA* as the annotated sources

(b) *LDC*, *ELRA*, and *Huggingface* as the annotated sources

Figure 5: Reconstructing Joshi et al. (2020) language classes with Wikipedia article count as the unannotated source and two configurations of annotated sources.



(a) Code Availability

(b) Data Availability

(c) Tool Availability

(d) Used Language

Figure 6: Sinhala NLP Percentage Cumulative analysis from the papers listed by de Silva (2021)

Joshi	Small				Mid				Large				Total
	Ex	En	St	In	Ex	En	St	In	Ex	En	St	In	
0	331	2146	1165	27	0	430	1676	164	0	11	109	75	6134
1	1	15	3	1	0	28	24	41	0	2	22	73	210
2	0	0	0	0	0	0	0	2	0	1	0	19	22
3	0	1	0	0	0	0	0	0	0	0	2	26	29
4	0	0	0	0	0	0	0	1	0	0	0	17	18
5	0	0	0	0	0	0	0	0	0	0	0	7	7
Total	332	2162	1168	28	0	458	1700	208	0	14	133	217	6420

Table 3: Confusion Matrix of Joshi et al. (2020) classes and Ethnologue language classes considering only *LDC* and *ELRA* as the annotated sources, where Ex=*Extinct*, En=*Endangered*, St=*Stable*, and In=*Institutional*.

Joshi	Small				Mid				Large				Total
	Ex	En	St	In	Ex	En	St	In	Ex	En	St	In	
0	331	2146	1165	27	0	430	1676	164	0	11	109	75	6134
1	1	12	3	1	0	19	23	24	0	2	18	27	130
2	0	3	0	0	0	9	1	18	0	1	4	61	97
3	0	1	0	0	0	0	0	1	0	0	2	26	30
4	0	0	0	0	0	0	0	1	0	0	0	21	22
5	0	0	0	0	0	0	0	0	0	0	0	7	7
Total	332	2162	1168	28	0	458	1700	208	0	14	133	217	6420

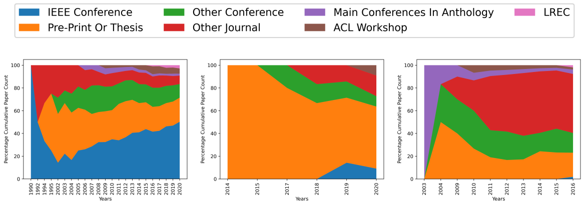
Table 4: Confusion Matrix of Joshi et al. (2020) classes and Ethnologue language classes considering *Huggingface*, *LDC*, and *ELRA* as the annotated sources, where Ex=*Extinct*, En=*Endangered*, St=*Stable*,and In=*Institutional*.

7 Discussion

1. What are the low-resource languages, and why are they low resourced? Most of the languages that lack data and pre-trained models and are missed out from technological platforms are either not institutional, or with small speaker populations. The institutional languages that lack resources are in the Global South. When there is no demand for language technologies due to unfavorable socio-economic conditions in the region, there would be a dearth of digital language resources and tools (Nekoto et al., 2020b). Another reason

Language	Anthology	Q1	Q2	Q3	Q4	Q5
Hausa	9	779	960	11	123	96
Sindhi	6	653	431	8	86	118
Sinhala	29	1130	644	14	146	187

Table 5: Amount of research publications for the languages Hausa, Sindhi, and Sinhala. Anthology - number of Anthology papers that mentioned this paper. Q1: “x”+ “natural language processing”, Q2: “x”+ “part of speech”, Q3: “x”+“grammar parsing”|“grammar parser”,Q4: “x”+ “question answering”, Q5: “x”+ “text classification”, where Q1-Q5 are Google scholar queries, and x = name of the language.



(a) Sinhala (b) Hausa (c) Sindhi

Figure 7: Cumulative percentage graphs showing where the NLP research of each language has been published.

could be the disconnection between different (indigenous) communities and the documentary linguistics community (Bird, 2020b). The fact that, most of these languages being from Global South, means that they do not have enough human resource to develop language resources (Nekoto et al., 2020a). Researchers in Global South who are working on low-resource languages being left out from the ACL forums, lack of interaction between local research communities, and reluctance to release the developed data resources, code, and models worsen this problem.

2. What can be done to take the low-resource languages out of the low-resource status? The starting point of developing NLP tools for languages is the availability of digital language content. For language content to be produced, the population should have a sufficient level of language, as well as computer literacy, plus there should be sufficient digital infrastructure within the country. For countries in the Global South, the governments may not have the bandwidth to fully satisfy these requirements, thus support of international and non-profit organization would be required.

Languages are vastly diverse with respect to their linguistic features (Dryer and Haspelmath, 2013), and linguistic aspects of some of those languages may be better understood by the local linguists. Thus, local language/linguistic researchers should take the lead for their languages.Given the fact that cross-lingual transfer is more effective between related languages and multilingual models built for regional languages have proven better than general models (Kakwani et al., 2020), communities within and across boarders working together to document and develop language resources would have a synergistic effect for all the involved languages. A recent success is the Masakhane project (Nekoto et al., 2020a). Given that many languages have practical limitations in creating data resources (e.g. not having enough speaker population), more research on zero-shot learning, few shot learning, transfer learning etc could help low resource languages.

ACL can focus on organizing shared data challenges, similar to shared tasks (Koehn et al., 2020). ACL also could take the lead in arranging more grants for researchers working in low resource languages. In fact, the existing funding schemes such as the NAACL Regional Americas fund[14] have produced positive impact (Ebrahimi et al., 2021). More D&I efforts, subsidies for researchers from global south to attend ACL venues, and above all creating/maintaining a forum of discussion related to the identified issues will be useful.

Finally, a comprehensive unambiguous list of languages and dialects in the world is needed. We noticed some inconsistencies between the language names used by Ethnologue, Joshi et al. (2020), etc.

8 Conclusion

The objective of this research was to provide a multi-facet analysis of the linguistic disparity in the world. We showed that this problem is due to socio-economic-linguistic factors. We provided some preliminary recommendations to get these languages out of *low-resourcefulness*, which we hope would be taken positively by the stakeholders. We hope there would be more frequent analysis of this sort, in particular to document the amount of research and NLP tools available for each language. In support of such efforts, we release our code to generate the visualizations shown in this paper[15].

[14] `https://bit.ly/NAACL_EmRe`

[15] Code Released After Acceptance

References

Antonios Anastasopoulos, Christopher Cox, Graham Neubig, and Hilaria Cruz. 2020. Endangered languages meet modern nlp. In *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 39–45.

Emily Bender. 2019. The benderrule: On naming the languages we study and why it matters. *The Gradient*.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Steven Bird. 2020a. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Steven Bird. 2020b. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519.

Andrew Cains. 2019. The geographic diversity of nlp conferences. *The Gradient*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Holger Schwenk, Ves Stoyanov, Adina Williams, and Samuel R Bowman. 2020. Xnli: Evaluating cross-lingual sentence representations. In *2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2475–2485. Association for Computational Linguistics.

Nisansa de Silva. 2021. Survey on publicly available sinhala natural language processing tools and research. *arXiv preprint arXiv:1906.02358v10*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

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

Gary F. Simons Eberhard, David M. and Charles D. Fennig, editors. 2021. *Ethnologue: Languages of the World*. Dallas, Texas: SIL International.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567, Online. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, et al. 2021. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *arXiv preprint arXiv:2104.08726*.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Wazir Ali Jamro. 2017. Sindhi language processing: A survey. In *2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT)*, pages 1–8. IEEE.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. inlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4948–4961.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

M Paul Lewis and Gary F Simons. 2010. Assessing endangerment: expanding fishman's gids.

META-NET. 2020. Meta-net white paper series: Key results and cross-language comparison. *META*.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020a. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi E Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, et al. 2020b. Participatory research for low-resourced machine translation: A case study in african languages. In *EMNLP (Findings)*.

George Orwell. 1945. *Animal Farm: A Fairy Story*. Secker and Warburg, London, England.

Ashwin Paranjape, Robert West, Leila Zia, and Jure Leskovec. 2016. Improving website hyperlink structure using server logs. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 615–624.

Veronica Perez-Rosas, Shihchen Kuo, William H Herman, Rada Mihalcea, et al. 2020. Upstage: Unsupervised context augmentation for utterance classification in patient-provider communication. In *Machine Learning for Healthcare Conference*, pages 895–912. PMLR.

Jishnu Ray Chowdhury, Cornelia Caragea, and Doina Caragea. 2019. Keyphrase extraction from disaster-related tweets. In *The world wide web conference*, pages 1555–1566.

Lane Schwartz, Emily Chen, Benjamin Hunt, and Sylvia LR Schreiner. 2019. Bootstrapping a neural morphological analyzer for st. lawrence island yupik from a finite-state transducer. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1.

Anil Kumar Singh. 2008. Natural language processing for less privileged languages: Where do we come from? where are we going? In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891.

Viraj Welgama, Dulip Lakmal Herath, Chamila Liyanage, Namal Udalamatta, Ruvan Weerasinghe, and Tissa Jayawardana. 2011. Towards a sinhala wordnet. In *Proceedings of the Conference on Human Language Technology for Development*.

Indeewari Wijesiri, Malaka Gallage, Buddhika Gunathilaka, Madhuranga Lakjeewa, Daya Wimalasuriya, Gihan Dias, Rohini Paranavithana, and Nisansa De Silva. 2014. Building a wordnet for sinhala. In *Proceedings of the seventh global wordnet conference*, pages 100–108.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130.

Rufai Yusuf Zakari, Zaharaddeen Karami Lawal, and Idris Abdulmumin. 2021. A systematic literature review of hausa natural language processing.

A Joshi et al. (2020) Class Descriptions

Class	Description	Language	
		Count	Examples
0	Have exceptionally limited resources, and have rarely been considered in language technologies.	2191	Slovene Sinhala
1	Have some unlabelled data; however, collecting labelled data is challenging.	222	Nepali Telugu
2	A small set of labeled datasets has been collected, and language support communities are there to support the language.	19	Zulu Irish
3	Has a strong web presence, and a cultural community that backs it. Have been highly benefited by unsupervised pre-training.	28	Afrikaans Urdu
4	Have a large amount of unlabeled data, and lesser, but still a significant amount of labelled data. have dedicated NLP communities researching these languages.	18	Russian Hindi
5	Have a dominant online presence. There have been massive investments in the development of resources and technologies.	7	English Japanese

Table 6: Language Categories identified by Joshi et al. (2020)

B Common Crawl Analysis

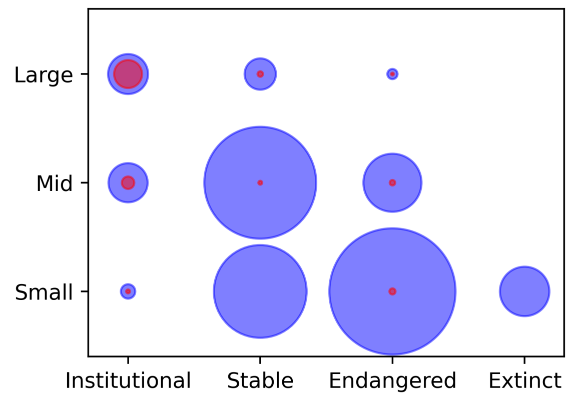


Figure 8: The 12 Ethnologue language classes where the size of each blue circle corresponds to the number of languages in that category and the size of each red circle corresponds to the coverage of that class in Common Crawl.

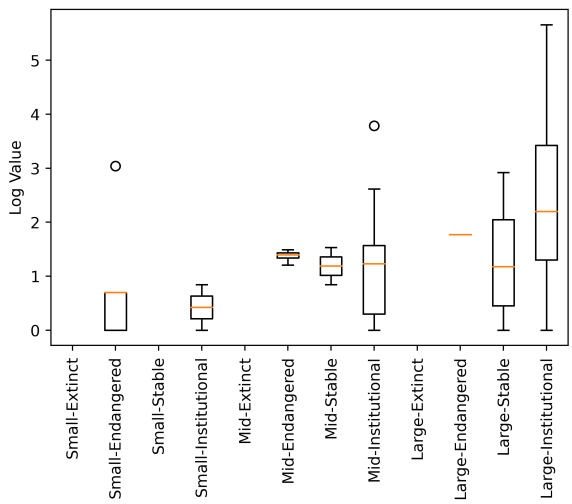


Figure 9: Boxplot showing Common Crawl data with the amounts corresponding to the Ethnologue language classes.

Class	CC	
	Count	%
Small-Extinct	0	0.00
Small-Endangered	4	0.19
Small-Stable	0	0.00
Small-Institutional	1	3.57
Mid-Endangered	4	0.87
Mid-Stable	2	0.12
Mid-Institutional	19	9.13
Large-Endangered	1	7.14
Large-Stable	4	3.01
Large-Institutional	100	46.08

Table 7: The Coverage of the 12 Ethnologue language classes in the Common Crawl. The Count column shows the number of languages in the relevant class covered by the Common Crawl and the % column shows that number as a percentage of the total number of languages in the class.

C ACL Publication History and Performance

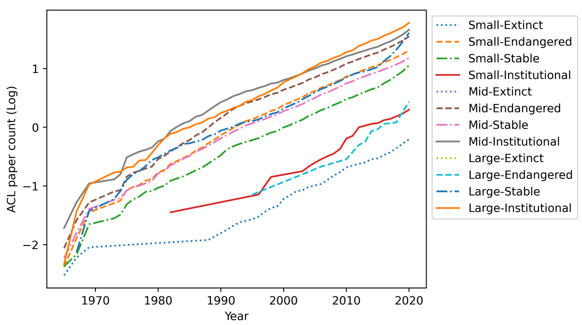
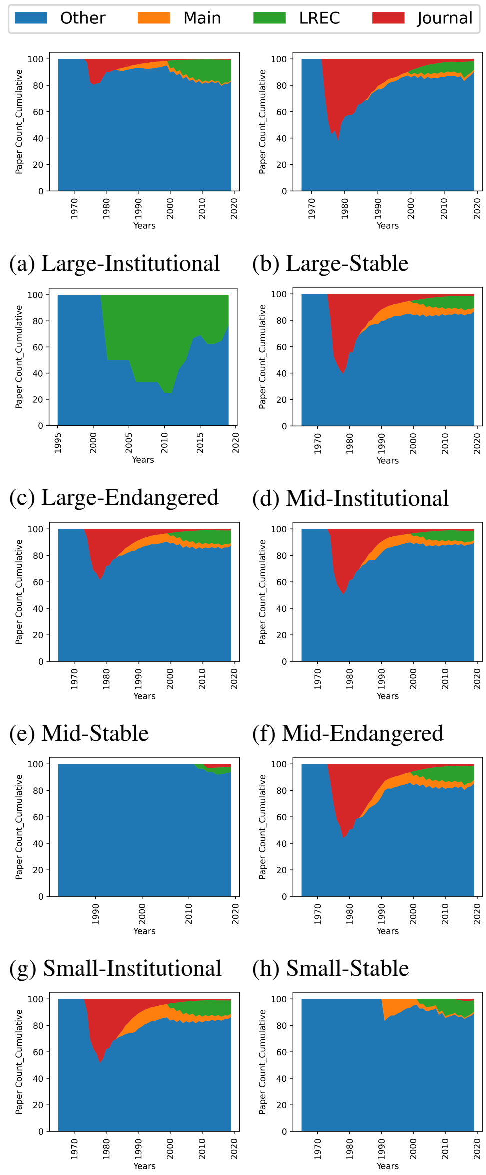


Figure 10: ACL publication count for the 12 Ethnologue language classes (cumulative class-normalized log)



(a) Large-Institutional (b) Large-Stable

(c) Large-Endangered (d) Mid-Institutional

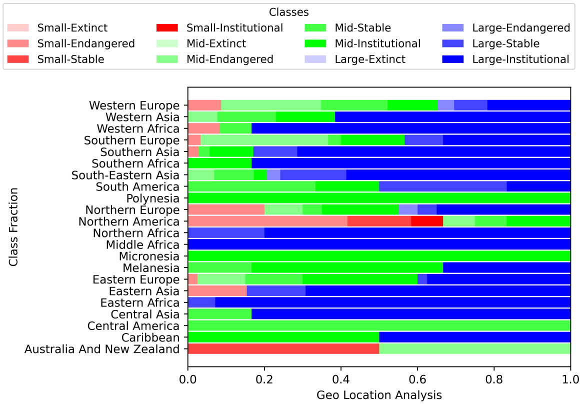
(e) Mid-Stable (f) Mid-Endangered

(g) Small-Institutional (h) Small-Stable

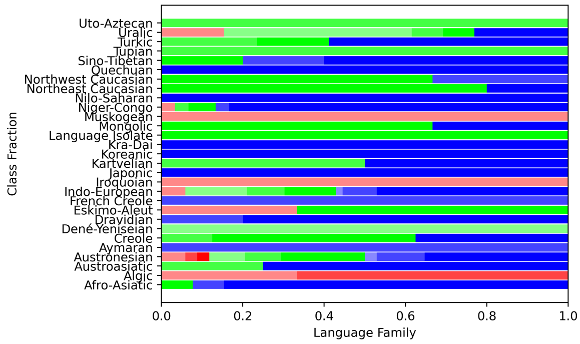
(i) Small-Endangered (j) Small-Extinct

Figure 11: ACL Participation of the languages belonging to the 12 Ethnologue language classes (Only the existing 10 classes shown here.)

D Wikipedia 12 Class Analysis



(a) Geological Location



(b) Language Families

Figure 12: The distribution of languages that have wikis among the 12 Ethnologue Classes