Measuring Linguistic Diversity During COVID-19

Jonathan Dunn
Department of Linguistics
University of Canterbury
Christchurch, New Zealand
jonathan.dunn@canterbury.ac.nz

Tom Coupe
Department of Economics
University of Canterbury
Christchurch, New Zealand
tom.coupe@canterbury.ac.nz

Benjamin Adams
Department of Computer Science and Software Engineering
University of Canterbury
Christchurch, New Zealand
benjamin.adams@canterbury.ac.nz

Abstract

Computational measures of linguistic diversity help us understand the linguistic landscape using digital language data. The contribution of this paper is to calibrate measures of linguistic diversity using restrictions on international travel resulting from the COVID-19 pandemic. Previous work has mapped the distribution of languages using geo-referenced social media and web data. The goal, however, has been to describe these corpora themselves rather than to make inferences about underlying populations. This paper shows that a difference-in-differences method based on the Herfindahl-Hirschman Index can identify the bias in digital corpora that is introduced by non-local populations. These methods tell us *where* significant changes have taken place and whether this leads to increased or decreased diversity. This is an important step in aligning digital corpora like social media with the real-world populations that have produced them.

1 Biases in digital language data

Data from social media and web-crawled sources has been used to map the distribution of both languages (Mocanu et al., 2013; Gonçalves and Sánchez, 2014; Lamanna et al., 2018; Dunn, 2020) and dialects (Eisenstein et al., 2014; Cook and Brinton, 2017; Dunn, 2019b,a; Grieve et al., 2019). This line of research is important because traditional methods have relied on census data and missionary reports (Eberhard et al., 2020; IMB, 2020), both of which are often out-of-date and can be inconsistent across countries. At the same time, we know that digital data sets do not necessarily reflect the underlying linguistic diversity in a country: the actual population of South Africa, for example, is not accurately represented by tweets from South Africa (Dunn and Adams, 2019).

This becomes an important problem as soon as we try to use computational linguistics to tell us about *people* or *language*. For example, if an application is using Twitter to track sentiment about COVID-19, that tracking is meaningless without good information about how well it represents the population. Or, if an application is using Twitter to study lexical choices, that study depends on a relationship between lexical choices on Twitter and lexical choices more generally. In other words, the more we use digital corpora for scientific purposes, the more we need to control for *bias* in that data. There are four sources of diversity-related bias that we need to take into account.

First, *production bias* occurs when one location (like the US) produces so much digital data that most corpora over-represent that location (Jurgens et al., 2017). For example, by default a corpus of English from the web or Twitter will mostly represent the US and the UK (Kulshrestha et al., 2012). It has been shown that this type of bias can be corrected using population-based sampling (Dunn and Adams, 2020) to enforce the representation of all relevant populations.

Second, *sampling bias* occurs when a subset of the population produces a disproportionate amount of the overall data. This type of bias has been shown to be closely related to economic measures: more wealthy populations produce more digital language per capita (Dunn and Adams, 2019). By default, a corpus will contain more samples representing wealthier members of the population. Thus,

Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, pages 1–10

https://doi.org/10.18653/v1/P17

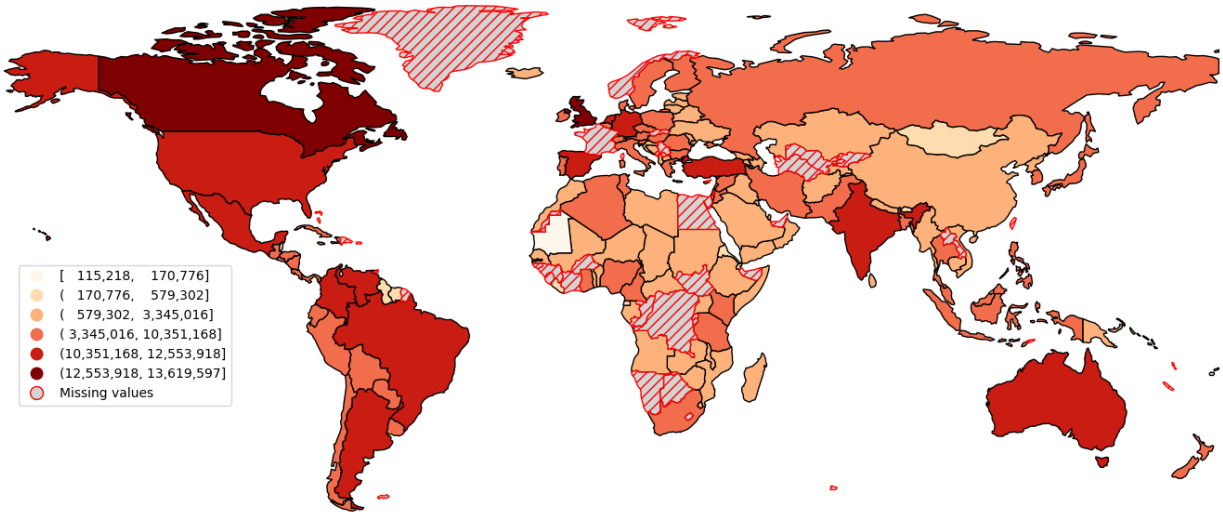

Figure 1: Number of observations per country.

this is similar to production bias, but with a demographic rather than a geographic scope.

Third, *non-local bias* is the problem of over-representing those people *in* a place who are not *from* that place: tourists, aid workers, students, short-term visitors, etc. For example, in countries with low per-capita GDP (i.e., where local populations often lack internet access) digital language data is likely to represent outsiders like aid workers. On the other hand, in countries with large numbers of international tourists (e.g., New Zealand), data sets are likely to instead be contaminated with samples from these tourists.

Fourth, *majority language bias* occurs when a multi-lingual population only uses some of its languages in digital contexts (Lackaff and Moner, 2016). Most often, majority languages like English and French are used online while minority languages are used in face-to-face contexts. The result is that even though an individual may be represented in a corpus, the full range of their linguistic behaviours is *not* represented. This is the only type of bias not quantified in this paper. For example, it is possible that changes in linguistic diversity are caused by a shift in behaviour, rather than a shift in population characteristics.

Of the three sources of bias that we examine here, non-local bias is the most difficult to uncover (Graham et al., 2014; Johnson et al., 2016). We can identify production bias when the amount of data per country exceeds that country's share of the global population. In this sense, the ideal corpus of English would equally represent each country according to the number of English speakers in that country. Within a country, we can measure the amount of sampling bias by looking at how economic measures like GDP and rates of internet access correspond with the amount of data per person. Thus, we could use median income by zip code to ensure that the US is properly represented. But non-local bias is more challenging because we need to know which samples from a place like New Zealand come from those speakers who are only passing through for a short time.

Only with widespread restrictions on international travel during the COVID-19 pandemic do we have access to a collection of digital language from which non-local populations are largely absent (Gössling et al., 2020; Hale et al., 2020). This paper uses changes in linguistic diversity during these travel restrictions, against a historical baseline, to calibrate computational measures that support language and population mapping. This is a part of the larger problem of estimating population characteristics from digital language data.

We start by describing the data used for the experiments in the paper (Section 2), drawn from Twitter over a two-year period. We then explore sources of bias in this data set by looking at production bias and sampling bias (in Section 3) and then developing a baseline of temporal variation in the data (in Section 4). We introduce a measure of geographic linguistic diversity (Section 5). Then we use this measure to find which countries and languages are most contaminated by non-local populations (in Section 6). Finally, we examine the results to find where the linguistic landscape has changed during the COVID-19 pandemic.

Region	N.	Pop	Data
Africa, Southern	12.28m	1.0%	2.0%
Africa, Sub	43.87m	10.1%	7.0%
Africa, North	16.60m	3.4%	2.7%
America, Brazil	10.96m	2.8%	1.8%
America, Central	66.12m	2.9%	10.6%
America, North	24.64m	4.8%	4.0%
America, South	77.79m	2.9%	12.5%
Asia, East	15.88m	22.3%	2.6%
Asia, Central	15.08m	2.7%	2.4%
Asia, South	30.06m	23.3%	4.8%
Asia, Southeast	31.88m	8.4%	5.1%
Europe, East	51.48m	2.4%	8.3%
Europe, Russia	9.38m	2.0%	1.5%
Europe, West	155.74m	5.7%	25.0%
Middle East	36.58m	4.5%	5.9%
Oceania	24.92m	0.8%	4.0%
Total	**623.33m**	**100%**	**100%**

Table 1: Distribution of data by region.

2 Data sources

We draw on Twitter data sampled globally from 10k cities over a 25-month period (July 2018 through August 2020). This city-based collection reduces production bias from the start (as opposed to collecting data by user or search term) because it forces non-central cities to be included. The cities are selected to represent the global population and all retweets are removed. This provides 623 million tweets, distributed across regions as shown in Table 1 with each region's share of the data and of the world's population.

This table provides a clear illustration of production bias. East Asia, for example, accounts for 22.3% of the world's population but only 2.6% of the data. We see the reverse in Western Europe, which provides 25% of the data but only 5.7% of the population. Population-based sampling is an effective method for correcting this bias (Dunn and Adams, 2020), if the goal is to produce a corpus representing the actual distribution of speakers. Our goal here is to find which countries contain data from non-local populations. To do this, we need to find out if the data has a stable geographic distribution that is driven by the underlying population.

The idNet language identification package is used to provide language labels (Dunn, 2020). Any tweet under 40 characters (after cleaning URLs and hashtags) is removed because of reduced identification accuracy below this threshold. The average tweets per month per country is visualized in Figure 1. Because we are looking at change over time by country, the data is binned into potentially small categories (e.g., Nigeria in July 2019). Both the table and the map show that countries in East Asia are under-represented. Thus, we use significance testing *within* countries when looking for change over time.

3 Demographics and language use

The next question is the degree to which the production of this data is driven by underlying populations (potential production bias) and by demographic factors like GDP (potential selection bias). We start, in Figure 3, by looking at the relationship between each country's population and share of the corpus. This expands on the region aggregations in Table 1 by dividing regions into countries. Each country is an observation that is represented by its average monthly data production and several demographic factors. Overall, there is a very significant correlation (Pearson) between population and the amount of data from each country (0.46). Thus, the number of people in a country is an important factor explaining how much data that country produces. While this is significant, however, it also means that there are many other factors that influence the geographic distribution of the data.

To better understand the factors influencing the geographic distribution of the data, we work with three variables: *population*, the number of people in each country; *internet population*, the number of internet users in each country; and *GDP*, a measure of each country's economic output (United Nations, 2011, 2017b,a). Figure 3 shows three regression plots in which these variables (on the y axis) are compared with the average monthly data production per country (given in number of tweets per month on the x axis).

In each case, there is a close relationship between data production and demographics, with several extreme outliers. For *population*, the outliers are China and India. Both are highly populated countries with significantly lower than expected data production (especially China). Both countries have relatively low rates of internet access: 38% for China and 11% for India; this lowers the total population in each country. Thus, although the populations are quite large, most of the population is not able to produce digital language data. For the influence of GDP, the outliers are the US and China.

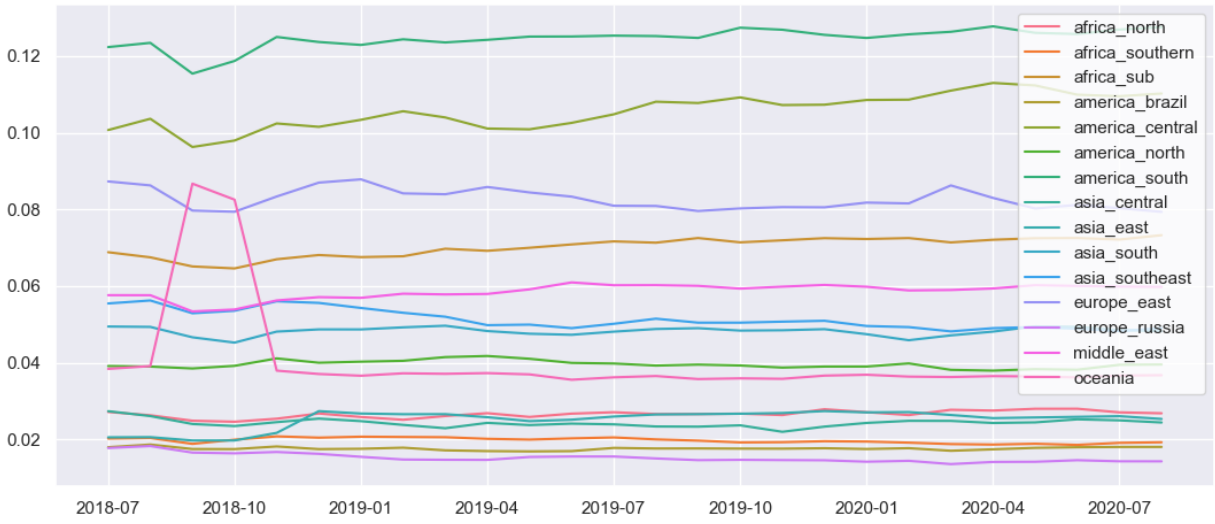

Figure 2: Geographic distribution of data by region by month.

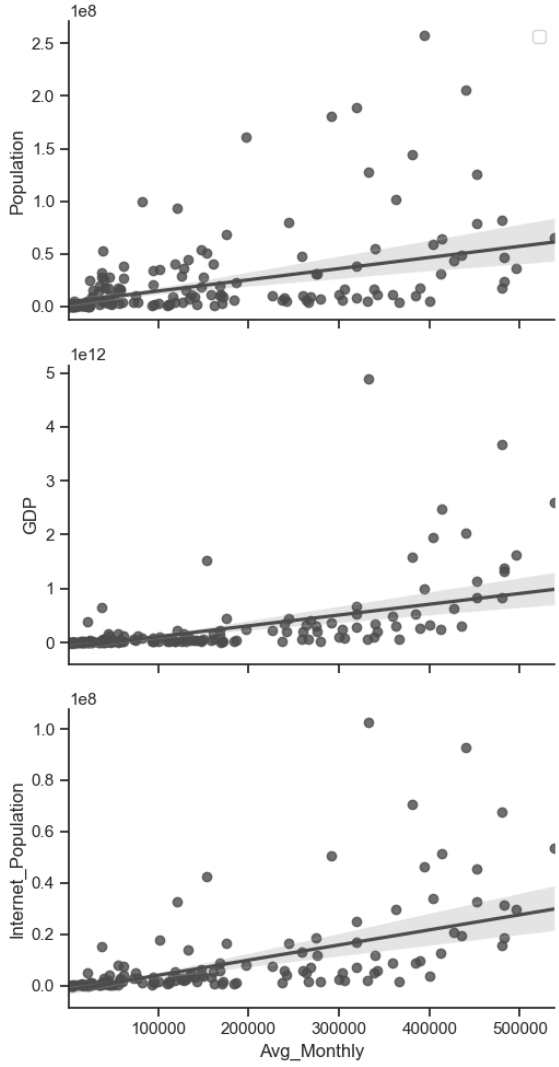

Figure 3: Relationship between data and demographic factors: *Population*, *Internet Access*, and *GDP* (With Outliers Removed).

For the US, in particular, the GDP is quite high: there seems to be a ceiling after which increased GDP is unlikely to influence digital behaviours. Further, that GDP is not evenly distributed across the entire population. For the influence of internet access, the outliers are again China and the US. With a few notable exceptions there is a relatively close relationship between data production and the demographic factors of each country.

With these three outliers removed (the US, China, India), there are very significant correlations between these three variables and the geographic distribution of the data: 0.46 (population), 0.61 (population with internet access), and 0.59 (GDP). This leaves some unexplained production factors. The most obvious missing factor here is social media platforms specific to given countries (e.g., Sina Weibo). These alternative platforms will siphon away enough users to distort the representation of a population given access only to other platforms. Further, Twitter is banned in China: because only some companies are allowed to use it through specific VPNs, the text is not representative of language use in China. Casual users of Twitter will use a VPN through another country which would distort this method of data collection.

Regardless, this section has shown that we can explain a significant portion of the geographic distribution of the data. This is important because we want to describe *populations* by observing *digital corpora*. If there is no relationship between the two in terms of distribution, it is difficult to make such inferences. What we have seen, however, is that there is a very significant relationship. What

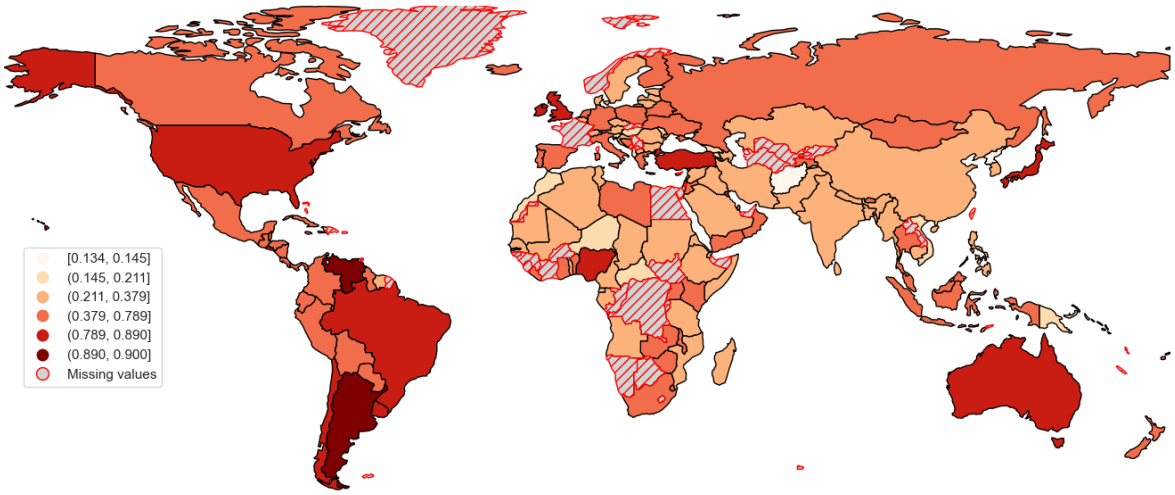

Figure 4: Herfindahl-Hirschman Index of the distribution of languages by country.

is the required threshold for establishing a relationship like this? We should think about this as a metric for evaluating digital corpora: data with a stronger relationship to demographic variables are more representative. The next question is whether this relationship remains stable over time: can we depend on these demographic factors across the entire period?

4 Controlling for temporal variation

The next question is whether these production factors are stable over time. Here we build a baseline for temporal variation: to what degree is the data subject to unrelated fluctuations that will reduce our ability to assign a cause-and-effect relationship to linguistic diversity during travel restrictions?

Although the same collection and processing methods are maintained over the two-year period, there is variation in the total number of observations (tweets) per month. There are many reasons why this might be the case. What matters to us, though, is the relative share of each country. In other words, the population does not change from month to month in the same way that the number of tweets changes. Regardless of the total amount of data collected per month, is the geographic distribution consistent? Figure 2 shows stability over time by representing the relative proportion of observations per region by month. Western Europe is removed for the sake of clarity, as it represents a significantly higher share (roughly 25%). The distribution of samples is consistent over time. The main exception is that, for a two-month period in 2018, there is much more data from Oceania.

We use a t-test to find out if the share of each region is stable over time. If the distribution changes significantly, then it may be hard to determine the cause of any individual change. None of the regions show a significant fluctuation; this is helpful because it shows that there is not random noise in the data that could interfere with measures of linguistic diversity. The difference-in-differences methods we use in Section 8 would control for such noise, but this gives us further confidence. We use a t-test, rather than a time-specific test like Dickey-Fuller, because we are interested in consistency rather than in non-stationarity. These results show that, in the aggregate, the distribution of samples remains constant. But how much variation within individual countries does this region-based measure disguise? To answer this, we look at the same t-test approach by country: do individual countries vary widely in their relative production? No countries show a significant change.

These findings show that we can largely focus on diversity, the distribution of languages within a country by month, rather than on the production of data over month as in Section 3. There is natural variation in the data, of course, and this is taken into account in our later approaches. For example, if we compute the correlation between population and language production (as in Section 3) for each month in isolation, there is no significant difference over time. This stability is important for creating a baseline against which to understand demographic changes during travel restrictions. Because the relationship between demographics and the data set remains stable, we can focus specifically on changes in linguistic diversity.

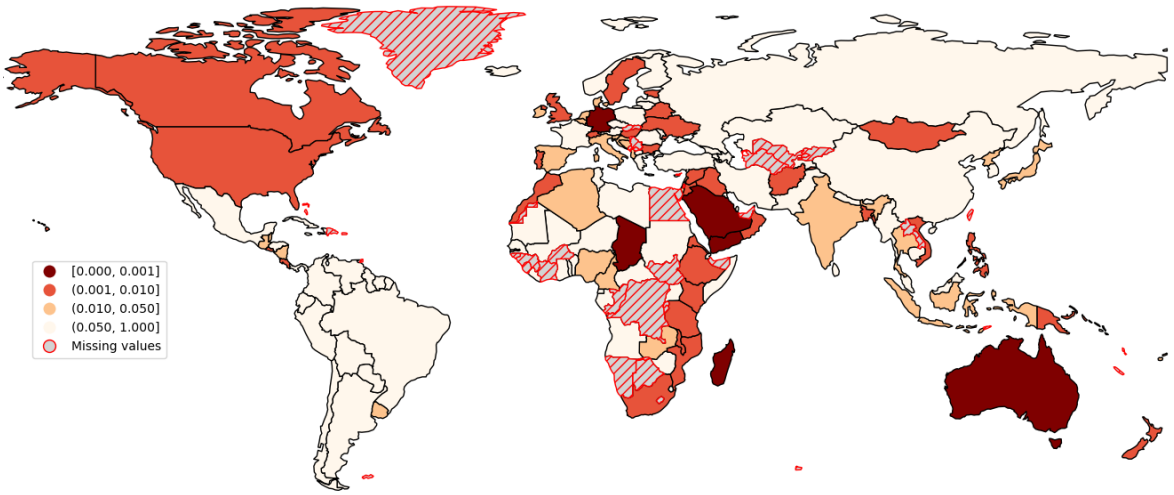

Figure 5: Countries with significant change in linguistic diversity during travel restrictions.

5 Measuring linguistic diversity

Linguistic diversity is an important part of accurate language and population mapping. The goal is to have a single measure that can tell us how much language contact is taking place and which communities are multi-lingual. To do this we must generalize across specific languages: linguistic diversity in the US might involve English and Spanish, but it might involve Portuguese and Spanish in Brazil.

We measure linguistic diversity as a probability distribution over languages for each country. Drawing on previous work on short-sample language identification, this paper includes 464 languages across 157 countries. For each country, then, we have a relatively accurate identification of which languages are used on Twitter. Given this probability distribution for each country, we compare countries using the Herfindahl-Hirschman Index (HHI) as shown in Figure 4. The HHI was developed in economics to measure market concentration: the more of a given industry is dominated by a small number of companies, the higher the HHI (Hirschman, 1945). The measure is derived using the sum of the square of shares, in this case the share of each language in each country. The higher the HHI (the darker red) for a country, the more one language dominates the linguistic landscape.

Thus, the HHI is higher when the distribution is centered around just a few languages. For example, in Table 2 we focus on three countries that show a range of linguistic diversity: Israel, India, and the US. Israel has the lowest HHI (0.207). Looking at the share of the top five languages, we see roughly equal usage of three languages (in the 20s)

	ISR	IND	USA
HHI	0.207	0.356	0.852
L1	27.3%	50.8%	92.3%
L2	25.9%	30.8%	2.6%
L3	23.5%	3.4%	0.6%
L4	7.5%	2.5%	0.6%
L5	5.3%	1.4%	0.4%

Table 2: Sample language distributions by country.

followed by two significant minority languages. This lower HHI reflects the fact that a number of languages are being used together: no language has a monopoly. On the other extreme, the US has one of the highest values for HHI (0.852). There is one very dominant language (92%), one significant minority language (2.6%), and a number of very insignificant languages. English has a metaphoric monopoly on the linguistic landscape of the US.

Figure 4 shows linguistic diversity across the world: lighter countries (like Israel) have a mix of languages while darker countries (like the US) are mostly monolingual. There are many linguistic landscapes around the world, ranging from multilingual to monolingual. This Figure 4 is a baseline representation, averaged across the entire time period (July 2018 to August 2020). It is possible that this averaged representation disguises temporal fluctuations. We have already seen that there are only a few changes in the share of data per country per month, and no significant change in the relationship between the data set and demographic factors like GDP. The question here is whether there is arbitrary variation in the linguistic diversity per country

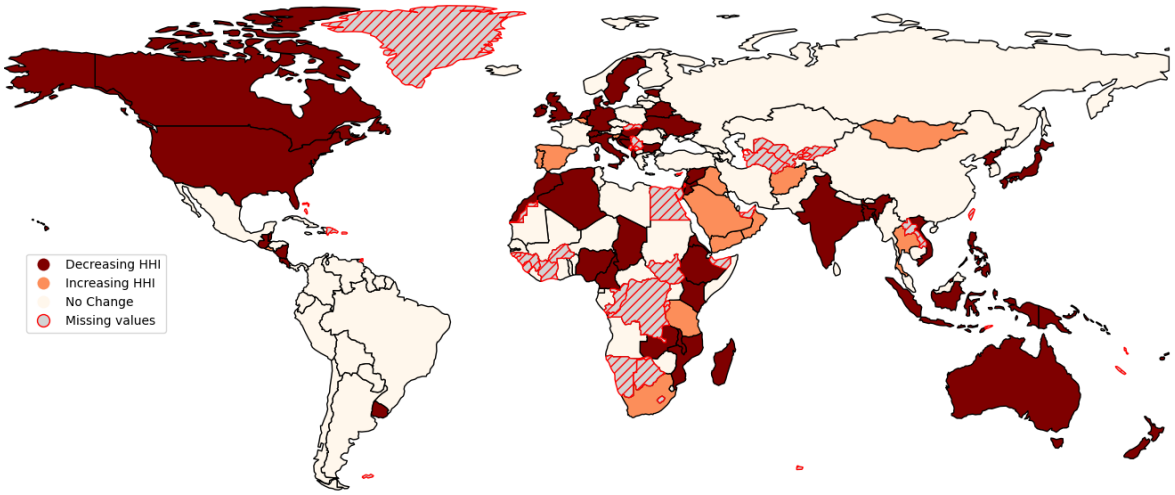

Figure 6: Increasing vs. decreasing HHI during travel restrictions.

per month. In other words, if Israel becomes significantly more diverse every three months, it will be difficult to find out what is causing those changes. We use a a t-test for the mean of each country to determine if each country's diversity is actually a single group. There are no significant fluctuations across the period as a whole.

6 Finding non-local populations

To what degree do countries change during travel restrictions resulting from COVID-19? We have a measure of diversity (the HHI) and data collected by month. The basic approach is to create two groups of samples: first, months during the pandemic (March through August, 2020); second, months not during the pandemic (March through August, 2019). These two groups are aligned by month so that seasonal fluctuations are taken into account (e.g., tourism high season in February for New Zealand and in July for Italy). Given these two groups of samples, we use a t-test for two independent samples to determine whether these groups are, in fact, different. If we reject the null hypothesis, it means that linguistic diversity during travel restrictions is significantly different than the seasonally-adjusted baseline.

The results show that 70 countries have a changed linguistic landscape during the pandemic. This is visualized in Figure 5, with p-values classed into highly significant (under 0.001), very significant (under 0.01), and significant (under 0.05). We see, for example, that the US and Canada undergo significant change, but not Mexico and South America. There are clear geographic patterns in linguistic change: North but not Central or South America; East Africa but not West Africa; South/east Asia but not East Asia; Europe but not Russia. We will examine in more detail how and why the linguistic landscape changes in Section 7.

These significant changes during international travel restrictions show that our measure (the HHI) and our data (tweets) offer a meaningful representation of underlying populations. If the data did not represent populations, we would not see the relationships examined in Section 3. There are no random fluctuations in the distribution of the data across countries or in the distribution of languages within countries. At the same time, given a massive social change (i.e., the COVID-19 pandemic), the measure clearly identifies changes in the linguistic landscape. Thus, the measure is both precise (not disguised by noise) and accurate (observing change where we expect it). The key point is that the change in diversity during the COVID-19 period is identifiable against the background noise.

A country's linguistic landscape could change by becoming more diverse (i.e., with more languages) or by becoming less diverse (i.e., with fewer languages). Which is causing the significant changes that we are observing? Figure 6 distinguishes between countries with an increasing HHI (becoming more monolingual) and a decreasing HHI (becoming more multilingual). We can think about two contexts in which this change can take place: a country like India might look more multilingual because non-local tourists who speak English are no longer creating noise in the data; or, a country like South Africa might look more monolingual

Country	Normal	COVID
Eritrea	63.16%	41.94%
Samoa	45.00%	30.18%
Cabo Verde	27.78%	16.63%
Equatorial Guinea	33.08%	24.40%
Madagascar	53.08%	44.87%
Kiribati	31.10%	23.56%
Tanzania	34.43%	27.35%
Mongolia	30.32%	23.52%
Chad	45.48%	39.71%
Sao Tome	12.57%	7.14%
Yemen	14.44%	9.20%

Table 3: Major reductions in English.

because its own English-speaking citizens abroad are returning home. The flow of international travellers changes the balance of locals and non-locals in both directions (leaving and coming home).

7 Identifying out-of-place populations

Our task now is to use these changes during travel restrictions to identify which populations are out-of-place in ordinary times. In other words, if India has decreasing English use during the pandemic period, then we know that English is over-represented in the country as a result of non-local populations. We find these languages by repeating the comparison of pandemic vs. normal periods per country per month, but now we look at the share of individual languages rather than the HHI (in countries with a significant change). We are only interested in languages which account for at least 1% of a country's usage. Less commonly used languages may be changing significantly but have less influence on a country's overall linguistic landscape.

We start by looking at countries where the use of English falls dramatically during the pandemic period, in Table 3. These dramatic reductions suggest that much of the population represented on Twitter is non-local: there is a change from 63% to 42% in Eritrea and from 53% to 44% English use in Madagascar. If the local population was well-represented on Twitter, we would not see this dramatic reduction in an international language. Thus, here we see an example of how digital data is biased towards non-local populations in countries where the local population has reduced internet access.

The influence of non-local populations returning home is shown for Russian and Arabic in Table 4. We see a major reduction in the use of Russian in countries like Ukraine that have had a strong Russian influence (from 54% to 50%). In both Ukraine and Belarus, there are other social and political factors that could influence the shift, since much of the population is bilingual (e.g., bilingual speakers in Ukraine putting aside the use of Russian for political purposes). But we also see similar changes in the use of Arabic. In Algeria it falls from 51% to 46% and in Morocco from 33% to 28%. These countries do not have the same political factors as Ukraine and Belarus, thus providing a clearer example of the exodus of non-local populations.

Country	Language	Normal	COVID
Belarus	Russian	69.05%	66.13%
Ukraine	Russian	54.60%	50.06%
Lithuania	Russian	20.09%	15.72%
Latvia	Russian	10.43%	8.26%
Algeria	Arabic	51.56%	46.77%
Morocco	Arabic	33.75%	28.53%
Israel	Arabic	27.75%	26.08%
Tunisia	Arabic	24.24%	19.65%
Bhutan	Arabic	6.25%	2.55%
Moldova	Arabic	2.71%	0.79%

Table 4: Major reductions in Russian and Arabic.

Country	Language	Normal	COVID
SAU	Arabic	70.10%	81.87%
SAU	English	12.18%	7.35%
SAU	Turkish	4.34%	2.12%
SAU	Greek	2.55%	1.65%
BEL	French	28.64%	34.72%
BEL	English	31.01%	26.83%
BEL	Dutch	27.08%	25.12%
BEL	German	2.26%	1.93%
BEL	Portuguese	1.51%	1.68%

Table 5: Changing landscape in Saudi Arabia and Belgium.

We get a different view by looking at the change of languages *within* a country, as with Belgium and Saudi Arabia in Table 5. In Saudi Arabia we see a rise in Arabic at the expense of English, Turkish, and Greek. This reflects the exodus of non-local tourists and workers; but it also likely reflects the return of Saudi Arabians from countries like Algeria and Morocco that is suggested by Table 4. In Belgium, we see a rise in French at the expense of English, Dutch, German, and Portuguese. This is a reflection of a reduction in non-local tourists.

However, we see the opposite effect of tourists

Country	Language	Normal	COVID
NZL	English	86.26%	84.13%
NZL	Spanish	2.13%	3.37%
NZL	Portuguese	2.30%	2.82%
NZL	Indonesian	0.89%	1.27%
AUS	English	89.51%	87.45%
AUS	Portuguese	1.83%	2.52%
AUS	Spanish	1.52%	2.08%
AUS	Japanese	0.99%	1.32%

Table 6: Changing landscape in Oceania.

leaving when we look at New Zealand and Australia, two countries which have had closed borders (Table 6). Here there is a *reduction* in English usage within English-majority countries that takes place when international tourists stop arriving. The situation here is that there are so many English-speaking tourists (i.e., from the US and UK) that local immigrant languages like Spanish and Portuguese (part of the long-term local population) are drowned out by non-local tourists using English. Another possible explanation is that immigrant populations are increasingly using Twitter to communicate with non-local populations (e.g., with family and friends in their previous country).

8 Sources of Change

This paper has shown that there is a significant change in the linguistic diversity of many countries *during* the travel restrictions caused by COVID-19. But to what degree are these changes *related* to the travel restrictions themselves? For example, we could imagine a population that is changing over time which we just happen to observe in mid-change. It could be the case that a country has been becoming less diverse over the past decade because of fewer incoming immigrants; the approach taken so far in this paper would misinterpret such macro-trends to be a direct result of COVID-19.

We use a difference-in-differences method (Card and Krueger, 1994) to correct for this. The basic idea behind a difference-in-differences approach is to conduct a *natural experiment* with a control group (here, data from 2018) and an effect group (here, data from 2020) differentiated by time. We have three months (July, August, September) that are shared across 2018, 2019, and 2020. So, using the same methods described above, we find out which countries have a significant change between 2019 and 2020. This is the period that takes place during travel restrictions. If travel restrictions influence linguistic diversity, we would expect such influence to take place during this period. We then find out if the countries which show a significant change in 2020 also show a significant change from 2018 to 2019. This provides a baseline: removing any country whose linguistic diversity was already in the process of changing.

Over this three-month period (July through September), 58 countries show a change in linguistic diversity during the pandemic. This is a smaller number than the main results reported above for two reasons: (i) the time span is shorter, giving less robust results and (ii) this particular time span came after some travel had resumed. Of these 58 countries that show a significant change in diversity, most (38) show no difference at all in the baseline period before the pandemic. Another eight show a much greater difference during the COVID-19 period (e.g., p-values of 0.03 vs 0.004 for baseline and COVID-19, respectively). This means that the pandemic has either created or has significantly contributed to 79.3% of the cases of changing linguistic diversity. The remaining 20.7% of changes, then, must have been created by macro-trends like immigration or changes in bilingual behaviour. The main conclusion from this difference-in-differences examination, however, is that most of these changes can be specifically connected to COVID-19.

9 Conclusions

The goal of this paper is to validate measures of linguistic diversity using changes in underlying populations during the COVID-19 pandemic. We have shown that there is a significant relationship between our data and the underlying population. Thus, what we are observing (tweets) can tell us about the people we want to study. At the same time, both the distribution of the data across countries and the distribution of languages within countries are stable. Thus, the data does not have random fluctuations that will get in the way. Using the HHI as a measure of diversity, there is a significant change in the linguistic landscape of 70 countries against a seasonally-adjusted baseline. This reflects non-local populations (e.g., the impact of tourists leaving a country or short-term visitors returning to their own countries). These results validate a measure of linguistic diversity that is based on digital language data and shows that we can correct for the bias introduced by non-local populations.

References

David Card and Alan Krueger. 1994. Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *American Economic Review*, 84.

Paul Cook and Laurel J Brinton. 2017. Building and evaluating web corpora representing national varieties of English. *Language Resources and Evaluation*, 51(3):643–662.

Jonathan Dunn. 2019a. Global Syntactic Variation in Seven Languages: Towards a Computational Dialectology. *Frontiers in Artificial Intelligence*, 2:1–15.

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

Jonathan Dunn. 2020. Mapping languages: the Corpus of Global Language Use. *Language Resources and Evaluation*.

Jonathan Dunn and Benjamin Adams. 2019. Mapping languages and demographics with georeferenced corpora. In *Geocomputation 2019*.

Jonathan Dunn and Benjamin Adams. 2020. Geographically-balanced gigaword corpora for 50 language varieties. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2528–2536. European Language Resources Association.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2020. *Ethnologue: Languages of the World. Twenty-third edition*. SIL International, Dallas, TX.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2014. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114.

Bruno Gonçalves and David Sánchez. 2014. Crowdsourcing dialect characterization through Twitter. *PloS one*, 9(11):e112074.

Stefan Gössling, Daniel Scott, and C Michael Hall. 2020. Pandemics, tourism and global change: a rapid assessment of COVID-19. *Journal of Sustainable Tourism*, pages 1–20.

Mark Graham, Scott Hale, and Devin Gaffney. 2014. Where in the World are You? Geolocation and Language Identification on Twitter. *The Professional Geographer*, 66(4):568–578.

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.

Thomas Hale, Anna Petherick, Toby Phillips, and Samuel Webster. 2020. Variation in government responses to COVID-19. *Blavatnik school of government working paper*, 31.

Albert O Hirschman. 1945. *National power and the structure of foreign trade*. Univ of California Press.

IMB. 2020. *People Groups Data, 2020-08*. International Missionary Board: Global Research, Richmond, VA.

Isaac L Johnson, Subhasree Sengupta, Johannes Schöning, and Brent Hecht. 2016. The geography and importance of localness in geotagged social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 515–526. Association for Computing Machinery.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating Dialectal Variability for Socially Equitable Language Identification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 51–57. Association for Computational Linguistics.

Juhi Kulshrestha, Farshad Kooti, Ashkan Nikravesh, and Krishna P Gummadi. 2012. Geographic dissection of the Twitter network. In *Sixth international AAAI conference on weblogs and social media*, pages 202–209. Association for the Advancement of Artificial Intelligence.

Derek Lackaff and William J Moner. 2016. Local languages, global networks: Mobile design for minority language users. In *Proceedings of the 34th ACM International Conference on the Design of Communication*, pages 1–9. Association for Computing Machinery.

Fabio Lamanna, Maxime Lenormand, María Henar Salas-Olmedo, Gustavo Romanillos, Bruno Gonçalves, and José J Ramasco. 2018. Immigrant community integration in world cities. *PloS one*, 13(3):e0191612.

Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. 2013. The Twitter of Babel: Mapping world languages through microblogging platforms. *PloS one*, 8(4):e61981.

United Nations. 2011. *Economic and Social Statistics on the Countries and Territories of the World, with Particular Reference to Childrens Well-Being*. United Nations Children's Fund.

United Nations. 2017a. *National Accounts Estimates of Main Aggregates. Per Capita GDP at Current Prices in US Dollars*. United Nations Statistics Division.

United Nations. 2017b. *World Population Prospects: The 2017 Revision, DVD Edition*. United Nations Population Division.