Validating and Exploring Large Geographic Corpora

Jonathan Dunn
University of Illinois Urbana-Champaign
Department of Linguistics
jedunn@illinois.edu

Abstract

This paper investigates the impact of corpus creation decisions on large multi-lingual geographic web corpora. Beginning with a 427 billion word corpus derived from the Common Crawl, three methods are used to improve the quality of sub-corpora representing specific language-country pairs like New Zealand English: (i) the agreement of independent language identification systems, (ii) hash-based deduplication, and (iii) location-specific outlier detection. The impact of each of these steps is then evaluated at the language level and the country level by using corpus similarity measures to compare each resulting corpus with baseline data sets. The goal is to understand the impact of upstream data cleaning decisions on downstream corpora with a specific focus on under-represented languages and populations. The evaluation shows that the validity of sub-corpora is improved with each stage of cleaning but that this improvement is unevenly distributed across languages and populations. This result shows how standard corpus creation techniques can accidentally exclude under-represented populations.

Keywords: geographic corpora, web corpora, outlier detection, deduplication, language identification

1. Multi-Lingual Geographic Corpora

This paper investigates the impact of corpus creation methods on large multi-lingual geographic corpora. A *geographic* corpus here is a collection of written language which represents the production of a particular place, defined at the country level. Thus, these geographic corpora might distinguish between Australian English and Indian English using a combination of language identification (to determine that each sample represents English) and geo-referencing (to determine that a sample represents Australia or India). The *place* which a corpus represents is really a proxy for the *population* which that corpus represents. In other words, a corpus from Australia would capture the production of people currently in Australia, whether long-term locals or recent immigrants. Such multi-lingual geographic corpora are important for NLP if the field endeavours to equally represent all languages and populations. A list of the largest non-English language-country sub-corpora in the final data set is shown in Table 4 at the end of this paper.[1]

This paper starts with an existing web corpus derived from the Common Crawl, the 427 billion word *Corpus of Global Language Use* (CGLU: Dunn 2020). The paper then implements a series of three improved data cleaning methods: *First*, the use of multiple language identification models to triangulate the actual language of ambiguous texts; *Second*, the use of hash-based deduplication to further remove samples dominated by non-authentic data (i.e., boilerplate texts which do not represent the linguistic production of individuals); *Third*, perplexity-based outlier detection which is specific to a text's place of origin (i.e., so that Swiss German is not seen as an outlier of other dialects of German).

Each of these improved cleaning methods is then evaluated across languages and across countries in order to understand both their wider impact on the corpus as a whole but also the distribution of their impact across different populations. In other words, a cleaning method which improves a corpus of Swiss German might at the same time degrade a corpus of Hindi from Fiji. The goal is to localize the impact of each of these methods. This paper is the first systematic investigation of how computational corpus creation methods influence the amount of linguistic diversity contained in large corpora.[2]

The basic idea behind the evaluation method is to compare the resulting geographic corpora (i.e., Swiss German on the web) to two types of baseline corpora: First, ground-truth corpora used to train language identification models; Second, geo-located tweets representing the same local populations (such as tweets in German from Zurich). The first ground-truth corpus evaluates the *language* the corpus represents and the second evaluates the *population* the corpus represents. Our hypothesis would be that each additional cleaning method should increase the similarity between the web corpus and these baseline corpora. One contribution of this work is the systematic use of baseline corpora of known quality to validate larger geographic corpora while not discarding samples from under-represented populations. The challenge is to maintain linguistic diversity while still removing noise.
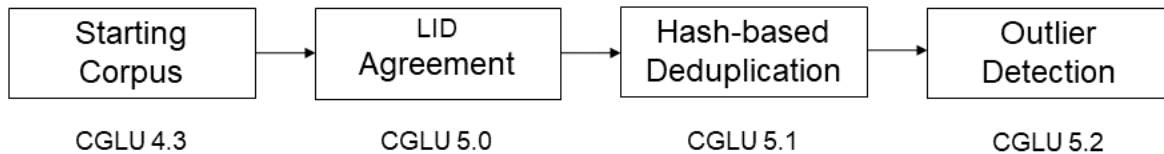
Figure 1: Sequence of cleaning methods from CGLU 4.3 to CGLU 5.2

Region	CGLU 4.3	CGLU 5.0	CGLU 5.1	CGLU 5.2	TWGLU
Africa, North	1,245 million	1,064 million	666 million	646 million	466 million
Africa, Southern	42 million	38 million	16 million	15 million	382 million
Africa, Sub	6,23 million	5,660 million	4,267 million	4,241 million	1,349 million
America, Brazil	2,265 million	1,955 million	751 million	718 million	492 million
America, Central	8,972 million	6,111 million	3,339 million	3,259 million	2,311 million
America, North	51,921 million	36,035 million	26,964 million	26,391 million	1,155 million
America, South	22,441 million	17,678 million	11,720 million	11,459 million	2,791 million
Asia, Central	7,185 million	5,960 million	4,155 million	4,053 million	473 million
Asia, East	49,521 million	37,695 million	7,901 million	7,389 million	924 million
Asia, South	15,147 million	10,701 million	7,515 million	7,341 million	2,738 million
Asia, Southeast	23,189 million	20,225 million	15,118 million	14,563 million	1,218 million
Europe, East	65,413 million	54,544 million	38,455 million	37,261 million	1,372 million
Europe, West	146,327 million	111,617 million	80,289 million	77,825 million	5,547 million
Europe, Russia	15,363 million	12,561 million	11,455 million	11,240 million	324 million
Middle East	11,606 million	10,378 million	7,239 million	7,102 million	1,224 million
Oceania	329 million	278 million	137 million	134 million	1,223 million
TOTAL	**427 billion**	**332 billion**	**219 billion**	**213 billion**	**23 billion**

Table 1: Size of corpora in number of words by region. CGLU 4.3 is the starting point. CGLU 5.0 includes two language identification models. CGLU 5.1 includes hash-based deduplication at the sentence level. CGLUV 5.2 includes location-sensitive outlier detection. TWGLU 3.0 is the baseline for geo-referencing.

The evaluation thus considers a sequence of four large geographic corpora, as shown in Table 1. The starting point contains 427 billion words (CGLU 4.3). After multiple language identification models are applied this is reduced to 332 billion words (CGLU 5.0). After hash-based deduplication at the sentence level this is reduced to 219 billion words (CGLU 5.1). And after location-specific outlier detection this is reduced to 213 billion words (CGLU 5.2). The size of the twitter baseline is 23 billion words (TWGLU). While this shows the overall magnitude of each additional cleaning step, the analysis in the paper is focused on two questions: First, what is the distribution of the impact of each step across languages and across countries? Second, do these cleaning steps increase the validity of the resulting geographic corpora?

This paper makes two primary contributions: In the first case, the analysis here provides a systematic evaluation of corpus creation methods across many languages and many populations. This is important because the uneven impact of methods shown in this paper implies that existing approaches to removing noise from large corpora also systematically remove sources of linguistic diversity. In the second case, this paper describes a new publicly-available multi-lingual geographic corpus which is cleaner and more usable than the original data from the Common Crawl. The general workflow for the successive versions of the corpus evaluated here is shown in Figure 1. This figure connects the specific cleaning steps with the relevant version of CGLU; these versions will be used throughout the paper as a shorthand. The three main steps are (i) enforcing language agreement (CGLU 5.0), (ii) improved deduplication (CGLU 5.1), and (iii) improved outlier detection (CGLU 5.2).

The paper is organized as follows: in Section 2 we consider related work on geographic corpora in particular and corpus validation more broadly. In Section 3 we consider the sources of data for the geographic web corpus as well as for the benchmark corpora used for purposes of comparison. We then describe and evaluate the impact of using independent language identification models to validate language labels (Section 4), of using hash-based deduplication (Section 5), and of using location-specific outlier detection (Section 6). Finally, in Section 7 we use the benchmark corpora to measure improvements to the corpus as a whole and in Section 8 we consider the wider implications for both multi-lingual and geographic corpora.

2. Related Work

There is a long history of using geographic corpora to represent language use by different populations, including country-specific corpora like the *British National Corpus* (BNC Consortium, 2007) and the *Corpus of Contemporary American English* (Davies, 2008). While these focus on individual countries, the *International Corpus of English* focuses instead of comparing a single language across multiple populations (Greenbaum, 1996). While these early corpora were somewhat limited in size, more recent corpora like the *Corpus of Global Web-based English* (Davies, 2013) include web-crawled data to create a total corpus of 1.9 billion words – somewhat small for NLP but significantly larger than previous geographic corpora. Multi-lingual geographic corpora soon followed, including the 427 billion word *Corpus of Global Language Use* (Dunn, 2020) and *GeoWAC* (Dunn and Adams, 2020), a collection of gigaword corpora for 50 languages, each balanced geographically to represent the real-world distribution of speakers for that language.

Geographic corpora explicitly represent different populations of speakers and thus have often been used to study linguistic variation. Such studies also provide a method of validating the corpora themselves. For instance, previous work has shown that both geographic web corpora (Cook and Brinton, 2017) and geo-referenced tweets (Grieve et al., 2019) can be used to replicate traditional studies of lexical variation across dialects. At the syntactic level, another line of work has shown that such corpora can be used to model grammatical variation across dialects (Dunn, 2018, 2019b,a, 2023a). Such work provides a validation of geographic corpora, at least in major languages like English and French, in the sense that they contain the linguistic variants expected from dialect surveys.

Another approach to validating geographic corpora relies on corpus similarity methods (Kilgarriff, 2001; Fothergill et al., 2016) to compare geo-referenced corpora to known ground-truth data from alternate sources. Recent work has shown that such measures can be updated for highly multilingual corpora (Li and Dunn, 2022) and also used to differentiate corpora from different sources or registers (Li et al., 2023). Recent work (Dunn, 2021) has applied these methods to geographic corpora by using geo-referenced tweets (with quite precise geographic meta-data) to validate geo-referenced web pages (which rely on less reliable top-level domain information). The question, again, is whether a geographic corpus actually represents the underlying population it is meant to represent. Another recent approach to exploring geo-referenced corpora relies on comparing named entities extracted from the text with the expected location of such entities (Faisal et al., 2022). While promising, such an approach relies on both the extraction of named entities and a reliable gazetteer of their location, both of which are problematic for low-resource languages and under-represented locations.

Recent audits of multi-lingual corpora have shown that many automated efforts create problematic corpora (Kreutzer et al., 2022). In part this is because of upstream issues in the data used for language identification models; given the enormity of digital sources like the web, language identification models with a relatively high precision and recall on formal texts can produce corpora which do not actually represent that language when applied to informal texts. This is one motivation for applying multiple language identification models to provide increased certainty about language labels.

This current work can also be situated relative to other corpus creation pipelines that start with the Common Crawl, such as CCNet (Wenzek et al., 2020). An important rationale is to measure the impact of each design decision, not only in the aggregate but also across languages and across populations. For instance, the outlier detection employed in CCNet relies on language models trained on Wikipedia, even though we know that there is a strong geographic skew in the contributors to Wikipedia (Graham et al., 2015) and even though there is a significant linguistic difference between Wikipedia's register and others (Li et al., 2023).

There is a trade-off here between (i) producing clean and usable corpora which contain utterances that are meaningful to human speakers while (ii) capturing linguistic variation and allowing for non-standard forms from under-represented populations. For instance, it has been shown that a large English corpus from the Common Crawl excludes certain populations as a result of cleaning decisions (Dodge et al., 2021). The goal of this paper, then, is to measure the impact of each stage of processing across languages and populations.

3. Data and Methods

The main data which serves as the starting point for this analysis is the full *Corpus of Global Language Use* (CGLU; Dunn 2020). This is a 427 billion word web corpus derived from the Common Crawl and geo-referenced mainly using information from top-level domains. The corpus is organized into 16 larger regions as shown in Table 1; the same geographic structure is maintained in this paper in order to enable the analysis of the impact of different cleaning steps.

Two sets of comparison corpora are used for validation. First, ground-truth language identification corpora are taken from seven independent sources: Bible translations (Brown, 2014), Global

Language	CGLU 4.3	CGLU 5.0	CGLU 5.1	CGLU 5.2	Change
Russian (rus)	5.96%	6.57%	8.73%	8.80%	+2.83%
Vietnamese (vie)	3.76%	4.60%	5.67%	5.59%	+1.82%
German (deu)	4.83%	5.28%	6.53%	6.53%	+1.69%
French (fra)	6.15%	6.38%	7.20%	7.17%	+1.02%
Dutch (nld)	2.23%	2.42%	3.11%	3.14%	+0.90%
Polish (pol)	1.51%	1.71%	2.26%	2.25%	+0.73%
Farsi (fas)	2.30%	2.78%	3.00%	3.03%	+0.73%
Finnish (fin)	1.02%	0.75%	0.80%	0.79%	-0.22%
Spanish (spa)	9.06%	8.23%	8.61%	8.55%	-0.51%
Portuguese (por)	1.45%	1.16%	0.95%	0.94%	-0.51%
Serbo-Croatian (hbs)	1.81%	1.72%	1.18%	1.19%	-0.62%
Japanese (jpn)	3.62%	3.88%	1.53%	1.46%	-2.16%
English (eng)	30.21%	26.62%	26.46%	26.61%	-3.59%
Chinese (zho)	5.74%	6.68%	1.16%	1.14%	-4.59%

Table 2: Size of corpora by language, given in percent of corpus in number of words after character segmentation. CGLUv4.3 is the starting point. CGLU5.0 includes two language identification models. CGLU5.1 includes hash-based deduplication at the sentence level. CGLUv5.2 includes location-specific outlier detection. CHANGE is the difference between the original corpus and the final corpus.

Voices News (Tiedemann, 2012), the JW 300 data set (Agić and Vulić, 2019), Open Subtitles (Lison and Tiedemann, 2016), QCRI Educational Domain (Tiedemann, 2012), Tatoeba Sentences (Tiedemann, 2012), and Wikipedia articles. Samples are drawn equally from these sources to avoid over-representing a specific register. Known inconsistencies for several languages in the JW 300 data are corrected (Kreutzer et al., 2022). These samples represent corpora from a known language. The second source of comparison corpora is tweets collected from the same countries as the original CGLU and divided into regions in the same manner (Dunn, 2019a; Dunn and Wong, 2022; Dunn, 2023b). The distribution of this corpora, totalling 23 billion words, is shown in Table 1. Because language identification is an essential component of preparing these types of corpora, the same models are used as for the agreement stage in Section 4. These samples represent corpora from a known location.

To compare each stage of cleaning with these benchmark corpora we use existing corpus similarity measures (Li et al., 2023; Li and Dunn, 2022). These measures rely on the frequency of character n-grams to differentiate between more or less similar corpora. Previous work has evaluated these in a cross-register and multi-lingual setting; here we rely directly on the models provided from this previous work.[3] These comparisons are limited to the 74 languages covered by the existing package.

4. Validating Language Labels

Classifying documents by language is an essential part of creating a web corpus. And yet the problem

of language identification itself remains a challenge for less-common languages (Brown, 2014), for non-standard varieties of languages (Jurgens et al., 2017), and for less-common or informal registers (Lui and Baldwin, 2011). Less-common languages often have lower precision or recall (Jauhiainen et al., 2022) and this can have downstream ramifications for corpus quality (Kreutzer et al., 2022). Our first validation step, then, is to use an independent language identification model to label the original CGLU data. The original paper used a feed-forward network for classification (Dunn, 2020). We employ a more recent fastText-based model (Dunn and Nijhof, 2022). Importantly, the second model represents all 464 languages that are represented by the first model.

Given two language labels for each document (from the original model and the more recent model), cases of agreement provide greater confidence that the language has been correctly identified. Noisy samples, especially, will be miscategorized by one or both models but are less likely to be miscategorized into the same label. The distribution of agreement rates across countries is shown in Figure 2, with darker red indicating high agreement and lighter red low agreement. We see from the map that the impact of this language validation step is not distributed equally: North America and Brazil, for instance, have much lower agreement rates than countries in Central and South America.

An alternate view is provided in Table 2 which shows the percentage of the web corpus which is composed of specific languages through each step. The difference between CGLU 4.3 and CGLU 5.0, then, reflects the constraint that all documents whose language labels disagree are removed. Some languages, like Russian and Viet-

---
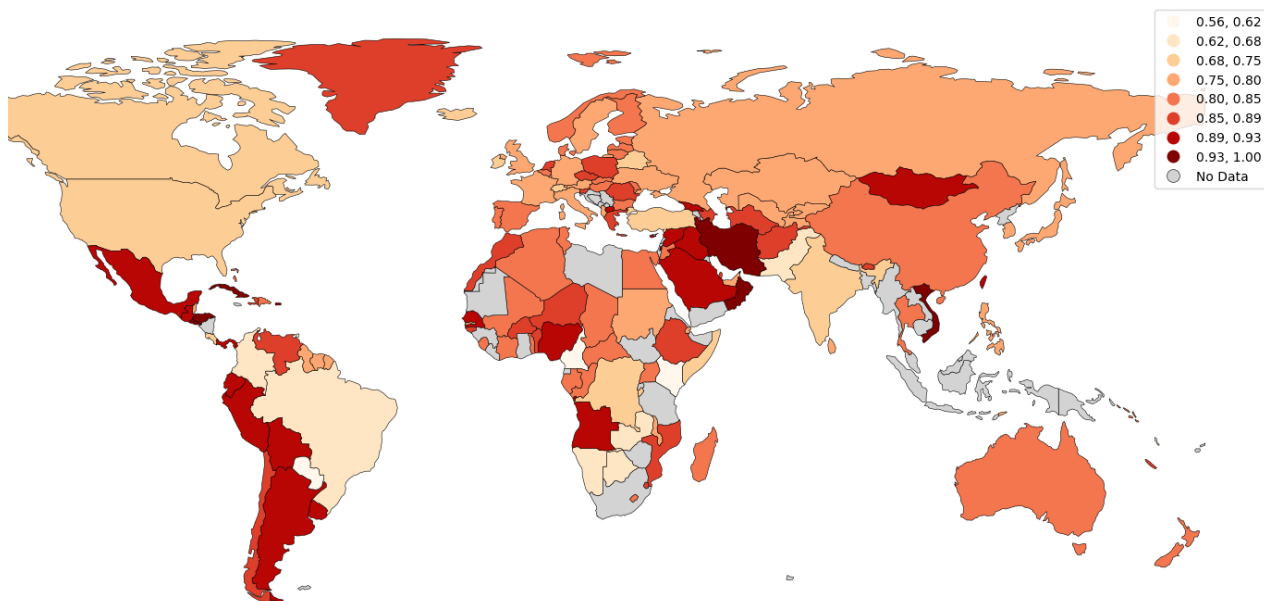[3]https://github.com/jonathandunn/corpus_similarity

Figure 2: Map showing agreement between language identification models by country. A value of 0.80 means that 80% of samples receive the same language label from each model.

namese, increase in their relative share of the data. Others, however, decrease: especially very common languages like English and Spanish.

A decrease in majority languages or in Western countries is acceptable insofar as these languages and populations are already over-represented in the corpus. As shown in Table 1, North America and Western Europe remain the single largest sources in the final corpus at 26.3 billion and 77.8 billion words, respectively. Enforcing agreement in language labels across models removes 95 billion words, or 22.2% of the original data.

5. Hash-based Deduplication

The second stage of cleaning investigated here is the removal of similar samples. The original CGLU corpus used a method based on exact matches at the level of paragraph tags, a conservative approach which avoids removing non-duplicates but has the potential to leave a large number of highly similar documents. Here we implement a more recent hash-based strategy (Wenzek et al., 2020), applied so that both line breaks and paragraph tags can separate samples. This approach first hashes each sample using the SHA-1 hashing function and then completely removes any samples which collide within that hashing space. As shown in Table 1, deduplication removes 113 billion words, or 34% of the data after language validation.

This hash-based depulication step has the biggest overall impact on the distribution of the corpus across languages and countries. We can measure this by viewing the corpus as a collec-

tion of sub-corpora representing specific language-country pairs (like Australian English). The Pearson correlation between the size of these sub-corpora after validating the language labels is 0.99 and highly significant. The correlation after deduplication is 0.93, still highly significant. And the correlation after outlier detection (see below) is again 0.99 and highly significant. Thus, deduplication is the stage which most alters the distribution of the corpus across languages and countries.

We can see why in reference to the relative percent of each language shown in Table 2: deduplication (the difference between 5.0 and 5.1) reduces Chinese from 6.68% to 1.16% and Japanese from 3.88% to 1.53%. English, which was disproportionately affected by the language validation step, has very little change during deduplication, going from 26.62% to 26.46%. If this were constrained to the usage of Chinese or Japanese in one or two countries, we might suspect that the data from those sources was problematic and unusually full of duplicates. The lower correlation, though, means that this pattern is spread across many individual countries. Thus, this indicates that the hash-based approach does not work equally well across languages and has the implicit side-effect of reducing some languages more than others. Future work is needed to investigate more equitable cross-linguistic deduplication methods.

6. Outlier Detection

The final cleaning method evaluated is outlier detection to remove noisy samples which are non-
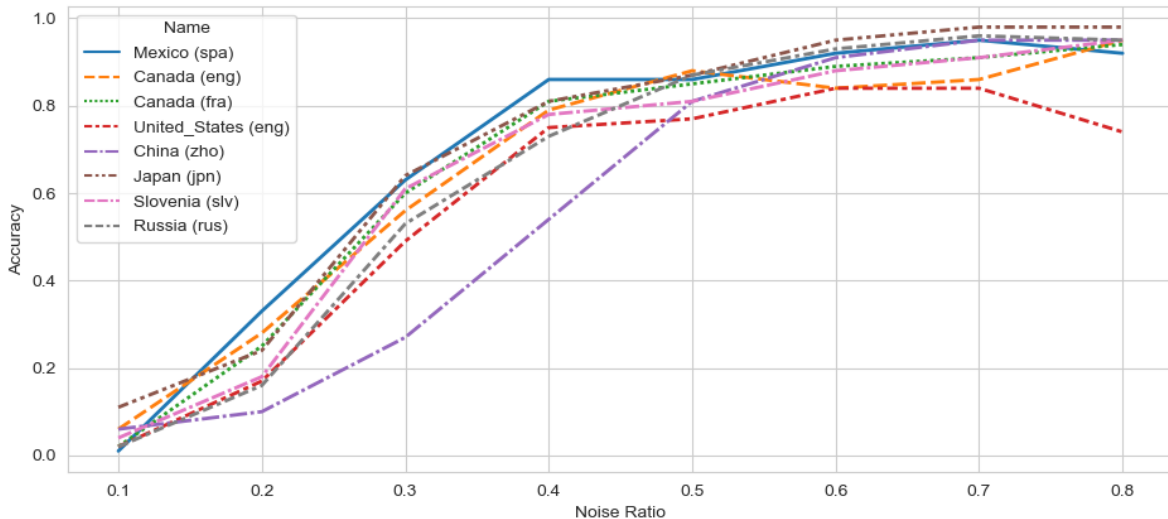
Figure 3: Accuracy of the outlier detection method for finding samples with injected noise. *Ratio* refers to the amount of noise added and *Accuracy* to the percent of such samples correctly identified.

Low	bed princess canopy new baby bed mosquito net cute baby princess canopy crib netting
Low	monster legends hack generator – monster legends hack using cheat engine monster
High	Table 10: Brand Fortified Wine market, Table 11: Brand Sparkling Wine market
High	Carnival Cruise Line Celebrity Cruises Costa Cruises Disney Cruise Cruises France

Table 3: Examples of outliers from the corpus of American English. *High* example are too predictable according to the model. *Low* examples are too unpredictable according to the model.

duplicate instances of the correct language. We again adapt and evaluate a recent approach that first trains a statistical language model for each language and then uses perplexity of documents to find outliers (Wenzek et al., 2020). The original approach trains a single model for each language using Wikipedia. Given that one goal of a geographic corpus is to include under-represented populations and low-resource varieties, using Wikipedia as the benchmark is likely to exclude many relevant documents. We therefore adapt this approach by instead training a language model for each language-country pair (like Swiss German) which contains at least 5 million words.[4]

The challenge, then, is that this creates a much larger number of models to be used for outlier detection. Instead of a statistical language model with a vocabulary from a sentence piece tokenizer, we use character-based skip-gram models with a fastText base. Specifically, we use a window size of 5 with character-ngrams of length 3 to 6, trained for two epochs on each language-country sub-corpora. This model is then used to measure the log probability of documents in the corpus given the learned embeddings, the specific measure used for outlier detection. This approach is more feasible for the large number of language-country sub-corpora that

need to be dealt with. Such an approach has previously been used for other document classification tasks (Taddy, 2015).

As in the original approach, then, each sample in the corpus is given a log probability so that the corpus itself is represented as a distribution of such values. We standardize these values using the Iglewicz and Hoaglin method of modifying the z-score to be robust to outliers shown in (1). This is comparable to the z-score but using the median rather than the mean and dividing by the median absolute deviation (MAD). Any sample with a standardized log probability which falls above 3 or below -3 is considered an outlier.[5]

$$M_i = \frac{0.6745(x_i - median(x))}{MAD} \quad (1)$$

An approach like this distinguishes between two types of outliers: those which are too unpredictable and those which are too predictable. Examples of both types are shown in Table 3. We evaluate the quality of outlier detection by artificially injecting noise into samples in the form of arbitrary words from other languages. The algorithm is them evaluated on its ability to accurately identify these noisy

---

[4]github.com/jonathandunn/common_crawl_corpus

[5]This method assumes a normal distribution of log probabilities within each corpus; figures in the supplementary material show that this assumption is met.
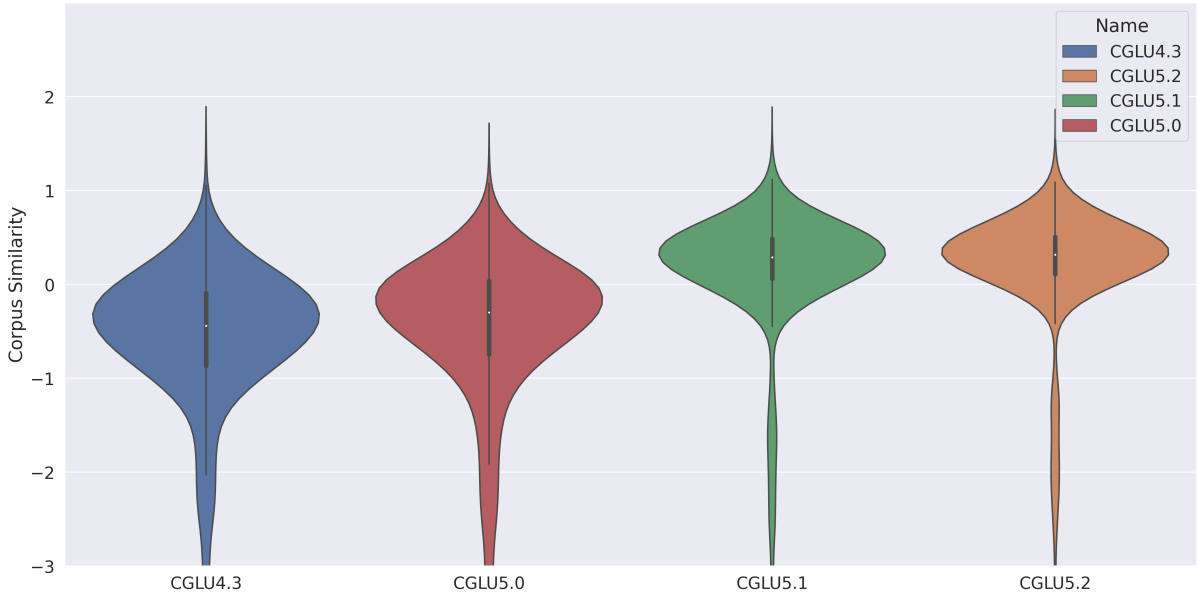
Figure 4: Similarity of the Swiss German corpus to the benchmark language identification corpus over each stage of cleaning. Higher values indicate more similar corpora. Significance of differences is tested using an ANOVA, here with a value of $p < 0.001$.

samples. Results for eight common language-country sub-corpora are shown in Figure 3. The y-axis represents prediction accuracy, with lines near the top being able to correctly identify most or all noisy samples. The x-axis represents the amount of noise inserted into the sample. At small amounts (similar to code-switching) the noisy samples are not discarded. As the ratio increases, however, almost all noisy samples are correctly identified. This evaluation shows, then, that location-specific outlier detection is able to accurately remove noisy samples from geographic corpora.

This stage of cleaning has the smallest impact on the corpus as a whole, as shown in Table 1: only 6.35 billion words are removed, or approximately 2.9% of the data. The correlation between the distribution of the corpus before and after outlier detection is 0.999 and highly significant. Thus, this cleaning method has the smallest impact overall and is the most evenly distributed across languages and countries. This is because the model used for identifying outliers is specific to each language-country sub-corpus.

7. Evaluation

We turn now to the evaluation of this series of cleaning methods, from validating the language labels to hash-based deduplication to outlier detection. The basic idea in the evaluation is to take more reliable benchmark corpora which are known to represent either (i) the language in question or (ii) the population in question. We then use corpus similarity measures (Li and Dunn, 2022) to determine if each

cleaning step makes the language-country sub-corpora more similar to these benchmarks. Being more similar to the language identification corpus would mean being cleaner and more valid samples of that language. Being more similar to the geo-referenced corpus would mean being cleaner and more valid samples of that particular location. We confine this evaluation to the 74 languages covered by these measures.

We start with the comparison with benchmark language identification data, visualizing the progression of Swiss German in Figure 4. The corpus similarity measure works by breaking the corpus into 500 equal-sized chunks (each containing 10,000 words). The violin plot is thus showing the distribution of values across many pairs of 500 chunks from the web corpus and the benchmark corpus. We use an ANOVA to test whether these are in fact distinct populations, here with a significance of $p < 0.001$. The violin plot shows two important features: first, the mean similarity of the corpus is increasing with each stage of cleaning; second, the tail of outliers decreases visually with each stage. This figure shows, then, that the corpus of Swiss German becomes more like expected German corpora at each stage of cleaning.

The next step compares the corpus instead with benchmark tweets. Here, in Figure 5 we use tweets in Spanish from Chile, evaluating whether this web corpus represents the Chilean Spanish dialect rather than other varieties of Spanish. Here we see a similar progression: the mean similarity is higher with each stage of cleaning and the presence of outliers is greatly reduced. As before, the
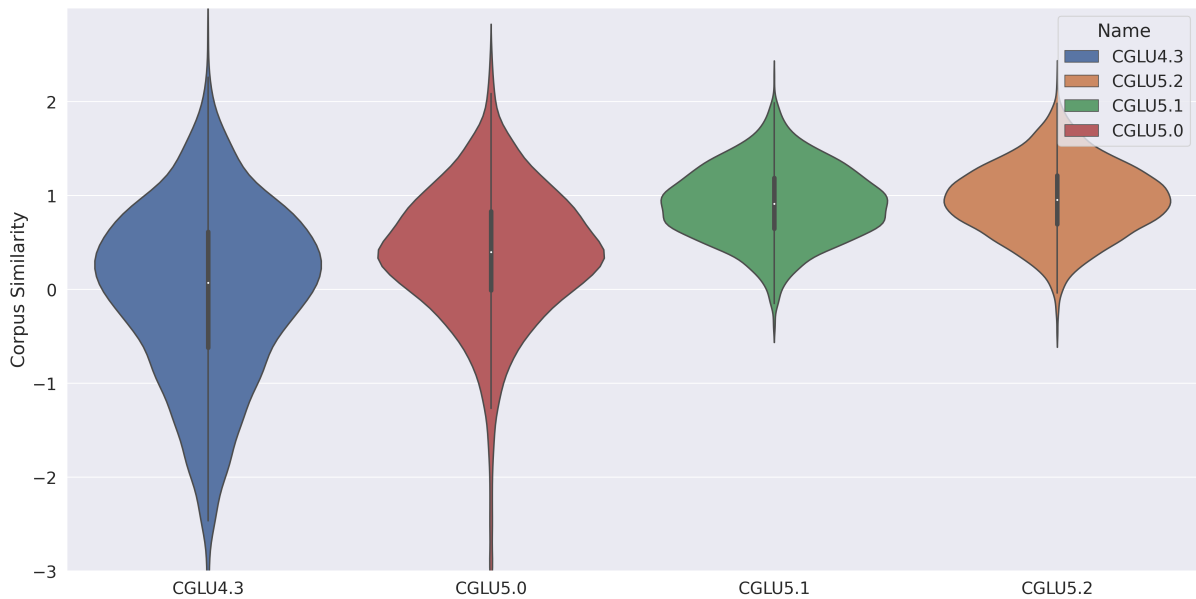
Figure 5: Similarity of the Chilean Spanish corpus to the benchmark corpus of tweets in Spanish from Chile over each stage of cleaning. Higher values indicate more similar corpora. Significance of differences is tested using an ANOVA, here with a value of $p < 0.001$.

difference is significant using an ANOVA test. The deduplication stage has the largest impact, as we would expect from the fact that this stage made the largest alteration to the corpora. The outlier detection stage, with the smallest amount of removed data, has a still visible but less impactful influence on the corpus similarity measures.

The full figures across all language-country sub-corpora are available in the supplementary material and online at the earthLings.io project. In most cases there is a similar improvement as shown in Swiss German in terms of both a higher mean and a more homogenous distribution, indicating a cleaner and more valid corpus. But this is not universally true. For example, Arabic shows an improvement in geo-referencing (against tweets) across countries in both attributes. But against the language identification benchmark Arabic does not improve in the mean value and only slightly in the reduced tail of outliers. English shows improvements across the board in Canada but somewhat less consistently in places like Mexico.

French, in contrast, shows significant improvements in geo-referencing but at the same time a slight decrease against language identification data when outside France. This implies that the language identification corpus does not adequately capture dialects of French so that accurate representations of Senegal French are actually less similar to European French. Japenese shows slight improvements in geo-referencing but no change against the language identification corpora; Chinese shows a similar pattern. Korean, in contrast, shows a significant improvement in both cases, as

do most other languages. These full figures are available in the supplementary material and are useful for exploring the full impact of these cleaning methods across different languages and populations. While almost all sub-corpora improve, the degree and character of improvement varies by language and population.

8. Discussion and Conclusions

This paper has systematically evaluated the impacts of cleaning methods on a large multi-lingual geographic web corpora, looking at the impact across both countries and languages. Three cleaning methods are evaluated in a specific order of application: (i) independent validation of language labels, (ii) hash-based deduplication, and (iii) location-specific outlier detection.

The first two cleaning methods have a very large impact on the corpus, removing 95 billion and then 113 billion words. The first stage leaves the corpus with much the same distribution across languages and countries (a Pearson correlation of 0.99), but has more impact in some areas than others. The deduplication step has the largest impact on the distribution of the corpus, with a significant reduction in East Asian languages like Chinese and Japanese. Outlier detection has the smallest impact and does not change the distribution of the corpus; yet an evaluation shows that it remains quite accurate in removing noisy samples. These findings are important because they indicate that corpus creation methods do not have a consistent impact across languages and populations.

Country	Lang.	Words (Mil)	Country	Lang.	Words (Mil)	Country	Lang.	Words (Mil)
Viet Nam	vie	11,824	Estonia	est	1,040	India	rus	121
France	fra	11,234	Serbia	hbs	1,011	Albania	sqi	117
Spain	spa	10,700	Slovenia	slv	1,001	Uzbekistan	uzb	116
Russia	rus	10,189	Peru	spa	769	Luxembourg	fra	115
Germany	deu	8,368	Ukraine	ukr	648	Honduras	spa	115
Italy	ita	7,180	Switzerland	fra	625	Colombia	rus	111
Iran	fas	6,327	Brazil	por	611	Switzerland	ita	111
Netherlands	nld	4,864	Colombia	spa	570	Palau	rus	104
Poland	pol	4,687	Azerbaijan	aze	466	Malaysia	ind	102
Romania	ron	4,295	Bosnia	hbs	394	Lithuania	rus	100
Czechia	ces	4,032	Cuba	spa	378	Ecuador	spa	97
Norway	nor	3,693	Taiwan	zho	365	USA	por	97
Sweden	swe	3,588	South Korea	kor	363	Colombia	deu	95
Slovakia	slk	3,566	Iceland	isl	286	Finland	swe	79
Denmark	dan	3,092	Moldova	ron	270	Colombia	ara	78
Austria	deu	2,971	Indonesia	ind	263	USA	spa	78
Japan	jpn	2,919	Georgia	kat	258	Belarus	bel	76
Chile	spa	2,849	Latvia	rus	258	USA	fra	74
Greece	ell	2,624	Colombia	fra	201	Morocco	ara	73
Hungary	hun	2,407	N. Macedonia	mkd	196	Spain	cat	66
Lithuania	lit	2,336	Kazakhstan	kaz	191	Liechtenstein	spa	66
Belarus	rus	2,247	Colombia	ind	166	Timor-Leste	spa	66
Kazakhstan	rus	2,144	Uzbekistan	rus	153	Gabon	fra	62
Bulgaria	bul	1,933	Azerbaijan	rus	152	Colombia	jpn	61
China	zho	1,931	Kyrgyzstan	rus	151	El Salvador	spa	59
Mexico	spa	1,844	Ecuador	rus	150	C. A. Rep.	fra	58
Switzerland	deu	1,824	Estonia	rus	150	Gabon	rus	53
Ukraine	rus	1,802	Moldova	rus	149	USA	ara	53
Belgium	nld	1,764	Slovakia	ces	143	USA	deu	53
Finland	fin	1,665	Morocco	fra	141	Dominican R.	spa	52
Latvia	lav	1,120	USA	rus	139	Slovakia	hun	52
Belgium	fra	1,095	Haiti	rus	134	Colombia	por	52
Portugal	por	1,091	Mongolia	mon	133	Vatican	ita	51
Croatia	hbs	1,080	UAE	ara	131	USA	ind	51
Canada	fra	1,062	Palestine	ara	127	Réunion	fra	51

Table 4: Size of CGLUV 5.2 for the 105 non-English language-country pairs with at least 50 million words. Each pair is a sub-corpus representing the use of a specific language in a specific country. *Size* is given in millions of words per corpus.

At the same time, it is essential to evaluate the impact of these methods because excluding documents from under-represented populations or low-resource varieties serves to further exclude these populations from the language technologies which depend on these corpora (Dodge et al., 2021). In other words, excluding a document because it belongs to a low-resource variety means that variety will remain under-represented. An important contribution of geographic corpora is that they are able to ensure equal representation across diverse populations and language varieties. For instance, Table 4 shows the 105 largest non-English language-country sub-corpora contained in the final dataset. This corpus improves our representation of linguistic diversity by including many populations around the world using many languages that are often under-represented.

The main contribution of this paper is not only a systematic evaluation of these methods but also a new version of the large multi-lingual geographic web corpus, now containing 213 billion words, which is measurably better than previous versions. As shown in Table 1, this corpus is organized by country and divided into 16 larger regions. While more wealthy populations from North America and Western Europe provide a large portion of this data, their combined contribution is below half of the final data set. This allows for significant representation of populations from Africa (nearly 5 billion words) and Asia (over 33 billion words), a large improvement on existing corpora.

9.   Ethics Statement

This paper uses written digital corpora to represent diverse populations of speakers around the world. While such geographic corpora are more representative of the world's population than non-geographic corpora, it remains the case the certain demographic segments within each geographic area are more likely to contribute to the corpus. Thus, this data should not be taken as an exhaustive representation of all sub-populations within each geographic location.

10.   Bibliographical References

Željko Agić and Ivan Vulić. 2019. JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210. Association for Computational Linguistics.

BNC Consortium. 2007. The British National Corpus, XML Edition. Oxford Text Archive.

Ralf Brown. 2014. Non-linear mapping for improved identification of 1300+ languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 627–632.

Paul Cook and Laurel Brinton. 2017. Building and Evaluating Web Corpora Representing National Varieties of English. *Language Resources and Evaluation*, 51(3):643–662.

Mark Davies. 2008. The Corpus of Contemporary American English (COCA). BYU Corpora.

Mark Davies. 2013. Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries (GloWbE). BYU Corpora.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan Dunn. 2018. Finding variants for construction-based dialectometry: A corpus-based approach to regional cxgs. *Cognitive Linguistics*, 29(2):275–311.

Jonathan Dunn. 2019a. Global Syntactic Variation in Seven Languages: Toward a Computational Dialectology. *Frontiers in Artificial Intelligence*, 2:15.

Jonathan Dunn. 2019b. Modeling Global Syntactic Variation in English Using Dialect Classification. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 42–53, Ann Arbor, Michigan. Association for Computational Linguistics.

Jonathan Dunn. 2020. Mapping languages: the Corpus of Global Language Use. *Language Resources and Evaluation*, 54:999–1018.

Jonathan Dunn. 2021. Representations of Language Varieties Are Reliable Given Corpus Similarity Measures. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 28–38. Association for Computational Linguistics.

Jonathan Dunn. 2023a. Syntactic variation across the grammar: Modelling a complex adaptive system. *Frontiers in Complex Systems*, 1.

Jonathan Dunn. 2023b. Syntactic variation across the grammar: modelling a complex adaptive system. *Frontiers in Complex Systems*, 1.

Jonathan Dunn and Benjamin Adams. 2020. Geographically-balanced Gigaword corpora for 50 language varieties. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2528–2536. European Language Resources Association.

Jonathan Dunn and Wikke Nijhof. 2022. Language identification for Austronesian languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6530–6539, Marseille, France. European Language Resources Association.

Jonathan Dunn and Sidney Wong. 2022. Stability of syntactic dialect classification over space and time. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 26–36, Gyeongju, Republic of Korea. International Committee on 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.

Richard Fothergill, Paul Cook, and Timothy Baldwin. 2016. Evaluating a topic modelling approach

to measuring corpus similarity. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation)*, pages 273–279, Portorož, Slovenia. European Language Resources Association.

Mark Graham, Ralph Straumann, and Bernie Hogan. 2015. Digital divisions of labor and informational magnetism: Mapping participation in wikipedia. *Annals of the Association of American Geographers*, 105(6):1158–1178.

Sidney Greenbaum. 1996. *Comparing English Worldwide: The International Corpus of English*. Clarendon Press, Oxford.

Jack Grieve, Chris Montgomery, Andrea Nini, Akira Murakami, and Diansheng Guo. 2019. Mapping Lexical Dialect Variation in British English Using Twitter. *Frontiers in Artificial Intelligence*, 2:11.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022. HeLI-OTS, off-the-shelf language identifier for text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3912–3922, Marseille, France. European Language Resources Association.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada. Association for Computational Linguistics.

Adam Kilgarriff. 2001. Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi.

2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Haipeng Li and Jonathan Dunn. 2022. Corpus similarity measures remain robust across diverse languages. *Lingua*, 275(103377).

Haipeng Li, Jonathan Dunn, and Andrea Nini. 2023. Register variation remains stable across 60 languages. *Corpus Linguistics and Linguistic Theory*, 19(3):397–426.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 923–929. European Language Resources Association.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Matt Taddy. 2015. Document classification by inversion of distributed language representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 45–49, Beijing, China. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the International Conference on Language Resources and Evaluation*, page 2214–2218. European Language Resources Association.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.