Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5035–5046
Marseille, 20-25 June 2022

Writing System and Speaker Metadata for 2,800+ Language Varieties

Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, Clara Rivera
Google Research
1600 Amphitheatre Parkway, Mountain View, CA 94043, USA
{dvanesch,tlucassen,ruder,icaswell,rivera}@google.com

Abstract

We describe an open-source dataset providing metadata for about 2,800 language varieties used in the world today. Specifically, the dataset provides the attested writing system(s) for each of these 2,800+ varieties, as well as an estimated speaker count for each variety. This dataset was developed through internal research and has been used for analyses around language technologies. This is the largest publicly-available, machine-readable resource with writing system and speaker information for the world's languages. We analyze the distribution of languages and writing systems in our data and compare it to their representation in current NLP. We hope the availability of this data will catalyze research in under-represented languages.

Keywords: multilingual, low-resource, natural language processing

1. Introduction

Today, language technologies are easily available in only a small minority of the world's 7,000+ language varieties. For example, technologies like automatic speech recognition and neural machine translation are available from commercial vendors in about 100 language varieties; even keyboards and spell-checkers, which are relatively straightforward to develop, are only available in about 1,000–1,500 varieties (Mager et al., 2018; van Esch et al., 2019; Kuhn et al., 2020).

Fortunately, the last few years have seen an impressive amount of activity towards making language technologies available in more languages, including community-driven initiatives (Ardila et al., 2020; ∀ et al., 2020; Mager et al., 2021; Mirzakhalov, 2021). At the same time, advances in self-supervised pre-training for text (Howard and Ruder, 2018; Devlin et al., 2019) and speech (Baevski et al., 2020) have led to the development of pretrained multilingual models (Conneau et al., 2020; Chung et al., 2021; Babu et al., 2021), significantly reducing the barriers to extending language technologies to new languages (Pfeiffer et al., 2020; Muller et al., 2021). Such technologies are supported by a blossoming ecosystem of easy-to-use open-source tooling and libraries, such as Hugging Face[1], but also more specialized tools like KeyMan[2] for keyboard development; the machine translation toolkit JoeyNMT (Kreutzer et al., 2019); and the speech recognition toolkit Elpis (Foley et al., 2018).

A challenge facing the expansion of language technology to more languages is a lack of easily accessible, machine-readable metadata about all language varieties of interest. While there are a number of available resources such as Ethnologue (Eberhard et al., 2021) and Glottolog (Hammarström et al., 2021) that provide information about the world's languages, these resources are either not publicly accessible (Ethnologue) or do not provide information about speaker numbers nor writing systems (Glottolog). However, speaker population estimates are important as they can be used to prioritize languages in order to maximize the benefit language technology confers to users (Blasi et al., 2021). Information about writing systems is crucial not only for product decisions like keyboard layouts, but also for modeling, as pre-trained models have been shown to be affected by a language's script (Muller et al., 2021; Rust et al., 2021; Pfeiffer et al., 2021).

We have developed an open dataset that provides such information.[3] To our knowledge, it is the largest publicly available resource that provides detailed writing system, speaker information, and endonyms for a large number of the world's languages (see Table 1). We believe that the vast majority of languages with more than 10K speakers and an ISO 639-3 language code is included in our dataset. Languages with fewer than 10K speakers for which we were able to determine an attested writing system are also included. As an illustration, Table 2 provides an excerpt.

Our analysis of the distribution of writing systems and languages in the data identifies 112 unique writing systems overall, with 54 being used by more than one language. Latin is the most common script and used by 2,316 languages as the sole writing system. The other most common scripts are Devanagari, Arabic, and

	# of language varieties	Speaker data	Writing system data	Open-source
Wikipedia list	100	✓	✗	✓
ISO 639-3	7,893	✗	✗	✓
Glottolog	8,549	✗	✗	✓
Ethnologue	7,459	✓	✗	✗
WALS	2,662	✗	✗	✓
Ours	2,831	✓	✓	✓

Table 1: Number of languages and information available in existing language resources compared to ours.

[1] https://huggingface.co/
[2] https://keyman.com/
[3] The data set is available at https://github.com/google-research/url-nlp.

ISO 639-3	BCP 47	Speakers (rounded)	Writing system	Name	Glottocode	Region (ISO 3166)
act	act	10K-100K	latn	Achterhoeks	acht1238	NL, DE
xon	xon	900,000	latn	Konkomba	konk1269	GH, TG
bsq	bsq	400,000	bass,latn	Bassa	nucl1418	LR, SL
ind	id	200,000,000	latn	Indonesian	indo1316	ID, NL, PH, SA, SG, US
wbp	wbp	<10K	latn	Warlpiri	warl1254	AU

Table 2: An excerpt from our data set, showing key information per language, including speaker counts and attested writing systems. Speaker counts are rounded and approximate, as described below. The writing systems we identified are given using ISO 15924 codes. Name and region metadata is pulled in from Glottolog.

Cyrillic, which are used by around 100 languages each. We furthermore identify around 2,000 languages with more than 10K speakers, around 1,200 languages with more than 100K, and around 400 languages with more than 1M speakers.

To demonstrate the kind of language technology analysis the data can enable, we compare the estimates of speaker numbers and writing systems in our data with their representation in current NLP, specifically to the scripts represented in vocabularies of pre-trained multilingual models and languages referenced in published NLP papers. In vocabularies, we observe that CJK and Indian writing systems are under-represented compared to less common scripts such as Hebrew or Greek. In NLP papers, we observe outsized research contributions to certain European languages as well as Inuktitut, Hawaiian, Faroese, and others.

The dataset is undoubtedly incomplete: certainly there will be more languages for which some writing system is attested that we were not able to identify. In addition, while we fully recognize the importance of technology to support revitalization or historical research, our work has been focused on cataloging living languages. We have also not been able to include sign languages in our survey at this time, but hope the dataset can be extended in the future. However, even with these limitations, we hope this data can help inform the expansion of language technology to new languages and catalyze further research.

2. Background and Related Work

Resources with language metadata There are several existing resources that provide information about the world's languages. Perhaps best known is Ethnologue (Eberhard et al., 2021), an annual publication that provides information about the number of speakers, locations, dialects, and endangerment status covering 7,139 living languages (7,459 in total).

Another resource is Glottolog (Hammarström et al., 2021), which provides information about "languoids" (Good and Hendryx-Parker, 2006), basically languages, dialects, and language families.[4] Each languoid is assigned a unique identifier, the Glottocode. For each languoid, Glottolog provides its genealogical classification, endangerment status, and a comprehensive collection of bibliographical data of descriptive work including grammars, dictionaries, word lists, texts, and so on.

The online World Atlas of Language Structures or WALS (Dryer and Haspelmath, 2013) is a large database of structural linguistic features published under Creative Commons license. It provides the geographical distribution of phonological, grammatical and lexical properties of 2,662 languages but lacks speaker numbers and writing systems information.

While it is not a dedicated resource specific to languages, the English-language Wikipedia provides a list of 44 languages with 40 million or more total speakers based on Ethnologue.[5] It additionally provides lists of languages by number of native speakers based on data from Ethnologue (for the top 91) and from the 2007 edition of the Swedish encyclopedia *Nationalencyklopedin* (for the top 100).[6]

Finally, the ISO 639-3 code set contains 7,893 three-letter codes.[7] It also categorizes each entry as 'Living', 'Historical', 'Ancient', 'Extinct', or 'Constructed'.

Language diversity in NLP Our work is closely aligned with the goal of increasing language diversity and representation in NLP research, which has recently received renewed attention (Bender, 2011; Joshi et al., 2020). Progress in this area has been driven by the release of new multilingual datasets and benchmarks including multilingual unlabeled (Xue et al., 2021) and labeled datasets (Goyal et al., 2021) spanning many languages, as well as benchmarks for under-represented languages such as Indonesian and African languages (Cahyawijaya et al., 2021; Adelani et al., 2021). At the same time, state-of-the-art pre-trained multilingual models (Conneau et al., 2020; Chung et al., 2021) cover around 100 different languages. These models, however, have been shown to perform poorly on languages with limited amount of pre-training data and on languages with non-Latin scripts (Hu et al., 2020; Pfeiffer et al., 2021; Muller et al., 2021; Bhat-

[4] https://glottolog.org/

[5] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

[6] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

[7] https://iso639-3.sil.org/code_tables/639/data

tacharjee et al., 2021). The absence of natural multi-script data during training requires NLP practitioners to use transliteration resources such as the indic-trans (Bhat et al., 2014) or transliterate[8] libraries, or the manually curated Dakshina dataset (Roark et al., 2020). Such resources are scarce and not available in the majority of under-represented languages, hindering the development of technologies in these languages.

3. Development Process

Our need for a dataset like this can be traced back to our initiative to make Gboard, Google's Android keyboard app, available in many more languages. As described in van Esch et al. (2019), a key challenge we faced was understanding which languages have an attested written presence online, and for those that do, what writing system(s) are in use. In addition, speaker population numbers were hard to come by in a centralized format outside of proprietary databases.

3.1. Writing Systems

Initially, our efforts focused on establishing what writing systems are in use for each of the world's languages. We developed a tool, described in Prasad et al. (2018), to analyze online texts found on websites such as Wikipedia, the Wikimedia Incubator[9], An Crúbadán (Scannell, 2007), JW.org, and PanLex[10]. Crucially, these all have language labels attached to their content, with e.g. Wikipedia subdomains indicating whether a given Wikipedia article is part of its English-language edition or its Dutch-language edition. We converted all language codes encountered as part of our analyses to the ISO 639-3 standard.

As described in Prasad et al. (2018) and Chua et al. (2018), we automatically computed various statistics about the characters, words, and n-grams available in these sources for each language. This gave us a good picture of what writing system(s) were attested for each language found in these sources. We then followed up with human review for each language to determine whether the writing systems that had been automatically determined were in fact correct, verifying the content on the text sources themselves, but also reviewing the academic literature for each language, e.g. as identified by Glottolog. This review effort was part of a larger program, in which our in-house linguists also updated our normalization rules for each language (Chua et al., 2018; Zupon et al., 2021) and designed the keyboard layouts for the 900+ language varieties supported in Gboard today (Breiner et al., 2019). We cataloged the writing system(s) for each language using the ISO 15924 standard.

Later on, we expanded our automatic and manual orthographic analyses to some additional highly multilingual sources, including Unilex[11], CorpusCrawler[12], Bible.is, and the LTI corpus for LangID[13] (Brown, 2014). We also manually added metadata on a number of languages where we found information on attested writing systems in academic publications, e.g. in research referenced on Glottolog (Hammarström et al., 2021), in a database on Indigenous language publishing[14] (Gref, 2016), and so on. We also consulted online resources like Omniglot[15], Sorosoro[16], the *Systèmes alphabétiques des langues africaines*[17], and the PanAfrican Localisation Resource Wiki[18].

While we performed a reasonably comprehensive search, we know there are more sources we have not yet analyzed. Most importantly, we did not have time to investigate deeply the repositories that archive linguistic fieldwork data—which by now probably cover more than half the world's languages (Seifart et al., 2018). DELAMAN[19] and OLAC[20] provide information on how to access to these resources. We believe analyzing these sources would probably uncover evidence of a written tradition existing for even more languages.

3.2. Speaker Population Estimates

At that point, we moved on from analyzing the orthographic situation of the world's languages and turned our attention to speaker population estimates. To help us design our expansion roadmap for Gboard, however, we also needed to know rough numbers of speakers—of course, ideally we would like to bring support to every language at the same time, but in practice this is never feasible due to resource constraints, and we wanted to maximize our impact, as also described in van Esch et al. (2019). Therefore, we attempted to attach estimated speaker population numbers for each language with orthographic metadata.

3.2.1. Creating Estimates

To create these estimates, we gathered public statistics from many sources, including national census reports, academic reference grammars, or simply the article for a given language on the English-language Wikipedia. We then manually corrected and edited where needed: for example, for some languages we had to add up speaker counts from different national censuses (e.g.

[8] https://pypi.org/project/transliterate/

[9] https://incubator.wikimedia.org/

[10] https://panlex.org/

[11] https://github.com/unicode-org/unilex

[12] https://github.com/google/corpuscrawler

[13] https://www.cs.cmu.edu/~ralf/langid.html

[14] https://emilygref.com/publishing-database/

[15] https://omniglot.com/

[16] https://sorosoro.org/

[17] https://llacan.cnrs.fr/phono/

[18] http://www.bisharat.net/wikidoc/pmwiki.php

[19] https://www.delaman.org/

[20] http://olac.ldc.upenn.edu/

Dutch in the Netherlands and in Belgium), for others we had to de-duplicate statistics that would otherwise yield an over-count, and so on. We did our best to arrive at reasonable numbers, but acknowledge that efforts to estimate the speaker population of any given language necessarily involve some amount of arbitrariness: as Good and Hendryx-Parker (2006) also describe, decisions made when cataloging languages are often contestable, because it is hard to determine based on any specific objective linguistic grounds how to group language varieties together. Still, we are confident that our dataset provides a reasonably good approximation, and is useful for analyzing broad trends and for language policy and planning.

3.2.2. Arbitrariness and Rounding

To mitigate this arbitrariness somewhat, we bucketed the speaker population estimates: for languages where our estimate was above 100M, we rounded to the nearest 10M; for languages where our estimate was between 10M and 100M, we rounded to the nearest 1M; and so on. This is in order to avoid giving a false sense of precision: for example, a source may indicate that a given language has 5,429,310 speakers, but we believe that such statistics are simply not knowable at such precision, and they are also bound to change very regularly (with new speakers learning a given language, other speakers passing away, and so on). In other words, we found it important to do some sort of rounding, so that all numbers in similar ranges have similar levels of precision, and to avoid a false sense of exactness.

3.2.3. L1, L2, and Multilingualism

Another factor to be aware of is that we attempted to include only information about the estimated number of first-language (L1) speakers. For example, English is spoken by many L1 speakers—e.g. in the UK, the US, Australia, and so on—but it has hundreds of millions of second-language (L2) speakers across the globe as well. Since L2 speaker statistics are even harder to come by than L1 speaker statistics, we decided to limit ourselves to L1 statistics only. This means that the speaker counts for English under-estimate the true number of proficient speakers—as is the case for other languages with similar usage profiles, e.g. Indonesian, which is also widely used as a lingua franca.

To be fair, even the L1 speaker statistics present their own complications: for example, in multilingual societies, there may not be a clear single L1 that can be assigned to a given speaker, and for such speakers it may indeed make sense to account for two L1s. Unfortunately, multilingualism is virtually impossible to disentangle based on most sources.

Future work could try to add L2 speaker counts, but since L2 statistics are rarely tracked centrally (e.g. by censuses), and since proficiency levels of L2 speakers can vary widely, providing L2 information will be challenging to do reliably and consistently. In practice, we believe that L1 speaker information alone is sufficient for most use cases: first, large lingua francas like English and Indonesian will still rank highly even based on the number of L1 speakers alone, and second, at least in our view, language technology should ideally work for people without having to resort to their L2.

3.2.4. Script proficiency

The speaker estimates given in our dataset reflect total speaker numbers, and are presented alongside an overview of the attested scripts for each language variety. Unfortunately, we were unable to estimate what percentage of speakers of a given language would be a proficient user of each of the writing systems in use for a language. We have been unable to find reliable sources that would let us account for language users who are proficient speakers and listeners, but who do not have reading and writing proficiency—either as L1 speakers due to literacy reasons or as L2 speakers.

This discrepancy between language and script user estimates is also relevant for languages with multiple writing systems attested, where one writing system may be in common use, such as the Latin script for the Javanese language, or Simplified Chinese characters for Mandarin, while another is in much less common use—e.g. the Javanese script for Javanese, or in a less extreme example, Traditional Chinese characters for Mandarin.

3.3. Final Preparation Steps

Once we had gathered information on the writing systems used for as many languages as possible, as well as the speaker count estimates, we took a number of steps to clean up and verify the dataset. For example, we scanned the dataset for any language codes marked as deprecated or spurious in the ISO 639-3 registry[21] or by the English-language Wikipedia[22]. We did find some of these codes in our dataset, since they had been present in some of the upstream sources we had used. We removed these entries from our dataset as needed.

3.3.1. Macrolanguages

Another complication was dealing with macrolanguages: the ISO 639-3 standard defines some language codes as being 'umbrella' codes that cover multiple varieties, each with their own ISO 639-3 code assigned as well. For example, the Akan language (spoken in Ghana) has ISO 639-3 code AKA assigned, which includes Fante FAT and Twi TWI. For our purposes, it would not make sense to have three entries: either there should be one entry for Akan, covering the aggregate speaker estimate, or there should be two independent entries, one each for Fante and Twi, each with the respective speaker estimates.

The choice of whether to include a macrolanguage code or the independent codes is a difficult one to make. The orthographic system in use tends to be the same

[21] https://iso639-3.sil.org/code_tables/deprecated_codes/data

[22] https://en.wikipedia.org/wiki/Spurious_languages

between a macrolanguage and the individual language codes grouped underneath it, but this is naturally not the case for speaker numbers. We have tended to use macrolanguage codes in order to align with other resources—e.g. listing Arabic as one, not many different varieties—but certainly there is room for future work to expand the dataset with additional detail here.
In addition to the concept of macrolanguages as used in ISO 639-3, there are also three-letter 'collective' language codes, which form part of the ISO 639-2 standard. For example, some sources contain text that is indicated to be in WEN, which is the collective ISO 639-2 code for the Sorbian languages, with ISO 639-3 codes DSB for Lower Sorbian and HSB for Upper Sorbian. We have not used such collective codes in our dataset.

3.3.2. Glottocodes and Names

Finally, we mapped our dataset to Glottocodes as used in Glottolog (Hammarström et al., 2021), which we then used to pull in language names from Glottolog. These names may not always be the preferred names for a given language; we provide them merely for convenience. Glottolog itself contains a rich collection of variant names which can be looked up using the Glottocodes. For convenience, we do include some very common variant names in a separate column in our dataset. Where the endonym for a given language—the name for the language in the language itself—is known to us, we have also included it in an additional column.

3.4. Languages not covered

Our final dataset covers about 2,800 language varieties. While this is the largest dataset we are aware of with machine-readable, publicly accessible metadata on writing systems and speaker estimates, this still means there are about 5,000 languages with ISO 639-3 codes, which are not yet covered. Out of these languages that are missing from our dataset, about 85% are marked as 'Living' in the ISO 639-3 code tables. Based on a spot-check of these 'Living' languages that are *not* included in our dataset, we believe our dataset covers the vast majority of language varieties with more than 10K speakers. This also means that based on the evidence we have seen, most languages with more than 10K speakers have at least some attested writing system. To be fair, we have also come across a few languages, which we believe to have more than 10K speakers, but which do not have any attested script, as best we could tell; see examples in Table 3.

ISO 639-3 code	Name from Glottolog
juy	Juray
kxk	Lahta-Zayein Karen
mvi	Miyako
rys	Yaeyama
yix	Axi Yi
onb	Western Ong-Be
ycl	Lolopo
ywt	Xishanba Lalo
cda	Choni
mvz	Mesqan
byo	Biyo

Table 3: Some languages we believe to have more than 10,000 speakers but without attested writing system.

4. Analysis

4.1. Writing systems

We identify 112 unique writing systems among the 2,831 languages in our data. We highlight the ten most common writing systems in Figure 1. 2,554 languages use the Latin script as one of their writing systems, and 2,316 languages use the Latin script as their sole writing system. After the Latin script, the most common scripts are the Devanagari, Arabic, and Cyrillic scripts. 54 writing systems are used by more than one language. We show the number of languages that use these writing systems in Figure 2. Beyond the most common writing systems, there is a relatively long tail of scripts that are used by a smaller number of languages. We show the detailed number of languages and names of writing systems in Appendix A.1.
We show the writing systems based on the combined number of speakers across languages using that script in Figure 3.[23] We note that determining the speakers using a given writing system is inherently inaccurate as speakers may use different writing systems to different degrees and varieties of a language may use different writing systems. In addition, speakers may also be bilingual or multilingual, which complicates quantifying the number of unique speakers per writing system. The numbers here should thus be taken as an approximation of the world's most commonly used writing systems. Compared to Figure 1, there are several scripts that are not used by many languages but that have a large number of users—for example, Han characters, the main script used to write Mandarin and other Chinese varieties as well as many scripts common in India.

4.2. Number of speakers

We analyze the number of speakers per language in Figure 4. In our dataset, we find around 2,000 languages with more than 10K speakers; around 1,200 languages with more than 100K; around 400 languages with more than 1M; and around 100 languages with more than 10M speakers.

5. Representation in Current NLP

We now show some examples of the kind of language-technology analysis that our dataset enables. Specifically, in the following, we compare the estimates of speaker numbers and writing systems with their representation in current NLP. We analyze the vocabulary

[23] We exclude rarely used English-specific phonemic scripts (Deseret, Shavian), historically used scripts (Mahajani), and transliteration scripts (Bopomofo).

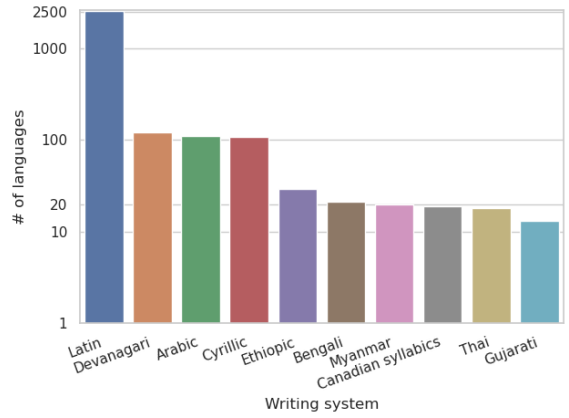

Figure 1: The ten most common writing systems in our data.

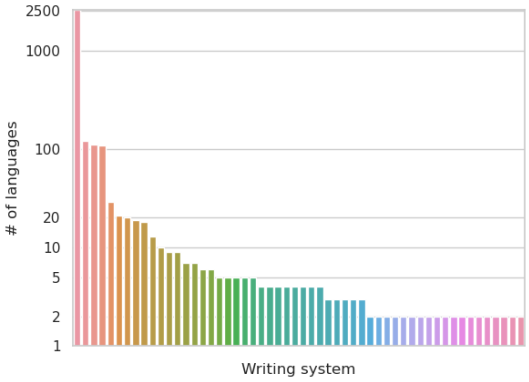

Figure 2: The 54 writing systems with more than one language in our data.

of pre-trained multilingual models as well as the languages mentioned in papers at NLP conferences.

5.1. Pre-trained Models

We analyze the representation of common writing systems in the vocabulary of state-of-the-art pre-trained multilingual models and how it compares to the distribution of writing systems across the world's languages. We analyze three representative state-of-the-art models: multilingual BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and multilingual T5 (Xue et al., 2021). Following Ács (2019), we group Unicode ranges[24] into writing systems and match the subwords in each model's vocabulary to each script.[25]

We show the vocabulary coverage of the top ten most common scripts across the three models in Figure 5. Most of the vocabulary of current models is dedicated to the Latin script. Compared to the number of speakers using a given system (see Table 3), Cyrillic is relatively over-represented while CJK languages are under-represented. Other commonly used scripts such as Devanagari, Bengali, and Gurmukhi are also under-represented in pre-trained models' vocabularies while less common scripts such as Hebrew (6M speakers), Armenian (6M), or Greek (15M) are relatively over-represented. We provide the detailed results for each model in Appendix A.2.

[24] https://www.ling.upenn.edu/courses/Spring_2003/ling538/UnicodeRanges.html

[25] We group Chinese (Han), Japanese, and Korean into a single CJK category.

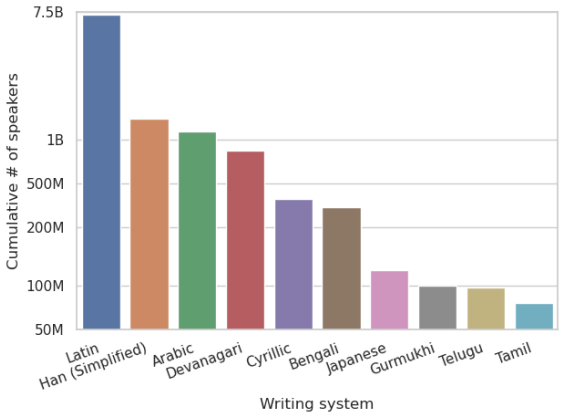

Figure 3: The ten writing systems based on the combined number of speakers across all languages using that writing system.

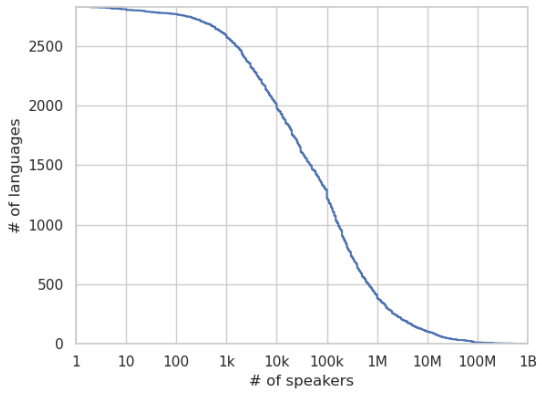

Figure 4: Empirical complementary cumulative distribution function (CDF) of speakers per language. The figure shows the number of languages that have at least N speakers in our data.

5.2. NLP Literature

Following (Joshi et al., 2020), we compile papers in the ACL Anthology (main conferences and workshops) to count the distribution of published works that reference the languages in our data. We first determine the 10 languages with the highest paper count per capita for the period of 2000–2020.[26] We show the number of

[26] We exclude several under-represented languages whose names are more commonly used in other contexts such as

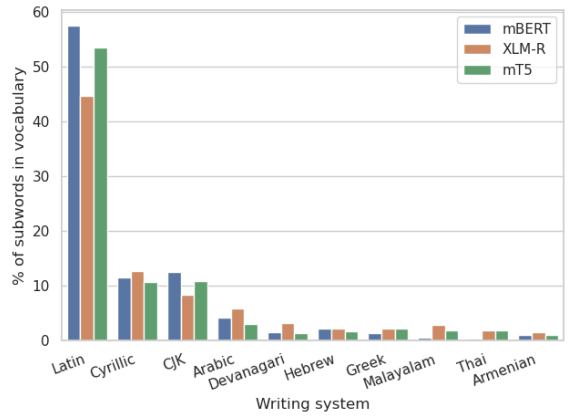

Figure 5: Representation of common scripts in pre-trained multilingual models.

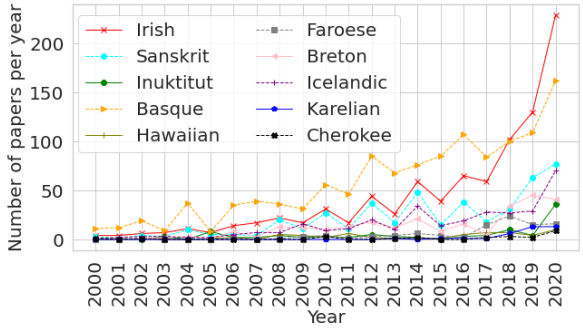

Figure 6: Distribution of papers referencing the languages with the highest papers-per-capita count.

papers for each language in Figure 6. A number of European languages with small speaker but dedicated research communities feature prominently while a number of other languages have also seen outsized research interest. The uptick in papers for Basque and Inuktitut in 2020 can be partially attributed to their inclusion in WMT 2020 News and Biomedial Translation tasks respectively.

To provide a more detailed overview of the languages with outsized research contributions relative to their speaker populations, we show the number of papers per million speakers for the 20 languages with the highest paper-per-capita count as well as some high-resource languages for comparison in Table 4.

Looking at the most under-researched languages relative to their speaker populations, we identify several varieties of Chinese including Jinyu, Min Nan, and Xiang Chinese; and several Indo-Aryan languages such as Rangpuri, Saraiki, and Chittagonian spoken in India or Bangladesh. We highlight a number of other extremely under-researched languages in Table 4, including two major languages of Indonesia and two African languages.

Overall, our analyses show that there are many languages with large speaker populations that are under-represented in current NLP research and systems, and consequently there is much headroom for developing language technology for these languages.

Pinyin (a language in Cameroon as well as the Chinese romanization system), Bench (an Ethiopian language), Ottawa (an indigenous Canadian language), Male (an Ethiopian language), Maria (an Indian language), etc.

Language	# of papers per million speakers	# of speakers (in millions)
Irish	5235	0.2
Sanskrit	3872	0.1
Inuktitut	2735	< 0.1
Basque	2430	0.5
Hawaiian	2068	< 0.1
Faroese	1515	< 0.1
Breton	1335	0.2
Icelandic	1063	0.3
Karelian	1000	< 0.1
Cherokee	1000	< 0.1
North Saami	960	< 0.1
Scottish Gaelic	771	< 0.1
Choctaw	673	< 0.1
Estonian	664	1
Plains Cree	647	< 0.1
Tuvalu	646	< 0.1
Romani	616	4
Corsican	600	< 0.1
Navajo	523	0.2
Czech	441	10
German	179	83
English	63	550
Arabic	29	180
Spanish	16	490
Chinese	11	1,000
Hausa	1.5	70
Sundanese	1.0	39
Bhojpuri	0.8	51
Sindhi	0.8	68
Javanese	0.7	85
Nigerian Pidgin	0.4	30

Table 4: Number of papers per million speakers for different languages.

6. Conclusion

We have created the largest publicly-available, machine-readable resource with writing system and speaker information for the world's languages to date. We have described how we developed this resource and have provided an analysis of the writing systems and speaker information it covers. We have also shown some examples of the kinds of language-technology analysis that our data enables. We hope the release of this data will facilitate and enable new research directions in under-represented languages.

7. Acknowledgements

We would like to thank Ankur Bapna, Sandy Ritchie, Vera Axelrod, and Françoise Beaufays for their helpful suggestions and their support. We thank Fajri Koto for help with the analysis of NLP papers. We would also like to thank all the linguists—including our linguist team at Google but even more importantly the thousands of linguists elsewhere—whose work to document, analyze, and catalog the world's languages has enabled us to produce this dataset.

8. Bibliographical References

Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., Mayhew, S., Azime, I. A., Muhammad, S. H., Emezue, C. C., Nakatumba-Nabende, J., Ogayo, P., Anuoluwapo, A., Gitau, C., Mbaye, D., Alabi, J., Yimam, S. M., Gwadabe, T. R., Ezeani, I., Niyongabo, R. A., Mukiibi, J., Otiende, V., Orife, I., David, D., Ngom, S., Adewumi, T., Rayson, P., Adeyemi, M., Muriuki, G., Anebi, E., Chukwuneke, C., Odu, N., Wairagala, E. P., Oyerinde, S., Siro, C., Bateesa, T. S., Oloyede, T., Wambui, Y., Akinode, V., Nabagereka, D., Katusiime, M., Awokoya, A., MBOUP, M., Gebreyohannes, D., Tilaye, H., Nwaike, K., Wolde, D., Faye, A., Sibanda, B., Ahia, O., Dossou, B. F. P., Ogueji, K., DIOP, T. I., Diallo, A., Akinfaderin, A., Marengereke, T., and Osei, S. (2021). MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., and Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May. European Language Resources Association.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., et al. (2021). Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Bender, E. M. (2011). On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology*, 6(3):1–26.

Bhat, I. A., Mujadia, V., Tammewar, A., Bhat, R. A., and Shrivastava, M. (2014). Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, pages 48–53.

Bhattacharjee, A., Hasan, T., Samin, K., Islam, M. S., Rahman, M. S., Iqbal, A., and Shahriyar, R. (2021). Banglabert: Combating embedding barrier in multilingual models for low-resource language understanding. *arXiv preprint arXiv:2101.00204*.

Blasi, D., Anastasopoulos, A., and Neubig, G. (2021). Systematic Inequalities in Language Technology Performance across the World's Languages. *arXiv preprint arXiv:2110.06733*.

Breiner, T., Nguyen, C., van Esch, D., and O'Brien, J. (2019). Automatic keyboard layout design for low-resource latin-script languages.

Brown, R. (2014). Non-linear mapping for improved identification of 1300+ languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 627–632, Doha, Qatar, October. Association for Computational Linguistics.

Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., Ruder, S., Lim, Z. Y., Bahar, S., Khodra, M., Purwarianti, A., and Fung, P. (2021). IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Chua, M., van Esch, D., Coccaro, N., Cho, E., Bhandari, S., and Jia, L. (2018). Text normalization infrastructure that scales to hundreds of language varieties. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*.

Chung, H. W., Févry, T., Tsai, H., Johnson, M., and Ruder, S. (2021). Rethinking Embedding Coupling in Pre-trained Language Models. In *Proceedings of ICLR 2021*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of ACL 2020*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL 2019*.

Matthew S. Dryer et al., editors. (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2021). *Ethnologue: Languages of the World. Twenty-fourth edition.* Dallas, Texas: SIL International.

Foley, B., Arnold, J., Coto-Solano, R., Durantin, G., Ellison, T. M., van Esch, D., Heath, S., Kratochvíl, F., Maxwell-Smith, Z., Nash, D., Olsson, O., Richards, M., San, N., Stoakes, H., Thieberger, N., and Wiles, J. (2018). Building speech recognition systems for language documentation: The coedl endangered language pipeline and inference system.

In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2018)*.

∀, Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., Tajudeen, K., Degila, K., Ogueji, K., Siminyu, K., Kreutzer, J., Webster, J., Ali, J. T., Abbott, J., Orife, I., Ezeani, I., Dangana, I. A., Kamper, H., Elsahar, H., Duru, G., Kioko, G., Espoir, M., van Biljon, E., Whitenack, D., Onyefuluchi, C., Emezue, C. C., Dossou, B. F. P., Sibanda, B., Bassey, B., Olabiyi, A., Ramkilowan, A., Öktem, A., Akinfaderin, A., and Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, November. Association for Computational Linguistics.

Good, J. and Hendryx-Parker, C. (2006). Modeling contested categorization in linguistic databases. In *Proceedings of the EMELD 2006 Workshop on Digital Language Documentation: Tools and standards: The state of the art*, pages 20–22.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzman, F., and Fan, A. (2021). The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation.

Gref, E. K. (2016). *Publishing in North American Indigenous Languages*. Ph.D. thesis, University of London.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2021). Glottolog 4.5.

Howard, J. and Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. In *Proceedings of ACL 2018*.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. In *Proceedings of ICML 2020*.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of ACL 2020*.

Kreutzer, J., Bastings, J., and Riezler, S. (2019). Joey NMT: A minimalist NMT toolkit for novices. 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): System Demonstrations*, pages 109–114, Hong Kong, China, November. Association for Computational Linguistics.

Kuhn, R., Davis, F., Désilets, A., Joanis, E., Kazantseva, A., Knowles, R., Littell, P., Lothian, D., Pine, A., Wolf, C. R., et al. (2020). The indigenous languages technology project at nrc canada: An empowerment-oriented approach to developing language software. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5866–5878.

Mager, M., Gutierrez-Vasques, X., Sierra, G., and Meza-Ruiz, I. (2018). Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Giménez-Lugo, G., Ramos, R., Meza Ruiz, I. V., Coto-Solano, R., Palmer, A., Mager-Hois, E., Chaudhary, V., Neubig, G., Vu, N. T., and Kann, K. (2021). Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online, June. Association for Computational Linguistics.

Mirzakhalov, J. (2021). *Turkic Interlingua: A Case Study of Machine Translation in Low-resource Languages*. Ph.D. thesis, University of South Florida.

Muller, B., Anastasopoulos, A., Sagot, B., and Seddah, D. (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, June. Association for Computational Linguistics.

Pfeiffer, J., Vulić, I., Gurevych, I., and Ruder, S. (2020). MAD-X: An Adapter-based Framework for Multi-task Cross-lingual Transfer. In *Proceedings of EMNLP 2020*.

Pfeiffer, J., Vulić, I., Gurevych, I., and Ruder, S. (2021). UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Prasad, M., Breiner, T., and van Esch, D. (2018). Mining training data for language modeling across the world's languages. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2018)*.

Roark, B., Wolf-Sonkin, L., Kirov, C., Mielke, S. J., Johny, C., Demirşahin, I., and Hall, K. (2020). Processing South Asian languages written in the Latin

script: the Dakshina dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 2413–2423.

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., and Gurevych, I. (2021). How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online, August. Association for Computational Linguistics.

Scannell, K. P. (2007). The crúbadán project: Corpus building for under-resourced languages. In *3rd Web as Corpus Workshop, 2007, Louvain-la-Neuve, Belgium*.

Seifart, F., Evans, N., Hammarstrom, H., and Levinson, S. C. (2018). Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

van Esch, D., Sarbar, E., Lucassen, T., O'Brien, J., Breiner, T., Prasad, M., Crew, E., Nguyen, C., and Beaufays, F. (2019). Writing Across the World's Languages: Deep Internationalization for Gboard, the Google Keyboard. *arXiv preprint arXiv:1912.01218*.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June. Association for Computational Linguistics.

Zupon, A., Crew, E., and Ritchie, S. (2021). Text normalization for low-resource languages of africa.

Ács, J. (2019). Exploring BERT's Vocabulary. `http://juditacs.github.io/2019/02/19/bert-tokenization-stats.html`.

A. Appendix

A.1. Common Writing Systems

We show the ISO 15924 code, names, and number of languages for writing systems covering more than one language in Table 5.

A.2. Writing Systems in Vocabularies of Pre-trained Language Models

We show the detailed breakdown of writing systems in the vocabularies of mBERT, XLM-R, and mT5 in Tables 6, 7, and 8 respectively.[27]

[27] Miscellaneous indicates punctuation, mathematical symbols, and various other kinds of symbols.

ISO 15924	Writing system	# of languages
latn	Latin	2554
deva	Devanagari	121
arab	Arabic	110
cyrl	Cyrillic	109
ethi	Ethiopic	29
beng	Bengali	21
mymr	Myanmar	20
cans	Canadian syllabics	19
thai	Thai	18
gujr	Gujarati	13
orya	Oriya	10
hans	Han (Simplified)	9
telu	Telugu	9
tibt	Tibetan	7
knda	Kannada	7
yiii	Yi	6
batk	Batak	6
mong	Mongolian	5
taml	Tamil	5
plrd	Miao (Pollard)	5
grek	Greek	5
takr	Takri	5
syrc	Syriac	4
geor	Georgian	4
tavt	Tai Viet	4
hebr	Hebrew	4
java	Javanese	4
guru	Gurmukhi	4
laoo	Lao	4
nkoo	N'Ko	4
tagb	Tagbanwa	3
hant	Han (Traditional)	3
kthi	Kaithi	3
lana	Tai Tham (Lanna)	3
bugi	Buginese	3
cham	Cham	2
khmr	Khmer	2
kali	Kayah Li	2
mlym	Malayalam	2
kana	Katakana	2
dupl	Duployan shorthand	2
glag	Glagolitic	2
kore	Korean	2
gran	Grantha	2
khoj	Khojki	2
sind	Sindhi	2
newa	Newa	2
shrd	Sharada	2
bali	Balinese	2
gonm	Masaram Gondi	2
tfng	Tifinagh (Berber)	2
lisu	Lisu (Fraser)	2
hmng	Pahawh Hmong	2
rjng	Rejang	2

Table 5: Common writing systems.

Writing system	# subwords	% subwords
Latin	68,725	57.49
CJK	14,934	12.49
Cyrillic	13,727	11.48
Arabic	4,874	4.08
Korean	3,275	2.74
Hebrew	2,482	2.08
Devanagari	1,852	1.55
Greek	1,567	1.31
Armenian	1,235	1.03
Bengali	946	0.79
Telugu	887	0.74
Tamil	832	0.70
Georgian	704	0.59
Kannada	653	0.55
Malayalam	565	0.47
Miscellaneous	563	0.47
Gurmukhi	406	0.34
Gujarati	404	0.34
Thai	370	0.31
Myanmar	271	0.23
Tibetan	40	0.03
Sinhala	12	0.01
Mongolian	4	0.00
Khmer	2	0.00

Table 6: Representation of writing systems in mBERT's vocabulary, in terms of the number and percentage of covered subwords.

Writing system	# subwords	% subwords
Latin	111,719	44.69
Cyrillic	31,672	12.67
CJK	20,627	8.25
Arabic	14,638	5.86
Devanagari	7,722	3.09
Malayalam	7,242	2.90
Korean	5,414	2.17
Hebrew	5,185	2.07
Greek	5,175	2.07
Thai	4,336	1.73
Georgian	3,770	1.51
Armenian	3,527	1.41
Sinhala	3,323	1.33
Telugu	3,236	1.29
Ethiopic	2,986	1.19
Kannada	2,773	1.11
Tamil	2,628	1.05
Bengali	2,499	1.00
Myanmar	2,455	0.98
Gujarati	2,265	0.91
Khmer	1,966	0.79
Oriya	1,844	0.74
Gurmukhi	1,675	0.67
Lao	1,615	0.65
Miscellaneous	1,608	0.64
Syriac	13	0.01
Canadian syllabics	5	0.00
Mongolian	5	0.00
Tibetan	3	0.00
Limbu	1	0.00

Table 7: Representation of writing systems in XLM-R's vocabulary, in terms of the number and percentage of covered subwords.

Writing system	# subwords	% subwords
Latin	133,651	53.44
CJK	27,189	10.87
Cyrillic	26,699	10.68
Arabic	7,422	2.97
Greek	5,252	2.10
Malayalam	4,722	1.89
Thai	4,490	1.80
Korean	4,131	1.65
Hebrew	4,123	1.65
Tamil	3,329	1.33
Devanagari	3,185	1.27
Myanmar	3,038	1.21
Georgian	2,581	1.03
Miscellaneous	2,527	1.01
Telugu	2,446	0.98
Armenian	2,272	0.91
Kannada	2,217	0.89
Khmer	2,100	0.84
Bengali	1,885	0.75
Sinhala	1,706	0.68
Lao	1,472	0.59
Gujarati	1,166	0.47
Ethiopic	1,030	0.41
Gurmukhi	631	0.25
Oriya	116	0.05
Thaana	108	0.04
Tibetan	99	0.04
Canadian syllabics	86	0.03
Mongolian	57	0.02
Syriac	52	0.02
Runic	28	0.01
Cherokee	25	0.01
Limbu	11	0.00
Tai Le	4	0.00
Buhid	2	0.00
Tagalog	2	0.00
Ogham	1	0.00

Table 8: Representation of writing systems in XLM-R's vocabulary, in terms of the number and percentage of covered subwords.