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

Surangika Ranathunga and **Nisansa de Silva**
Department of Computer Science and Engineering
University of Moratuwa, Katubedda 10400, Sri Lanka
{surangika, nisansaDds}@cse.mrt.ac.lk

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. We show that simply categorising languages considering data availability may not be always correct. 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, GDP and the speaker population of languages and provide possible reasons for this disparity, along with some suggestions to overcome the same.

1 Introduction

Even after more than fifty years since the inception of the fields of Computational Linguistics (CL) and Natural Language Processing (NLP), 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. 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 speakers of low-resource languages from reaping the benefits of NLP in areas such as healthcare (Perez-Rosas et al., 2020), disaster response (Ray Chowdhury et al., 2019), law (Ratnayaka et al., 2020), and education (Taghipour and Ng, 2016).

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., 2018; Ebrahimi et al., 2022), benchmarks (Hu et al., 2020) and techniques that favour low-resource languages (Schwartz et al., 2019); all of which, are very promising developments. However, the problem is not fully solved, and this disparity should be quantified to understand the gravity of the problem (Khanuja et al., 2022). Such an understanding is the first step in developing solutions to solve the problem (Grützner-Zahn and Rehm, 2022).

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. Hedderich et al. (2021) considered the availability of auxiliary resources such as lexicons in addition. Faisal et al. (2022) estimated the level of language speaker representation in dataset content. 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 datasets available in the LDC[2] and ELRA[3] data repositories.

However, such a data-centric perspective tends

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)

[2] https://catalog.ldc.upenn.edu/

[3] http://catalog.elra.info/en-us/

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 823–848
November 20–23, 2022.

to overlook 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 models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020). Moreover, such a narrow view 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 speaker population.

This paper 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[4]) (Eberhard et al., 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 simply categorising languages using data availability as done by Joshi et al. (2020) can be misleading.** We also show that many languages are neglected with respect to all the considered 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 and GDP. We also provide possible reasons for this disparity, along with some recommendations to eradicate the same.

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. The language list we extracted, as well as the selection criteria are in Appendix A.

Ethnologue languages are categorised into 12 classes, based on 2 variables: *Population* and *Vitality*. *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 et al., 2021). *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).

We plotted the languages in a 12-point grid, according to vitality and number of speaker population. The size of the outer circles corresponds to the number of languages in one category. According to Figure 1, a large number of languages are endangered with small speaker populations, or stable but with mid or small speaker population numbers. Note that two classes do not have any representation in this grid. Therefore, hereafter we only refer to the remaining 10 classes.

3 Resource & Tool Support Distribution

We analyse how languages in the Ethnologue categories are being treated with respect to data (annotated and un-annotated) availability, inclusion in multilingual web-based platforms and inclusion in pre-trained multilingual models. This dataset was extracted in October-November, 2021. The dataset preparation process is given in Appendix B.

3.1 Un-annotated Data Availability

There are two possible sources: Wikipedia data and CommonCrawl. However, the latter covers only 160 languages[5], compared to the 318 languages in Wikipedia (excluding the 7 constructed languages). Thus, we focus on Wikipedia data as the source of un-annotated data. The CommonCrawl data analysis is briefly reported in Appendix C.

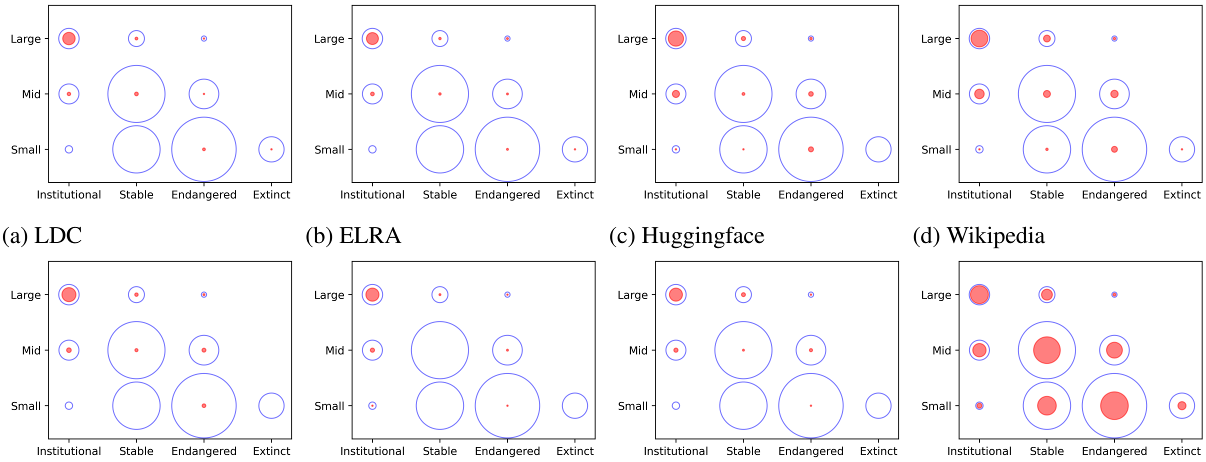
3.2 Annotated Data Availability

Although Joshi et al. (2020) used *LDC* and *ELRA* to retrieve the number of annotated datasets, not all datasets in these sources are available for free, and there are membership charges. This can be quite a disadvantage for researchers working under severe financial constraints. Thus not many languages have their datasets in these repositories. In order to highlight that categorising languages while having incomplete information about datasets gives a wrong picture (see Section 5), we selected another public data repository - *Huggingface* data sets[6]. Huggingface is known to be sparse, and the data has to be accessed via an API. On the positive side, despite being launched in 2021, it has more datasets than ELRA and LDC. Huggingface datasets are categorised according to language and task. Many existing datasets, such as those hosted in OPUS (Tiedemann and Thottingal, 2020), have

[4] https://bit.ly/3kJircB

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

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



(a) LDC (b) ELRA (c) Huggingface (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 outer 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.

been already linked to Huggingface. Other possible data repositories include Zenodo[7] and CLARIN[8]. However, these do not have a language-wise categorisation or have a smaller number of 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, and reinforces the legitimacy of a language (CBC, 2022). Conversely, the languages that are not supported will be less and less used (Bird, 2020). 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 (Twitter supports 36 languages).

3.4 Pre-trained Multilingual Model Coverage

mBERT (trained with Wikipedia data) and *XLM-R* (trained with CommonCrawl data) are the most popular models as of today. These models are quite effective in zero-shot and few-shot NLP tasks (Hu et al., 2020; Lauscher et al., 2020). They mostly perform better for languages that are included in the pre-training stage (Muller et al., 2021) and outperform their monolingual counterparts for low resource languages (Wu and Dredze, 2020). Considering the above facts, and noting that training multilingual models is computationally expensive, languages that are included in mBERT and XLM-R would have an edge over those that are not.

4 Aggregated Analysis

4.1 Overview

Inner circles in 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. Note that the language categorisation shown in the bottom part of Table 2 is newly created by us, according to Joshi et al. (2020)'s categories (see Table 5 in Appendix D).

It is evident that language resource creation and technology availability have been mostly centred around institutional languages with high speaker populations, while small and endangered languages have mostly been ignored.

4.2 Data Availability

Table 1 shows that Wikipedia has some coverage for all existing categories, including some extinct languages, which may be partly due to research efforts[9] (Paranjape et al., 2016). However, LDC, ELRA and Huggingface 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, which is more decentralized.

[7] https://zenodo.org/
[8] https://www.clarin.eu/content/data
[9] https://stanford.io/3mXQK0Z

Class	LDC		ELRA		Huggingface		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 10 existing 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-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-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 10 existing 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 $100(u_{i,j}/n_j)$

4.3 Inclusion in Web-based Platforms and Pre-trained Models

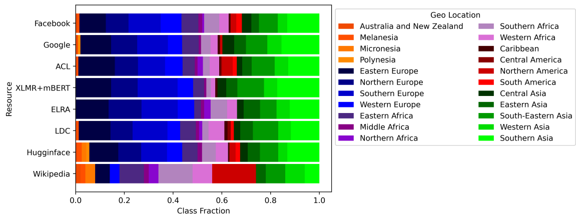
In Table 2 we observe that Facebook and Google platforms mainly cover institutional languages, with a negligible representation of other languages. The same is observed for the coverage in the pre-trained multilingual models *mBERT* and *XLM-R*, released by Google and Facebook, respectively. Note that such models suffer from 'curse of multilinguality' (Conneau et al., 2020), and the number of languages in the models has to be bound.

4.4 Impact of Socio-Econo-linguistic Factors

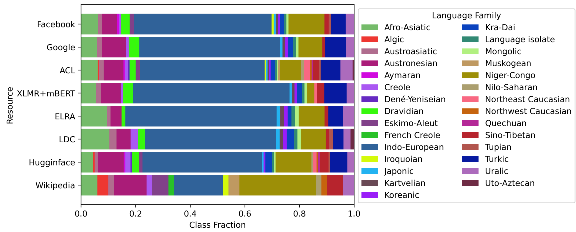
Figures 2a and 2b visualise the coverage of these different platforms and resources with respect to the geographical location and family of a language. We can see that all these criteria are biased towards the *Indo-European* family and the *Europe* region. This is not surprising, given the emphasis placed on language resource development in the European region (META-NET, 2020).

Further analysis on the languages covered by *mBERT* and *XLM-R* models shows that the language selection has indeed been motivated by the speaker population and geographical location. Most of the languages included in these models are *Large-Institutional*. As shown in Figure 10 in Appendix E, 75% of non-Large-Institutional languages included in either XLM-R or mBERT are from Europe, and the rest are from Asia. All these Asian languages are either *Mid-Institutional* or *Large-Stable*. On the other hand, most of the Large-Institutional languages omitted from these models are in the African region (51%). This also explains the observation made by Hu et al. (2020), where pre-trained multilingual models perform better for Indo-European languages.

Interestingly, Wikipedia has been more democratic compared to other resources, mainly because content creation is de-centralized (More analysis in Appendix F). LDC and ELRA data sources are more concentrated in the Europe area. In contrast, Huggingface is more distributed. This affirms the importance of free data repositories.



(a) By Geographical Location of the Language Origin



(b) By Language Families

Figure 2: The Distribution of Resources[10]

However, Figure 1 only can be misleading, as the amount of data varies across languages even within the same category. We derived the box plots shown in Figure 3, which uncovered a noticeable disparity between language categories. Aside from the inter-class disparities, Figure 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 and used Pearson correlation. As can be seen, the number of Wikipedia articles available has a *moderate correlation* (0.561474) to the GDP represented by the speakers of that language[11]. Blasi et al. (2022) found a similar correlation, between population and GDP, and the number of research papers per language. Here we show that the same GDP impact can be seen in the size of Wikipedia [12].

4.5 Task-wise and Size-wise Analysis

We also carried out a preliminary analysis of NLP task-wise data availability in HuggingFace. Results are shown in Table 6 in Appendix H. Despite this task categorisation being extremely noisy, there are some interesting observations. Popular NLP tasks such as translation, text classification, text generation and text retrieval have the highest counts, at least for Large-Institutional category. For all the tasks, dataset availability is the highest for large-Institutional, followed by Mid-Institutional.

As for the size of datasets, we are only aware of OPUS, which records the number of sentences per language. According to the results in Table 7 in Appendix I, not only the number of datasets, but the amount of data samples also depends on the language class.

5 Revisiting Data Availability-based Language Categorisation

In order to analyse the robustness of using annotated data availability to categorise languages, we recreated Joshi et al. (2020)'s language category plot. We plot the availability of annotated data in LDC and ELRA against the unannotated wiki data in 5a[13]. In 5b we plot the same graph including the HuggingFace datasets as well.

We note a clear relationship between Joshi et al. (2020) categories, and the Ethnologue classes. As shown in Tables 8 and 9 in Appendix K, all the *Extinct* languages as well a vast majority 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*.
Although both graphs have the same trends, as shown in Figure 5 and the discussion in Appendix K, 87 languages have changed their class (84 are promotions) when Huggingface is considered. Interestingly, class of Welsh changes from 1 to 3, and Azerbijanis changes from 1 to 4. This cautions us not to rely on a hard categorisation based on a partial set of data repositories.

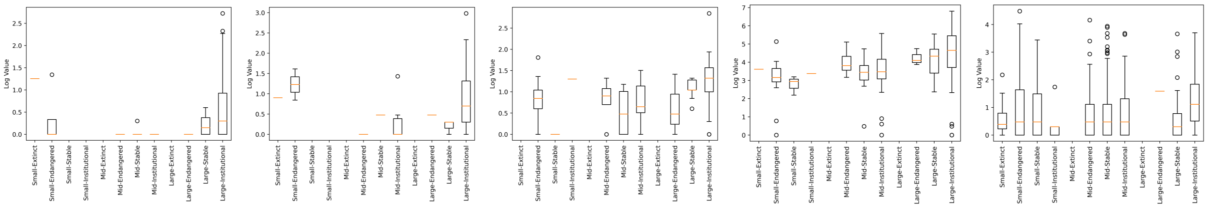
To further explain the limitations of a language categorisation that relies on annotated datasets de-

[10]Larger versions are available in Appendix J.

[11]GDP, population of a country and the percentage of language speakers of a country are extracted from https://www.worlddata.info/. Missing entries were identified from Wikipedia and Ethnologue. The GDP for a given language is calculated by a variation of Blasi et al. (2022) where a GDP of each country is first distributed proportionally among languages spoken as L1 in that country and then the GDP of the language is calculated by summing the aforementioned portions. The colour of each data point is taken according to the class in Ethnologue.

[12]An equivalent analysis between population and the number of Wikipedia articles is in Appendix G.

[13]Different to (Joshi et al., 2020), we considered the number of *Wikipedia articles*, as considering *pages* could be misleading due to admin-pages such as user pages and talk pages.



(a) LDC (b) ELRA (c) Huggingface (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)

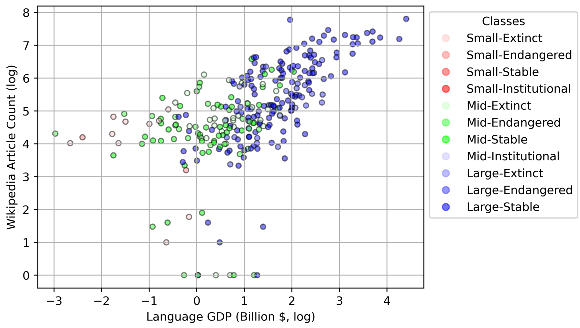


Figure 4: Language GDP in Billions of Dollars (log) vs Wikipedia Article Count (log).

rived only from a set of repositories, consider Gorontalo and Gujarati languages. Both belong to class 1 in Joshi et al. (2020)'s categorisation. Gorontalo is a mid-endangered language with 1 million L1 speakers. It is not in Google keyboard or Facebook language list, nor is it included in pre-trained multilingual models. In contrast, Gujarati is a large institutional language with 56 million L1 speakers. It is included in all of the above three lists. In addition, `gujarati + "Natural Language Processing"` query returns 1960 results in Google scholar, and has 189 papers in ACL anthology corpus extracted by Rohatgi (2022). The corresponding query for Gorontalo returns only 81 results, and 0 results in Rohatgi (2022)'s corpus. Bird (2022) builds a similar argument by comparing Tamil (75 million speakers) and Cree (75,000 speakers).

6 Amount of Research Conducted for Different Languages

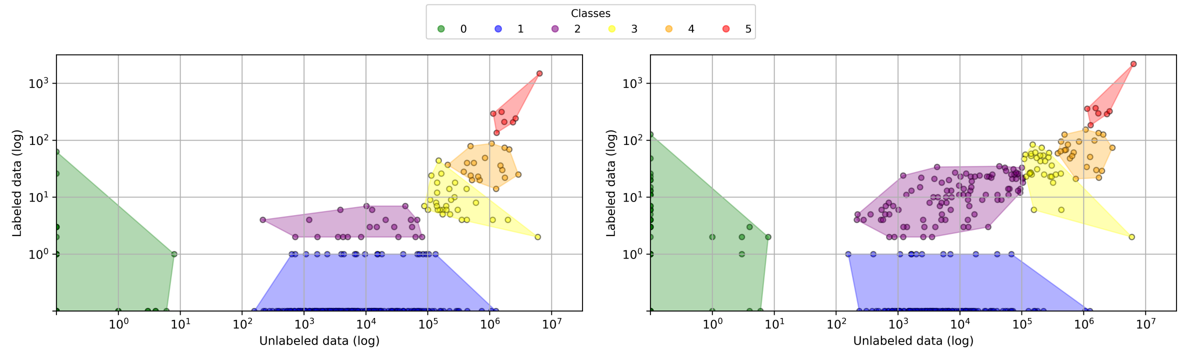
We use the research papers published in ACL Anthology curated in Rohatgi (2022)'s corpus, which contains full papers and their metadata of all Anthology papers upto now[14]. Figure 1h shows that ACL Anthology, even when considering LREC and workshops associated with ACL, has less coverage for languages other than those belonging to the *Large-Institutional* category. As further shown in Appendix L, research papers in ACL anthology for categories other than *Large-Institutional* category comes mainly from LREC and workshops. This observation aligns with what Joshi et al. (2020) reported in their conference-language inclusion analysis. However, interestingly, our results 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. While this could mean that datasets are re-used across research, it could mean the data used in these papers might be in personal/institutional repositories, or the data might have not been released at all.

In order to further validate this hypothesis, we went through a random set of 50 papers extracted from ACL Anthology 2020. However, only 16 papers presented new datasets. Since the number is not enough to conduct a deeper analysis, we extracted the first 100 papers from LREC 2022 proceedings. Our assumption was LREC papers would be more focused on presenting new datasets. Out of the 56 LREC papers that presented new datasets, only 5 (9%) have published their data in public repositories. 80% papers indicated that they have released the data in personal or public repositories. The process to collect this data, as well as the visualizations are given in Appendix M.

We also conducted a mini survey (https://forms.gle/FbWhChAeBE5KBvsQ8) among NLP researchers[15]. The survey questions and the responses from 81 participants in 31 countries are given in Appendix N. First and foremost, the results further confirm that categorising languages considering only a few data repositories is mis-

[14]We extract the full text from the beginning of abstract to the beginning of references excluding acknowledgements.

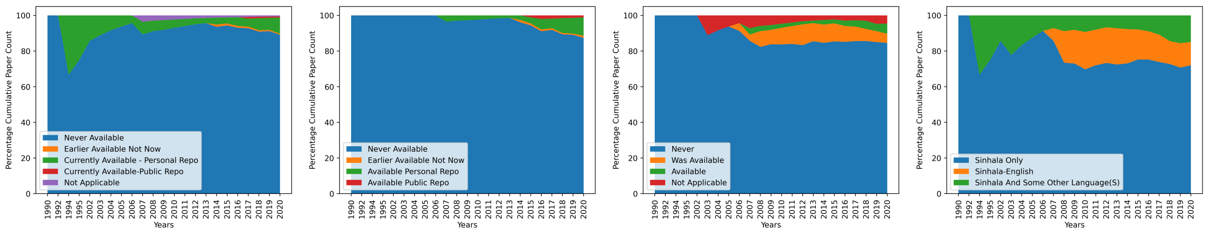
[15]By sending the survey participation request via public mailing lists, private interest groups and personal contacts



(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

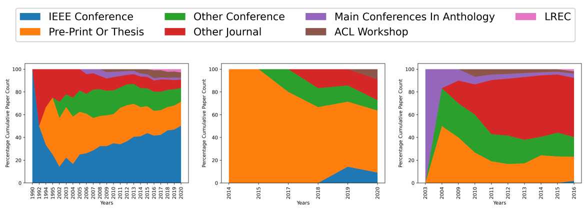
(c) Tool Availability

(d) Used Language

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

leading, as there are many such repositories - the repository selection depends on personal, as well as institutional choices. It is also interesting to note that there is a noticeable number of respondents who are not aware of such data repositories. It also explains why the language count is higher in ACL Anthology compared to language counts in ELRA/LDC/HuggingFace - researchers mostly prefer to keep their data in their personal repositories.

In order to further understand where papers of languages traditionally known as low-resource languages are published, we carried out a language-specific analysis. We identified three survey papers: Sinhala (de Silva, 2021), Sindhi (Jamro, 2017), and Hausa (Zakari et al., 2021) (all are large-institutional languages, with Joshi et al. (2020)'s category being 0, 1 and 2, respectively). We noted down the publishing venues of the research papers cited in these surveys. These results are plotted in Figure 7. We see that apart from the ACL venues, there are: IEEE conferences, other conferences (not IEEE or ACL anthology), other journals (not in ACL anthology) and pre-prints/thesis/white papers/reports. While different languages show different patterns (e.g. Sinhala mostly gets published in regional IEEE conferences, while Sindhi gets published in other (regional) journals) there is one common observation - there is extremely low number of papers in anthology, even for LREC and workshops published in ACL Anthology. A further look confirms that most of the other conferences and journals are either local or regional. Further, we carried



(a) Sinhala

(b) Hausa

(c) Sindhi

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

out the Google scholar queries shown in Table 3 in order 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 has been shown that Google scholar results have false positives (Ranathunga et al., 2021), the difference between ACL numbers and scholar numbers is significant.

Language	Anthology	Q1	Q2	Q3	Q4	Q5
Hausa	94	779	960	11	123	96
Sindhi	35	653	431	8	86	118
Sinhala	100	1130	644	14	146	187

Table 3: 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.

7 Case Study: Sinhala

In Joshi et al. (2020)'s language categorisation, the class of Sinhala is ambiguous - while Sinhala is categorised as class 0, its synonymous term 'Sinhalese' is categorised into class 1. Despite its exact category, Sinhala has been considered a low-resource language even in recent research (Guzmán et al., 2019; Sarioglu Kayi et al., 2020). In contrary, Sinhala has its presence in Wikipedia, Huggingface, Google keyboard, Facebook, as well as XLM-R. So why is Sinhala still considered low-resource?

We went through all the Sinhala NLP papers cited in de Silva (2021)'s survey paper to get an idea about the datasets presented in each of the papers, whether the code and data are publicly available and whether any language tool has been released. Figure 6 visualizes this information. Only 11.43% of papers has data set publicly released (10.29% in personal repositories, 1.14% in public repositories) and only 9.71% of papers have code publicly released. Only 5.71% have released tools.

Working behind closed doors has shown its negative consequences - within a small time span, two research groups started working on Sinhala WordNet (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. de Silva (2021)'s author graph highlights another problem - the researchers seem to be working in silos, with almost zero interaction between research groups. On the positive side, recently, the use of pre-trained multilingual models has shown its benefit (Rathnayake et al., 2022; Thillainathan et al., 2021; Dhananjaya et al., 2022).

8 Discussion

We analysed the linguistic disparity in a global scale. Thus, inevitably, the analysis was limited to only a set of factors, which could be determined by the freely available data. In contrast, the EU-funded European Language Equality (ELE) project (Grützner-Zahn and Rehm, 2022) categorised European languages with respect to language resources, tools, as well as contextual factors such as economic and financial factors. This analysis is very comprehensive, however, it does not shed any light on the vast majority of the languages in the world. An ambitious project would be to extend this effort in a global scale.

In order to highlight the importance of carrying out frequent analysis of linguistic disparity, we recorded the number of Wikipedia articles and Huggingface dataset counts as of July 2022. As shown in Tables 11 and 12 in Appendix O, 611 new datasets were added to *Large-Institutional* category alone, within less than an year. However, for the small-extinct/endangered/stable/institutional classes altogether, only 9 datasets have been added. This trend of rich getting richer is a concern as this shows that the average interest still lies with the few languages that are already enjoying an abundance of datasets. As for Wikipedia, an astounding number of articles have been added to *Large-Institutional* category. Many other language categories have also received articles, suggesting community involvement in content creation. It would be interesting to check whether this content addition impacts the Ethnologue categorisation, however, we lack historical Ethnologue data to conduct this analysis.

We highlighted that the inclusion of a language in a pre-trained multilingual model provides an added advantage for a language. However, not many languages are included in the available models. At least for the languages where text data is there, pre-trained multilingual models should be publicly released. While doing so, models including related languages would be more beneficial (Khanuja et al., 2022; Kakwani et al., 2020).

Many languages are missing in Wikipedia or CommonCrawl. Thus, community engagement should be promoted and funded to improve language-specific Wikipedias. Wikimedia grant scheme is one useful lifeline [16]. Bapna et al. (2022) reported the possibility to web-mine data for 1500 languages. We hope this data will be publicly available. For spoken languages that do not have any text (Bird, 2022), extra effort is needed to collect speech data. There should be initiatives (preferably funded, for languages in Global South) to create annotated data, even in small quantities, for languages that have monolingual data.

Inuktitut, a mid-institutional language with about 40,000 speakers has been recently included in Facebook, with the support from a local learning center (CBC, 2022). This is welcome news - collaborations between locals and tech giants can facilitate the inclusion of languages in the web platforms. However, Inuktitut is a North American language. Adding an African language to Facebook or Google language list may face more challenges.

Not all authors have added data to public repositories, which also have limitations. Particularly, many do not have language or task-wise categorisation of data, and meta data is not collected. We hope ACL can take the initiative to setup a repository that does not have the limitations identified in our survey. A similar initiative is preferable to create an infrastructure to host language tools.

As NLP researchers from Global South, we have our own interpretation of the reasons for many languages having research papers in non-ACL venues. Many reviewers in ACL conferences are sceptical of techniques tested only on a language not popularly known. With time, authors stay away from submitting to these venues, as they anticipate the possible outcome. While there are several workshops welcoming low-resource language research, most of them are non-indexed. This is a concern in institutions that take indexed publications as a measure of academic success. Travelling to ACL venues is expensive for researchers from the Global South, and many conferences are held in countries with high visa restrictions. Thus, hybrid events with less expensive online versions are a blessing for such researchers. Blasi et al. (2022) found no evidence that research papers dealing with more languages in their evaluation having any advantage over those that do not when considering the number of citations, which means researchers have no incentive to test their systems in many languages. Organising multilingual shared tasks and more recognition for papers presenting multilingual datasets might help alleviating this problem.

Finally, we showed the need to discuss the full situation of languages used in research with respect to the socio-economic status as well as resource availability, rather than saying the language is low-resource just by considering data availability.

These are the limitations of this study: The use of language names is not consistent across different data sources. We put every effort to use a uniform language list across data sources, however there can be a few languages that we missed. We used the logic by Blasi et al. (2022) to check the existence of a language name in a paper. Thus, the extracted data may have some noise, so does Google scholar search. As already mentioned, task-wise dataset analysis is extremely noisy.

In order to carry out better analysis in the future, we recommend: (1) Creating a map of synonyms of languages, (2) a widely accepted list of NLP tasks, (3) NLP papers adhering to the Bender rule (Bender, 2019) and (4) recording the meta data of the datasets reported in repositories and in research papers (Data statements (Bender and Friedman, 2018) would be a good starting point).

9 Conclusion

The objective of this research was to provide a multi-facet analysis of the linguistic disparity in the world. We showed that such an analysis provides a more detailed view of the linguistic disparity, rather than depending on the dataset (particularly annotated) availability. 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 support of such efforts, we release our code to generate the visualisations shown in this paper as well as the relevant data[17].

10 Acknowledgement

This publication was funded by the publication support scheme of the SRC of University of Moratuwa (UoM). We thank Graham Neubig and Antonis Anastasopoulos for providing their code to process ACL anthology, Shravan Nayak for providing OPUS data counts, Undergraduates of the department of CSE, UoM and survey participants.

[16] https://meta.wikimedia.org/wiki/Grants:Start

[17] https://bit.ly/AACL2022SomeLanguages

11 Ethical Impacts (Responsible NLP)

We employed three workers to manually enter statistics into a spreadsheet. One was an undergraduate, the other two were graduates. One was a male, and the other two were females. However, this demographic information was not recorded, as it is not needed for the task. We gave them initial instructions verbally over a meeting, and demonstrated the data extraction process. They worked remotely. They were compensated on an hourly rate. Payment rates were according to the approved rates of the university.

The survey was anonymous. We did not collect the email addresses of the participants. The only demographic information we collected was the country of residence. The individual responses have not been publicly released. Only the aggregated results are included in this paper. The participants have discussed limitations of individual data repositories. However, such specific comments are not included in this paper.

The language list we created is publicly available. We mentioned the sources we used to extract data. The limitations in data collection and processing were listed in the discussion. Our code to generate visualisations is publicly available, for the same visualisations to be developed in the future.

We believe that our study provided valuable insights to the linguistic disparity in a global scale, which would be useful in formulating action plans to mitigate this disparity.

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, Barcelona, Spain (Online). International Committee for Computational Linguistics.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. 2022. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*.

Emily Bender. 2019. The #Benderrule: On naming the languages we study and why it matters. *The Gradient*, 14.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

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. 2020. 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. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

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

News CBC. 2022. 'Reinforces the legitimacy of our language': Inuktitut officially available on Facebook desktop. `https://bit.ly/3IJraWW`.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. 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 *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.

Vinura Dhananjaya, Piyumal Demotte, Surangika Ranathunga, and Sanath Jayasena. 2022. BERTifying Sinhala - a comprehensive analysis of pre-trained language models for Sinhala text classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7377–7385, Marseille, France. European Language Resources Association.

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

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Fahim Faisal, Yinkai Wang, and Antonios Anastasopoulos. 2022. Dataset geography: Mapping language data to language users. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3381–3411, Dublin, Ireland. Association for Computational Linguistics.

Annika Grützner-Zahn and Georg Rehm. 2022. Introducing the digital language equality metric: Contextual factors. In *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, pages 13–26, Marseille, France. European Language Resources Association.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

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.

Armin Hoenen and Marc D Rahn. 2021. Migration of small and endangered languages into the Wikipedia. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 2, pages 41–47.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multitask 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, Online. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Simran Khanuja, Sebastian Ruder, and Partha Talukdar. 2022. Evaluating inclusivity, equity, and accessibility of NLP technology: A case study for Indian languages. *arXiv preprint arXiv:2205.12676*.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

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, Online. Association for Computational Linguistics.

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*.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

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.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2021. Neural Machine Translation for low-resource languages: A survey. *arXiv preprint arXiv:2106.15115*.

Himashi Rathnayake, Janani Sumanapala, Raveesha Rukshani, and Surangika Ranathunga. 2022. Adapter based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification.

Gathika Ratnayaka, Nisansa de Silva, Amal Shehan Perera, and Ramesh Pathirana. 2020. Effective approach to develop a sentiment annotator for legal domain in a low resource setting. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 252–260, Hanoi, Vietnam. Association for Computational Linguistics.

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

Shaurya Rohatgi. 2022. ACL Anthology Corpus with Full Text. Github.

Efsun Sarioglu Kayi, Linyong Nan, Bohan Qu, Mona Diab, and Kathleen McKeown. 2020. Detecting urgency status of crisis tweets: A transfer learning approach for low resource languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4693–4703, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lane Schwartz, Emily Chen, Benjamin Hunt, and Sylvia L.R. Schreiner. 2019. Bootstrapping a neural morphological analyzer for St. Lawrence Island Yupik from a finite-state transducer. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 87–96, Honolulu. Association for Computational Linguistics.

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, Austin, Texas. Association for Computational Linguistics.

Sarubi Thillainathan, Surangika Ranathunga, and Sanath Jayasena. 2021. Fine-tuning self-supervised multilingual sequence-to-sequence models for extremely low-resource NMT. In *2021 Moratuwa Engineering Research Conference (MERCon)*, pages 432–437. IEEE.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

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, Tartu, Estonia. University of Tartu Press.

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, Online. Association for Computational Linguistics.

Rufai Yusuf Zakari, Zaharaddeen Karami Lawal, and Idris Abdulmumin. 2021. A systematic literature review of Hausa Natural Language Processing.

A Language List used in the Study

When looking at the list of languages used by Joshi et al. (2020), we noticed that it was quite inconsistent. It had dialects and alternate names of languages as separate entities. For example, it contained *Sinhala* as well as *Sinhalese*. The former is the correct name of the language. The latter is the name of the ethnicity of the people who speak *Sinhala*. While there are online sources that erroneously use *Sinhalese* as the name of the language, it would not suit a research on language to use this term. In addition to that, this also meant that the resources listed for the *Sinhala* are distributed among the two alternate names. This resulted in Joshi et al. (2020) categorising *Sinhala* as a class 0 language and *Sinhalese* as a class 1 language. Moreover, Joshi et al. (2020)'s list covers less than half of the languages in the world. Shortfalls such as this motivated us to look elsewhere for a more reliable and consistent source for creating our language list.

We used Ethnologue as our primary source for creating the language list. They list information on 7139 living language entries[18] in the world, including dialects. Ethnologue also lists some dialects and minor schisms within languages as separate entities. However, they are consistent in reporting them. For example, for *German*, they cleanly list *German, Pennsylvania*, *German, Standard*, and *German, Swiss*. Thus, when we were collecting language names from them, we could simply take the term that precedes the comma.

While this was an efficient strategy to automatically reduce dependencies, when we proceeded to prepare data sets as explained in Appendix B with the 'list of Wikipedias'[19], it was evident that some cases that are represented as a single language in Ethnologue has multiple entries in Wikipedia due to them being functionally distinct. An example of this is *Norwegian*, which has only one entry in Ethnologue[20] but separate Wikipedias for *Norwegian (Bokmål)*[21] and *Norwegian (Nynorsk)*[22]. In these cases, we added distinct entries for the differing languages. When a singular language in Ethnologue was split this way, the resultant languages were given the class of the source language. Given that all such splits (rather predictably) happened with *Large* languages, the margin of error is still within safe values given the vast difference between the threshold value for the *Large* class and the *Mid* class. Some languages have multiple names, and there were instances where different data sources were using different names. When a language in (say) Wikipedia was not is Ethnologue, we did a web search to check for the alternative names. We used the Ethnologue version of language names.

After these steps we compiled a list of 6420 unique languages to derive our language list, which we have made publicly available [23] for the benefit of future language researchers.

B Dataset Preparation

The 'list of Wikipedias' page in Wikipedia records the statistics of wiki pages in different languages[24]. We manually recorded the number of Wikipedia articles per language, according to this wiki page. CommonCrawl also has explicitly listed the number of HTML web pages per language[25], which we manually recorded. We manually recorded the dataset statistics from LDC, ELRA and Huggingface. In all these repositories, datasets are grouped by language.

The L1 speakers for a language was extracted from the infobox[26] of the corresponding Wikipedia page. There were few cases, where for some small languages, the number of L1 speakers were not mentioned in the infobox but were mentioned somewhere in the body text. This information was meticulously and manually gathered. The total speaker counts for the Language GDP in Billions of Dollars (log) vs Wikipedia Article Count (log) analysis shown in Figure 4, as already mentioned in the main body text of this paper, were collected from the publicly available website *worlddata*[27] along with the corresponding information on GDP and percentage of language speakers of each country. The Ethnologue size (*Large*, *Mid*, and *Small*) as well as the Ethnologue Vitality (*Institutional*, *Sta-*

[18] https://www.ethnologue.com/browse/names
[19] https://bit.ly/Wikipedias_Details_table
[20] https://www.ethnologue.com/language/nor
[21] https://no.wikipedia.org/wiki/
[22] https://nn.wikipedia.org/wiki/
[23] https://bit.ly/AACL2022LangList
[24] https://bit.ly/Wikipedias_Details_table
[25] https://commoncrawl.github.io/cc-crawl-statistics/plots/language
[26] https://en.Wikipedia.org/wiki/Help:Infobox
[27] https://www.worlddata.info/

ble, *Endangered*, and *Extinct*) were of course, manually collected from Ethnologue. The language family information as well as the geographical origin of the languages were also collected from the Wikipedia infoboxes of the relevant languages. The count of ACL publications mentioning the relevant language was obtained executing the algorithm proposed by Blasi et al. (2022) on the full ACL text dataset published by Rohatgi (2022). The *Huggingface* dataset counts for both November 2021 and July 2022 were manually collected from the *Huggingface* dataset search web interface[28].

Facebook language list was manually extracted according to the instructions in their Help Centre web page[29]. The language list supported by Google was manually extracted from the Google Translate web page [30]. We selected the statistics in the 'Type' column'. Conneau et al. (2020) has reported the list of languages covered in XLm-R. mBERT language list was manually extracted from its github repository[31].

C CommonCrawl Analysis

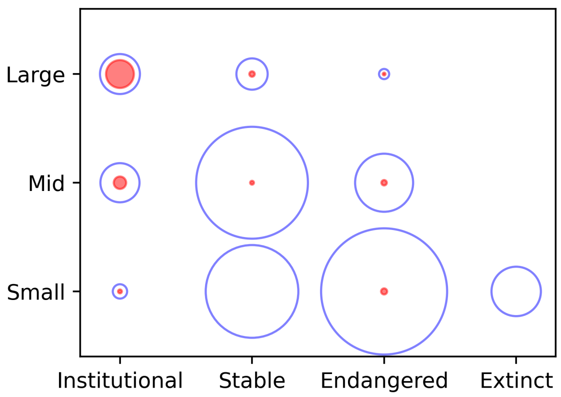


Figure 8: The 12 Ethnologue language classes where the size of each blue circle corresponds to the number of languages in that category and the size of each red circle corresponds to the coverage of that class in CommonCrawl.

As shown in Figure 8, CommonCrawl also covers mainly *large-institutional* and *mid-institutional* languages. Some language categories have no presence at all. Table 4 shows the gravity of this problem - out of the 160 languages present in CommonCrawl, 100 come from *large-institutional* category alone. Even *large-endangered* and *large-stable* categories do not have a significant presence in the web, despite a large population using those languages. This behaviour continues to Fig 9 where it can be observed that other than *Large-Institutional*, all other classes display a disappointing spread.

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

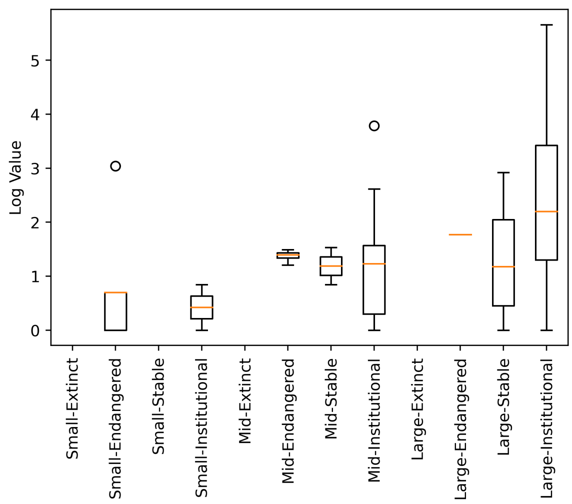


Figure 9: Boxplot showing CommonCrawl data with the amounts corresponding to the 12 Ethnologue language classes.

D Joshi et al. (2020)'s Class Descriptions

This is the language categorisation originally proposed by Joshi et al. (2020). Note that the number

[28]https://huggingface.co/datasets
[29]https://www.facebook.com/help/327850733950290
[30]https://translate.google.com/intl/en/about/languages/
[31]https://github.com/google-research/bert/blob/master/multilingual.md

Class	Description	Language Count	Language 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 labelled 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 unlabelled 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 5: Language Categories identified by Joshi et al. (2020)

of languages reported here are the numbers originally reported by them. This categorisation is done considering the number of Wikipedia pages and the total of ELRA and LDC datasets per language.

E Analysis of language Coverage in XLM-R and mBERT

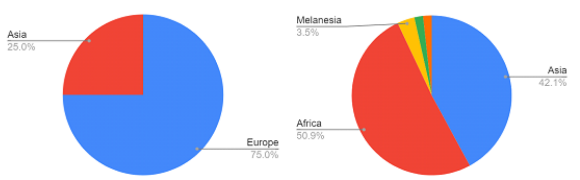


Figure 10: (a) Where the non-Large-Institutional languages included in XLM-R and mBERT models reside. (b) Where the Large-Institutional languages NOT included in XLM-R and mBERT reside.

F Wikipedia 12 Class Analysis

We conducted an analysis on the size of Wikipedias in each of the languages that have a Wikipedia in the relevant language. The first of the analysis, shown in Fig 12, shows the distribution of the languages belonging to the 12 Ethnologue language classes by the geographical origin of each of the languages. It is very important to note that, this means languages with colonial histories such as English, French, Spanish, Portuguese are counted for *Western Europe* and not for locations that they have colonised and displaced the local languages. The reason for this is to show the disparity of prevalence of languages on Wikipedia where all things equal and free in the sense that, any person with knowledge in an under represented language or otherwise may go and write articles at no cost. But it seems, that is not happening. Consider specially the case of *North America*, *South America*, *Australia and New Zealand*. When the colonial languages are taken off consideration from those areas and we look at the state of native languages, we see that they are being under utilised.

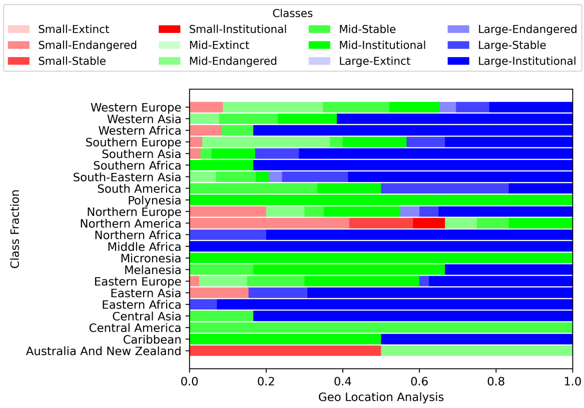


Figure 11: The distribution of languages that have wikis among the 12 Ethnologue Classes - By Geographical Location

The second analysis, shown in Figure 12, is similar to the first in set up but instead of geographical location, focuses on the language family. Most analysis done for language are commonly dominated by languages in the *Indo-European* family given the wide global spread that family of languages enjoy. In our analysis, we have taken that pressure off the other language families and tried to look at them in an equal footing. By doing this we make a number of interesting observations. The *Afro-Asiatic* group with contains *Arabic* and *Hebrew* seem to enjoy a spread skewed towards

Institutionally supported languages. The same pattern but with a slightly weaker bias can be observed from the *Dravidian* family of languages native to the southern part of India. We also note that the language families such as *Koreanic* and *Japonic* which carry only the eponymous languages also enjoying complete *Institutional* status.

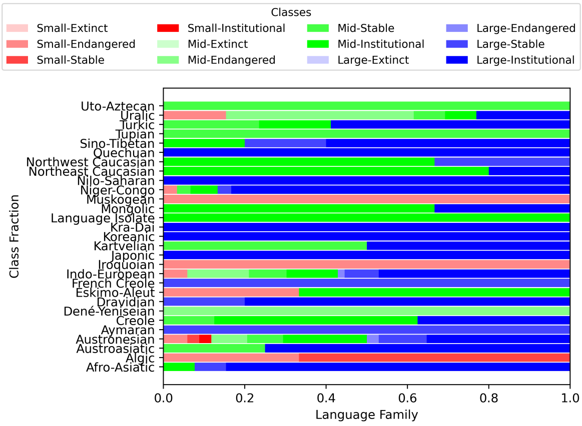


Figure 12: The distribution of languages that have wikis among the 12 Ethnologue Classes - By Language Families

These observations further re-enforce our earlier claims on the impact of resource distribution and support has on the ability of future research in a given language as Wikipedia is one of the most used language sources for NLP. Therefore, whose language has a seat at the Wikipedia table then partially influences, whose language gets a seat at the NLP research table. If we are to lift some of these languages out of resource and research poverty, starting it with building the relevant Wikipedia is a rational place to start given that it has a low barrier to entry and has an already established ecosystem with editor tools, translator tools, and most importantly collaborative community help.

G Impact of Population on the Wikipedia Article Count

We plotted the graph shown in Figure 13 and used Pearson correlation. As can be seen, the number of Wikipedia articles available has a *moderate correlation* (0.518789) to the population that speaks the language. 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. When a language is spoken as L1 in more than one geographical area, the numbers reported in Wikipadia are summed.

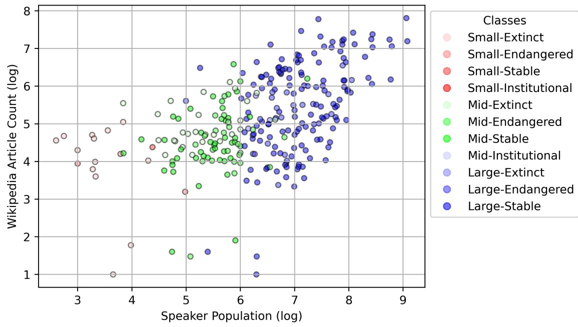


Figure 13: Speaker Population (log) vs Wikipedia Article Count (log).

H HuggingFace Datasets Task and Language Analysis

In Table 6 we show the datasets that are tagged with languages and tasks on HuggingFace classified to the Ethnologue language classes. From the get go, it is evident that all the languages are not represented. We observe that only 8 Ethnologue classes: *Large-Institutional*,*Large-Stable*, *Large-Endangered Mid-Institutional*, *Mid-Stable*, *Mid-Endangered*, *Small-Stable*,*Small-Endangered* have any data sets tagged with their member languages.

Even if we disregard *Large-Extinct* and *Mid-Extinct* which are missing in all other analyses, this still comes short for *Small-Institutional* and and *Small-Extinct*. On the other end, we note that the following 50 tasks has zero languages tagged on their data sets: *information-retrieval*, *zero-shot-retrieval*, *zero-shot-information-retrieval*, *time-series-forecasting*, *computer-vision*, *reasoning*, *paraphrasing*, *code-generation*, *tts*, *image*, *image-retrieval*, *image-captioning*, *text-generation-other-code-modeling*, *Code Generation*, *Translation*, *Text2Text generation*, *text-to-slide*, *paraphrase detection*, *Summarization*, *cross-language-transcription*, *grammatical error correction*, *named-entity-disambiguation*, *textual-entailment*, *natural-language-inference*, *query-paraphrasing*, *text-regression*, *entity-extraction*, *unpaired-image-to-image-translation*, *generative-modelling*, *Token Classification*, *caption-retrieval*, *gpt-3*, *crowd-sourced*, *sequence2sequence*, *Inclusive Language*, *Text Neutralization*, *super-resolution*, *image-enhancement*, *speech-synthesis*, *data-integration*, *Language-model*, *Automatic-Speech-Recognition*,

Task	Large-Institutional	Large-Stable	Large-Endangered	Mid-Institutional	Mid-Stable	Mid-Endangered	Small-Stable	Small-Endangered
translation	1579	12	1	123	17	39	2	20
text-classification	896	6	0	35	6	14	0	10
text-generation	687	6	0	52	12	18	1	8
fill-mask	597	6	0	50	12	18	1	8
token-classification	469	5	0	24	5	9	0	6
question-answering	487	3	0	5	0	0	0	1
conditional-text-generation	387	3	0	32	6	7	0	3
text-retrieval	179	3	0	7	0	5	0	2
text2text-generation	183	0	0	2	0	1	0	2
other	137	2	0	7	3	2	0	1
image-to-text	125	2	0	6	0	5	0	2
summarization	118	0	0	0	0	0	0	0
automatic-speech-recognition	101	0	0	7	1	1	0	0
multiple-choice	104	1	0	1	0	0	0	0
speech-processing	74	0	0	6	3	2	0	0
zero-shot-classification	59	0	0	0	0	0	0	0
table-question-answering	58	0	0	0	0	0	0	0
tabular-classification	57	0	0	0	0	0	0	0
audio-classification	45	0	0	4	0	1	0	0
sequence-modeling	36	0	0	2	0	0	0	0
structure-prediction	35	0	0	0	0	0	0	0
image-classification	25	0	0	0	0	0	0	0
conversational	16	0	0	0	0	0	0	0
sentence-similarity	13	0	0	0	0	0	0	0
tabular-to-text	12	0	0	0	0	0	0	0
table-to-text	10	0	0	0	0	0	0	0
paraphrase-mining	8	0	0	0	0	0	0	0
object-detection	7	0	0	0	0	0	0	0
text-scoring	7	0	0	0	0	0	0	0
commonsense reasoning	4	0	0	0	0	0	0	0
coreference resolution	4	0	0	0	0	0	0	0
sentiment-analysis	4	0	0	0	0	0	0	0
question-generation	4	0	0	0	0	0	0	0
image-to-image	3	0	0	0	0	0	0	0
text-to-image	3	0	0	0	0	0	0	0
email subject	3	0	0	0	0	0	0	0
one liner summary	3	0	0	0	0	0	0	0
topic modeling	3	0	0	0	0	0	0	0
symbolic-regression	3	0	0	0	0	0	0	0
text_classification	3	0	0	0	0	0	0	0
meeting title	3	0	0	0	0	0	0	0
visual-question-answering	3	0	0	0	0	0	0	0
machine-translation	3	0	0	0	0	0	0	0
text-mining	3	0	0	0	0	0	0	0
image-segmentation	3	0	0	0	0	0	0	0
classification	3	0	0	0	0	0	0	0
masked-auto-encoding	2	0	0	0	0	0	0	0
closed-domain-abstrative-qa	2	0	0	0	0	0	0	0
dialog-response-generation	2	0	0	0	0	0	0	0
extractive-qa	2	0	0	0	0	0	0	0
neural-machine-translation	2	0	0	0	0	0	0	0
rendered-language-modelling	2	0	0	0	0	0	0	0
abstractive-qa	2	0	0	0	0	0	0	0
language-modelling	2	0	0	0	0	0	0	0
long-texts	2	0	0	0	0	0	0	0
other-test	1	0	0	1	0	0	0	0
feature-extraction	2	0	0	0	0	0	0	0
other-text-to-structured	2	0	0	0	0	0	0	0
text-understanding	1	0	0	0	0	0	0	0
commonsense-reasoning	1	0	0	0	0	0	0	0
moral-reasoning	1	0	0	0	0	0	0	0
social-reasoning	1	0	0	0	0	0	0	0
style-transfer	1	0	0	0	0	0	0	0
task-dialogue	1	0	0	0	0	0	0	0
natural-language-understanding	1	0	0	0	0	0	0	0
text-comprehension	1	0	0	0	0	0	0	0
story-generation	1	0	0	0	0	0	0	0
natural-language-generation	1	0	0	0	0	0	0	0
data-to-text	1	0	0	0	0	0	0	0
MultiLabel Text Classification	1	0	0	0	0	0	0	0
commonsense-generation	1	0	0	0	0	0	0	0
sequence-modelling	1	0	0	0	0	0	0	0
open-dialogue	1	0	0	0	0	0	0	0
patents	1	0	0	0	0	0	0	0
deduplication	1	0	0	0	0	0	0	0
Information Retrieval	1	0	0	0	0	0	0	0
named-entity-recognition	1	0	0	0	0	0	0	0
simplification	1	0	0	0	0	0	0	0
video-captionning	1	0	0	0	0	0	0	0
text-generation-other-common-sense-inference	1	0	0	0	0	0	0	0
text-generation-other-discourse-analysis	1	0	0	0	0	0	0	0
other-text-to-tabular	1	0	0	0	0	0	0	0
other-text-search	1	0	0	0	0	0	0	0
question-pairing	1	0	0	0	0	0	0	0
Semantic Search	1	0	0	0	0	0	0	0
question_answering	1	0	0	0	0	0	0	0
Evaluation of language models	1	0	0	0	0	0	0	0
masked-language-modeling	1	0	0	0	0	0	0	0
multi-class classification	1	0	0	0	0	0	0	0
topic-classification	1	0	0	0	0	0	0	0
paraphrase	1	0	0	0	0	0	0	0
language-modeling	1	0	0	0	0	0	0	0
machine translation	1	0	0	0	0	0	0	0
text-to-speech	1	0	0	0	0	0	0	0
image-generation	1	0	0	0	0	0	0	0

Table 6: Datasets for different task-language category combinations (Excluding the 50 tasks that are not tagged with any language).

influence-attribution, *question-answering-retrieval*, *text*, *linear-regression*, *syntactic-evaluation*, *text classification*, *text tagging*, *named entity recognition*.

Now this does not imply that all of these are not text based tasks. Some of them, (e.g., *image*) may fall into that category. But some, (e.g., *Text Neutralization*, *Text2Text generation*) are ostensibly text based tasks. So is *Translation* which a variant capitalisation of *translation* which is the highest language tagged task. What we can say here, given how HuggingFace search gives the intersection of the labels, is that, this must be an artefact of how users tag their data sets on HuggingFace. It seems some users tag their task, but have not taken steps to tag the languages in their data set.

Therefore, it is vital that before using the HuggingFace tags for any meta-analysis on the NLP domain datasets, a large-scale data-clean up task be done on them. While the task still seem to be manually tractable, with the speed of growth shown by HuggingFace datasets, it is conceivable that it would soon cease to be so. Alternatively, it can be suggested to introduce a levelled tag system to HuggingFace where the top level tag is selected from a pre-set collection of tags set by HuggingFace while the lower level tag can be typed-in by the person who upload the data set.

I OPUS Data

We extracted the number of sentences available for each language listed in OPUS as shown in Table 7.

Language Class	Data Set Count
Large-Institutional	1.556114e+10
Large-Stable	3.216824e+07
Mid-Institutional	6.123440e+07
Mid-Stable	4.243600e+04
Mid-Endangered	7.833096e+06
Small-Institutional	1.104000e+03
Small-Stable	1.200500e+04
Small-Endangered	1.278468e+06
Small-Extinct	8.000000e+00

Table 7: OPUS Data Set Counts

J The Distribution of Resources

We have added larger versions of Fig 2a and Fig 2b at Fig 14 and Fig 15 respectively.

K Impact of using Huggingface as a Data Source

When *Huggingface* data sets were introduced, 87 languages changed their class. Out of this, 84 were promotions. The three demotions are Afrikaans, Bosnian, and Croatian. The full list of class changes are given below. The list header gives the *Ethnologue* language class followed by the Joshi et al. (2020) class shift in parenthesis. The cases where language classes are demoted are indicated by an “*” at the end of the list header.

- ***Large-Institutional* (1 → 2):** Akan, Albanian, Assamese, Bamanankan, Bikol, Burmese, Chichewa, Chuvash, Fulah, Ganda, Gujarati, Igbo, Javanese, Kannada, Kashmiri, Kinyarwanda, kurdish (kurmanji), Kyrgyz, Limburgish, Lingala, Maithili, Malagasy, Malayalam, Nepali, Quechua, Rundi, Sango, Shan, Shona, Sindhi, Sinhala, Somali, Southern Sotho, Swati, Tajik, Telugu, Tibetan, Tsonga, Turkmen, and Venda.
- ***Large-Stable* (1 → 2):** Aymara, Scots, Sicilian, and Sunda.
- ***Mid-Institutional* (1 → 2):** Abkhaz, Avar, Bislama, Chamorro, Dzongkha, Faroese, Fijian, Inuktitut, Luxembourgish, Ossetic, Romansh, Samoan, Scottish Gaelic, Tahitian, Yakut, and Yiddish.
- ***Mid-Stable* (1 → 2):** GuaranÃ.
- ***Mid-Endangered* (1 → 2):** Aragonese, Breton, Corsican, Maori, Navajo, Occitan, Sardinian, Udmurt, and Walloon.
- ***Small-Endangered* (1 → 2):** Cornish, Manx, and Pali.
- ***Large-Institutional* (1 → 3):** Armenian, Chechen, Esperanto, Macedonian, and Tatar.
- ***Mid-Institutional* (1 → 3):** Welsh.
- ***Large-Institutional* (1 → 4):** Azerbaijani.
- ***Large-Institutional* (3 → 2)*:** Afrikaans and Bosnian.
- ***Large-Institutional* (3 → 4):** Indonesian, Norwegian, Romanian and Ukrainian.
- ***Large-Institutional* (4 → 3)*:** Croatian.

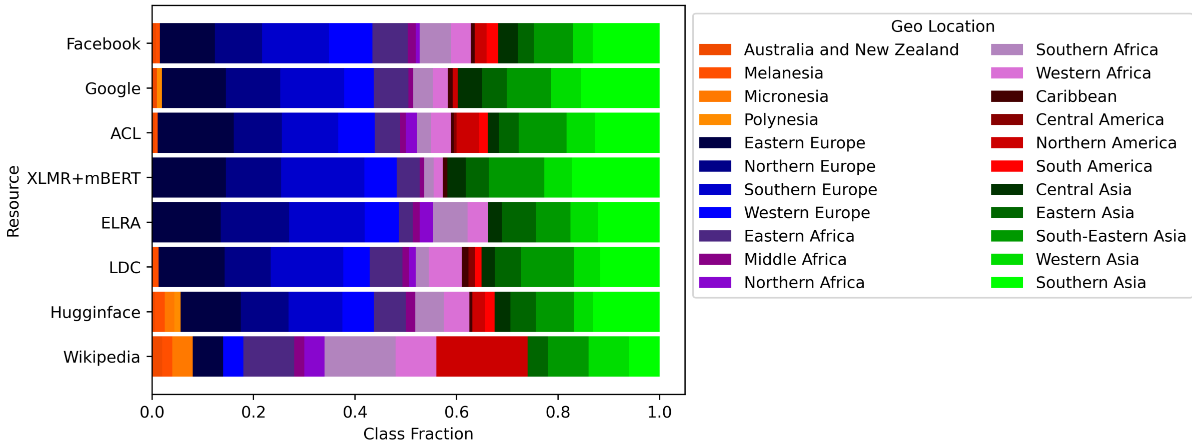


Figure 14: By Geographical Location of the Language Origin

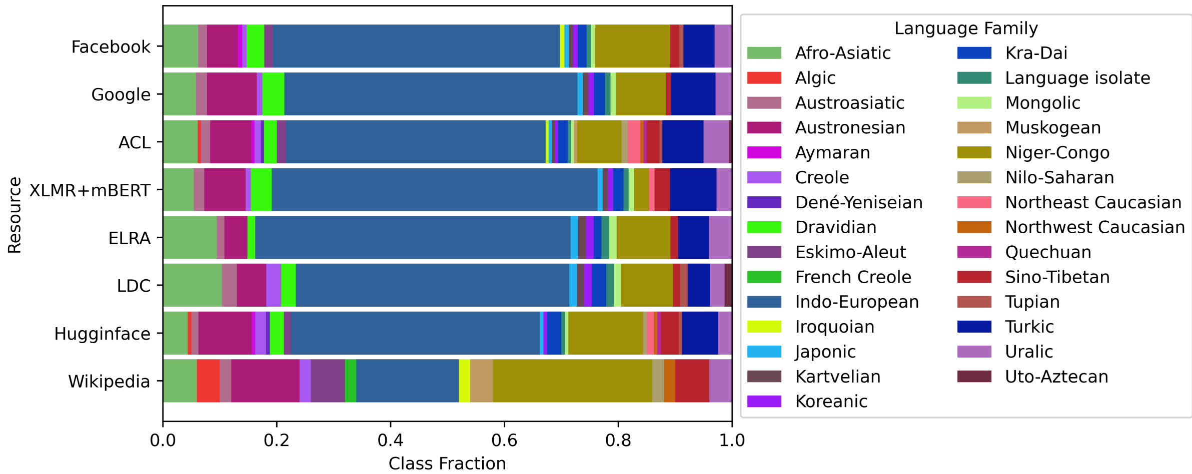


Figure 15: By Language Families

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 8: 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 9: 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*.

We show the confusion Matrix of Joshi et al. (2020) classes and the 12 Ethnologue language classes resluting when the Joshi et al. (2020) classes are derived only considering *LDC* and *ELRA* as the annotated sources in Table 8.

Then we show the same confusion Matrix but considering *Huggingface* in addition to *LDC* and *ELRA* as the annotated sources in Table 9. The information in Table 8 corresponds to Fig 5a while the information in Table 9 corresponds to Fig 5b. We can clearly see some of the promotions and demotions that we discussed above. One very easy to spot transition is the promotion of the three *Small-Endangered* languages: *Cornish*, *Manx*, and *Pali* from class 1 to class 2. Note how in the *Small-Endangered* column of Table 8, there are 15 languages in class 1 and 0 languages in class 2. Then in the *Small-Endangered* column of Table 9, there are 12 languages in class 1 and 3 languages in class 2 attesting to the promotion of the aforementioned languages.

L ACL Publication History and Performance

As shown in Figure16 (considering all the publications in ACL Anthology), there is a continuous increase of publications for all categories. There are some interesting observations here - (1) research on some language categories started much later than categories such as large-institutional and (2) the number of papers for large-institutional is less than some other categories. We believe this is the impact of workshops. As mentioned by Bender (2019), many research that focused on English did not bother to mention the language in the paper as it is assumed *de facto*.

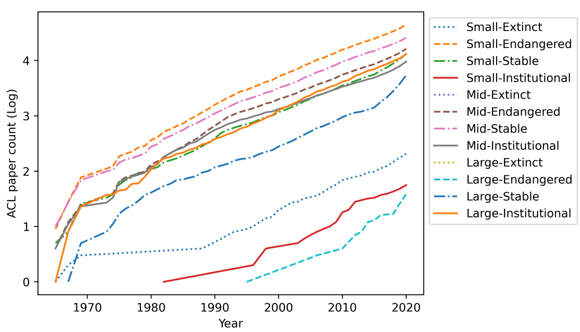


Figure 16: ACL publication count for the 12 Ethnologue language classes (cumulative log)

Figure 17 shows a breakdown of mentions in the abstracts of ACL Anthology publications. Here, *Main* venues include (1) Annual Meeting of the Association for Computational Linguistics, (2) North American Chapter of the Association for Computational Linguistics, (3) European Chapter of the Association for Computational Linguistics, (4) Empirical Methods in Natural Language Processing, (5) International Conference on Computational Linguistics, (6) Conference on Computational Natural Language Learning (7) International Workshop on Semantic Evaluation, (8) Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, and (9) Conference on Computational Natural Language Learning. *Journals* include (1) Transactions of the Association for Computational Linguistics and (2) Computational Linguistics. *Other* category means everything except the aforementioned conferences/journals and *LREC*. We have given *LREC* a separate category as it is a venue where a considerable amount of researchers in under-resourced languages target. This decision is especially justified by the observations in Fig 17j. It can be seen that despite the language category, most of the papers that mention a language name are in workshops. Interestingly, only *LREC* and *other* category has coverage for large-endangered languages.

M Analysis on Where NLP Researchers Publish their Datasets

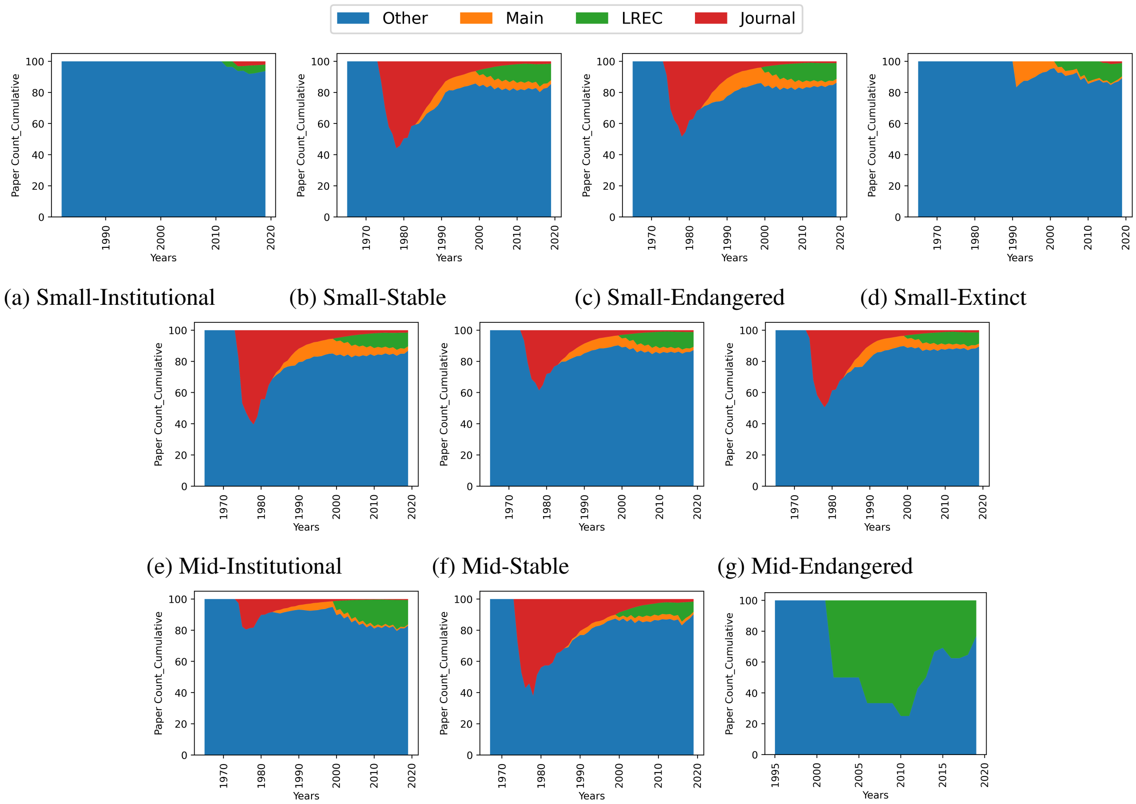
M.1 How the Analysis was Carried out

We first checked the Dataset section of each paper. If the paper has used a dataset, we recorded whether it is a new dataset presented in the paper. If so, we check whether the dataset has been published. We mainly checked the Abstract, Introduction Dataset and Conclusion sections to see if information related to dataset publishing has been given. If not, we do a search using keywords such as data, corpus, publicly, share, release, free and available. This analysis was manually carried out.

M.2 Dataset Publication Details

As mentioned above, we first identify whether a paper has created a new dataset. Then we note down whether the dataset has been released in any of the following forms:

- Via personal repository (github, personal web page, Google drive, etc)
- Via institutional repository (github, institutional website, etc). We also note whether the



(a) Small-Institutional (b) Small-Stable (c) Small-Endangered (d) Small-Extinct

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

(h) Large-Institutional (i) Large-Stable (j) Large-Endangered

Figure 17: ACL Abstract Participation of the languages belonging to the 12 Ethnologue language classes (Only the existing 10 classes shown here.)

dataset is available freely or based on request. In some papers, this is clearly mentioned. For others, we visited the corresponding website and checked.

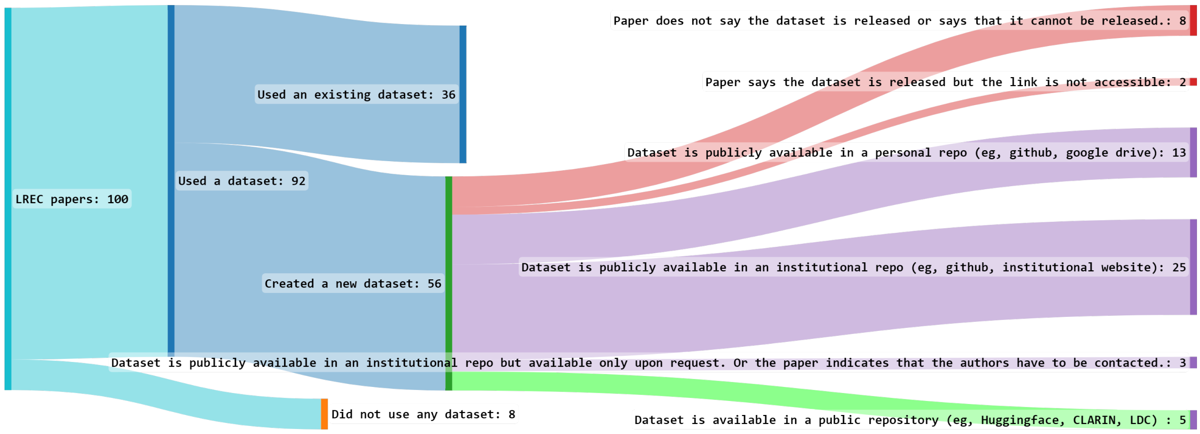
- via a public repository (ELRA, LDC, HuggingFace, CLARIN, etc)

If a link to any of the above has not been given, or if the paper explicitly mentions that the dataset cannot be released, we consider the dataset not released. Results are shown in Figure 18.

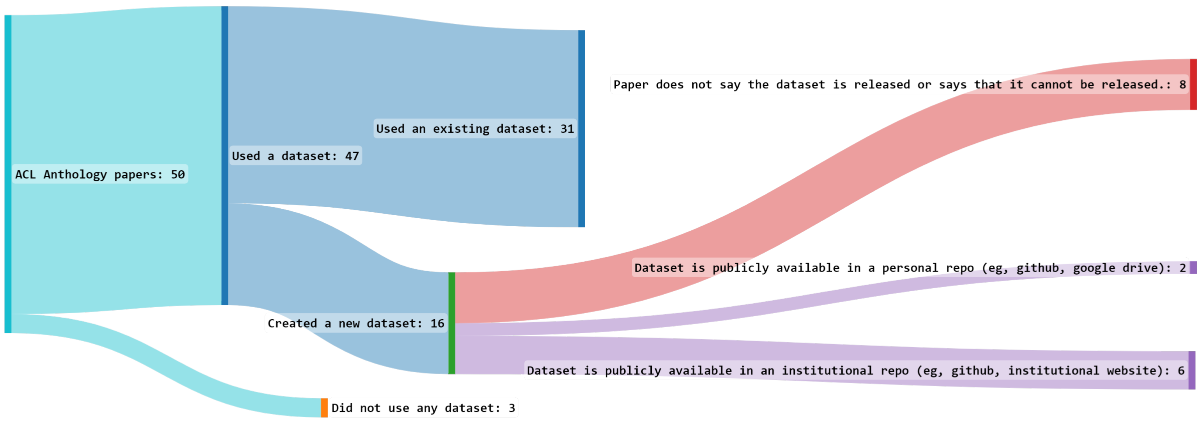
N Survey Results

Given below are the survey questions that we have used:

1. Have you ever kept a dataset you created ONLY in a private repo? Please select the most appropriate answer. (Results in Fig 19)
2. If your answer was ‘yes’ to the above question, please select all that applies. (Results in Fig 20)
3. Have you ever made your dataset conditionally available? (e.g. signing NDA, expected a request to release data). Please select the most appropriate answer.(Results in Fig 21)
4. If your answer was ‘yes’ to the above question, please select all that applies. (Results in Fig 22)
5. Have you ever publicly made your dataset available? Please select the most appropriate answer. (Results in Fig 23)
6. If yes, where did you publish your dataset? Please select all that applies. (Results in Fig 24)
7. If you have ever used a public repository (free or paid) to release data, what are they? select all that applies. (Results in Fig 25)
8. If you are not using data repositories such as Huggingface, Kaggle and OSF, what are the reasons for that? Please select all that applies (Results in Table 10)



(a) Analysis based on LREC papers



(b) Analysis based on ACL Anthology papers

Figure 18: Information of the use and release of data used in NLP research papers

9. Country that you are/were residing when you created most of your datasets (select the most relevant country) (Results in Fig 26)

Figure 19 shows a very positive trend - most researchers are releasing their dataset publicly. As per Figure 20, the main reason for not publicly releasing the data is the privacy concerns. This is understandable, as text corpora deals with information written by/about people and organizations. It is interesting to see that the second common reason for not releasing the dataset is the researcher not being confident about the dataset quality. This is a worrying situation, as the corresponding publication has already been made public and the claims in the paper may not be entirely correct.

In their meta-study on parallel language data sets, Kreutzer et al. (2022) did observe that even the publicly available datasets have various quality issues. In that light, when these two ideas are put together, the conclusions we can draw here become more dire. If we are to hypothesise that

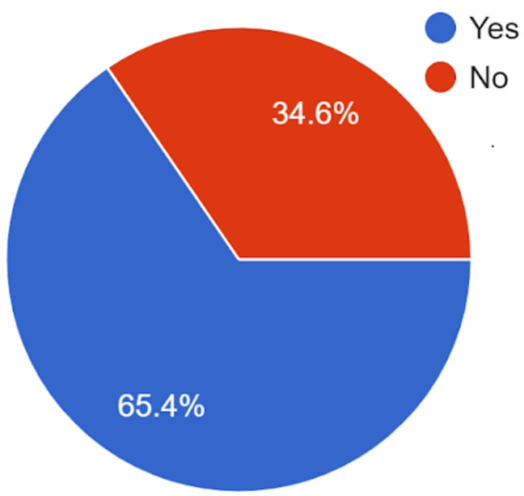


Figure 19: Distribution of researchers who published and did not publish data

the datasets that are released by the researchers that were confident of their data sets, and studies such as Kreutzer et al. (2022) find them lacking of quality, the work where the researchers themselves were not confident of the releasing data may be of highly questionable result. It is also worth encouraging researchers to publicly release their datasets, because some seem not to release the datasets just out of personal preference.

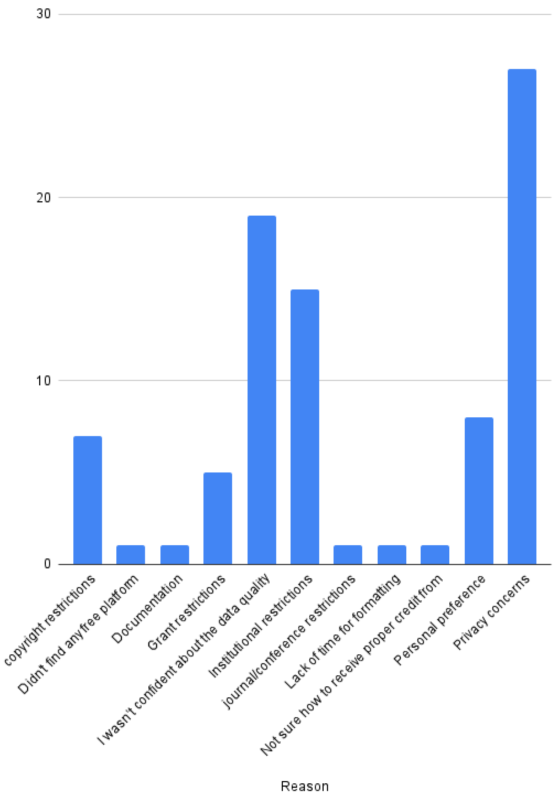


Figure 20: Reasons for not publishing the data

Conditionally releasing the datasets also has a similar trend (see Figure 21). Figure 22 indicates that the reasons for conditionally releasing the datasets follows a similar trend to that of not releasing datasets. Institutional restrictions is also notable. We believe this is due to the institution investing in the dataset, or the dataset adding a competitive advantage to the institution. de Silva (2021) also criticised the institutional barriers as a major reason for Sinhala NLP tools and data sets are not publicly shared. Our survey results in Figure 22 re-affirms this observation but in a more generic manner, by the self-admission of NLP researchers on a wide range of languages.

Figure 23 paints a very promising picture - about 90% of the researchers have made their data publicly available at some point of time. What varies is how they publish their datasets. According to

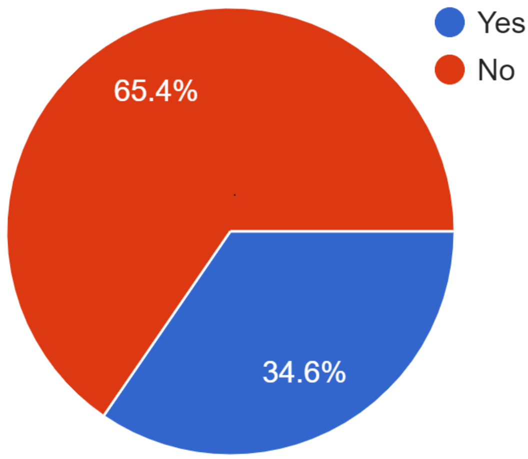


Figure 21: Openly released vs conditionally released

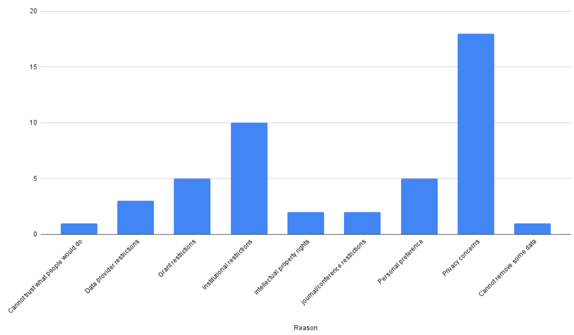


Figure 22: Reasons for conditionally releasing data

Figure 24, most of the researchers still prefer to release their datasets via their personal repository (e.g. Github repository of GoogleDrive). A considerable number released their datasets via their institutional repository, which could be due to institutional policies. It is worth noting that although it is lesser than those who release their data via their personal repositories, a decent number of researchers release their data via public repositories as well. This has a contradiction to what we found out by analysing LREC submissions, where only 9% of the papers have indicated that the dataset has been released via a public repository. We suspect that this is due to the researchers adding their datasets first to their personal repository, and then to the public repository after publishing their paper.

The next noticeable fact is number of options that are available to publicly release a dataset (see Figure 25). Out of the 15 possible repositories, HuggingFace has been the most famous choice- this justifies our selection of the same to explain the impact of data repository in determining the resourcefulness of a language. The other famous

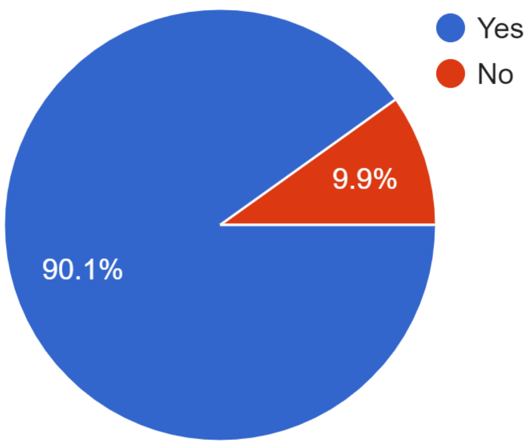


Figure 23: Distribution of researchers who made their data freely available

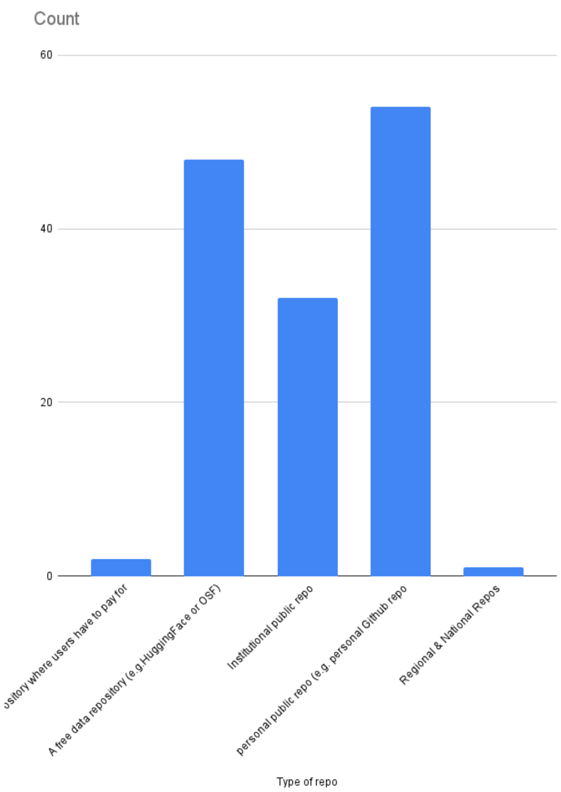


Figure 24: How datasets are publicly released

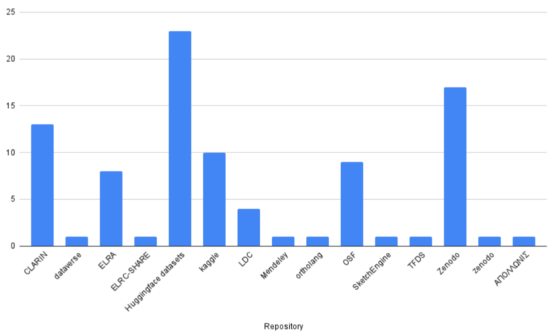


Figure 25: Where datasets are published

repositories are Zenodo, CLARIN, Kaggle and OSF (in the given order). Interestingly, ELRA and LDC, the two repositories selected by Joshi et al. (2020) are further down in the preference list.

In Table 10, we identify the reasons for researchers to not use the public repositories. It is surprising to see that there are several researchers who have not heard of such data repositories. A look into the individual responses did not indicate that these researchers belong to any particular geographical region. Given that there are 21 researchers who indicated that they cannot be bothered about adding data to public repositories, more awareness on the benefits of using public repositories should be carried out. Furthermore, availability of a repository that mitigates the limitations of the existing repositories would be a catalyst to encourage researchers.

Reason	Response Count
Accessing data through such repositories is difficult	5
Control: it's easy to modify if it's personal/institute	1
Data was already released via my personal/institutional repo. so I could not be bothered to publish into another repo	21
Repository is maintained by a private company interested in Machine Learning	2
I do not trust those repositories would last long	5
Some repositories do not issue DOI	1
I was not aware of such free data repositories	13
Such repos store older versions of datasets	1
Too many different repositories. Unsure where the data will be found by other researchers	1

Table 10: Reasons for not using public repositories

Similarly, on the other end, these replies may also help those organisations and non-profits who maintain public repositories to augment the way they approach researchers to utilise their services. Specifically note the complaint of accessing data through such repositories being difficult. This could be taken as a call to improve the user interfaces and the overall experience of the repositories. The doubt of some researchers on how long the repositories would last is also an interesting point in this perspective. It seems given the choice between the institute of the researcher and a public repository run by a third party, some researchers are not confident of the continued existence of the repository. Thus this is a call for the repositories to inform the researchers of their policies on what happens to the hosted datasets upon a possible cessation of operations. Providing the researchers of such assurances about reliability, accessibility, and longevity may incentivise them to consider public

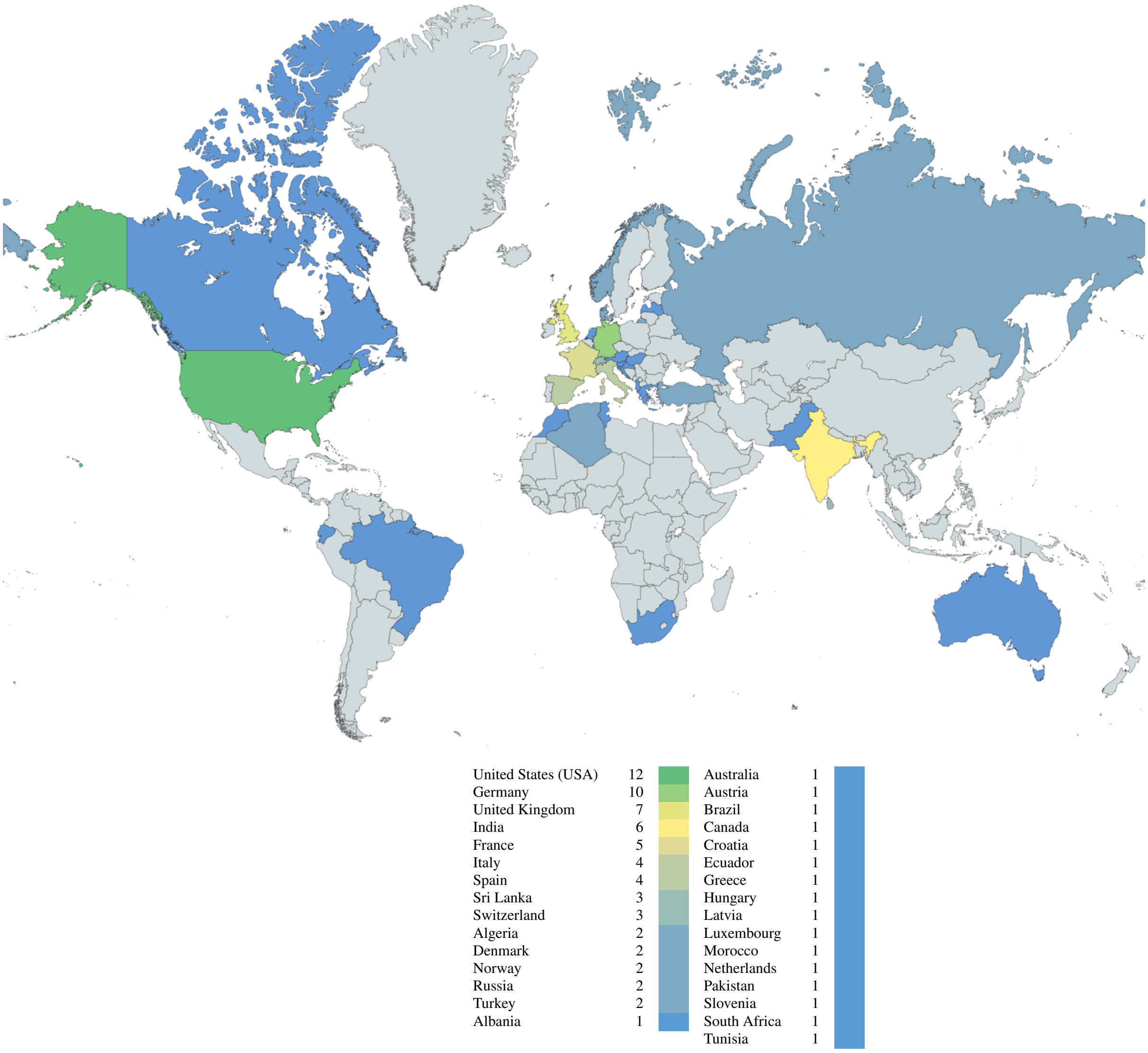


Figure 26: Countries at which the researchers who have uploaded their data sets have conducted their research

data repositories in the future.

We show where each of the respondents of our survey marked as the country that they were residing when they created most of their datasets in Figure 26. It is unsurprising that the highest number of respondents are from the United States of America. The fact that personal contacts of the authors were also sent the survey explains the relative high number Sri Lanka has in the results. However the mot noticeable absentee is East Asia including China where a large portion of human population is concentrated and a considerable amount of language research is done. This might be an indication that researchers from these areas are under represented in the public mailing lists and private interest groups to which we sent our survey. We can postulate that one reason may be that aforementioned public mailing lists and private interest groups to which we sent out survey use English as the operational language. The researchers from East Asia (especially China) may use insular lists and groups that operate in the local language. This previously unforeseen divide may stand in the way of collaborations in the NLP field.

O Language Resource Increase Over Time

Tables 11 and 12 record the number of annotated and unannotated (respectively) dataset increase from November 2021 to July 2022. The *Difference* column shows the growth in number and each of the normalised columns carries the value obtained by dividing the values in adjoining *count* column by the the number in the *count* column for the relevant class. Both tables show a similar trend, even after normalising to the class size - Large-institutional

Class		Nov 2021		Jul 2022		Difference	
Name	Count	Count	Normalised	Count	Normalised	Count	Normalised
Small-Extinct	332	0	0.00	0	0.00	0	0.00
Small-Endangered	2162	38	0.02	45	0.02	7	0.00
Small-Stable	1168	1	0.00	3	0.00	2	0.00
Small-Institutional	28	0	0.00	0	0.00	0	0.00
Mid-Extinct	0	0	0.00	0	0.00	0	0.00
Mid-Endangered	458	86	0.19	101	0.22	15	0.03
Mid-Stable	1700	24	0.01	34	0.02	10	0.01
Mid-Institutional	208	228	1.10	310	1.49	82	0.39
Large-Extinct	0	0	0.00	0	0.00	0	0.00
Large-Endangered	14	27	1.93	31	2.21	4	0.29
Large-Stable	133	51	0.38	76	0.57	25	0.19
Large-Institutional	217	3529	16.26	4140	19.08	611	2.82

Table 11: The number of datasets available in *Huggingface* for the 12 Ethnologue language classes in November 2021 compared with July 2022.

Class		Nov 2021		Jul 2022		Difference	
Name	Count	Count	Normalised	Count	Normalised	Count	Normalised
Small-Extinct	332	0	0.00	4176	12.58	4176	12.58
Small-Endangered	2162	3849	1.78	180106	83.31	176257	81.52
Small-Stable	1168	1036	0.89	2958	2.53	1922	1.65
Small-Institutional	28	0	0.00	2455	87.68	2455	87.68
Mid-Extinct	0	0	0.00	0	0.00	0	0.00
Mid-Endangered	458	18028	39.36	650903	1421.19	632875	1381.82
Mid-Stable	1700	8903	5.24	171688	100.99	162785	95.76
Mid-Institutional	208	366882	1763.86	1058393	5088.43	691511	3324.57
Large-Extinct	0	0	0.00	0	0.00	0	0.00
Large-Endangered	14	0	0.00	77070	5505.00	77070	5505.00
Large-Stable	133	22124	166.35	1085994	8165.37	1063870	7999.02
Large-Institutional	217	1243317	5729.57	54612595	251670.94	53369278	245941.37

Table 12: The number of datasets available in *Wikipedia* for the 12 Ethnologue language classes in November 2021 compared with July 2022.

category has been added with more data. Similarly, the extinct languages seem to be forever forgotten. Annotated dataset count for Mid-institutional languages have increased by a noticeable number. On the other hand, focus on 'small' languages is negligible, if not zero. This trend of rich getting richer is a cause for concern for those who are interested in developing and using data sets to and from low-resourced languages as this shows that the average interest still lies with the few languages that are already enjoying an abundance of datasets.

In contrast, most categories show a growth in Wikipedia article counts. Particularly of interest is the mid-endangered category, which has a noticeable gain. This hints at some community efforts to increase the digital content for these languages that took place recently. As observed by Hoenen and Rahn (2021), some members of the communities of endangered languages have taken to Wikipedia as a means of conserving traditional knowledge, and oral traditions in the source language.