*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 2528–2536
Marseille, 11–16 May 2020

# Geographically-Balanced Gigaword Corpora for 50 Language Varieties

**Jonathan Dunn, Benjamin Adams**
University of Canterbury
Christchurch, New Zealand
{jonathan.dunn, benjamin.adams}@canterbury.ac.nz

### Abstract

While text corpora have been steadily increasing in overall size, even very large corpora are not designed to represent global population demographics. For example, recent work has shown that existing English gigaword corpora over-represent inner-circle varieties from the US and the UK (Dunn, 2019b). To correct implicit geographic and demographic biases, this paper uses country-level population demographics to guide the construction of gigaword web corpora. The resulting corpora explicitly match the ground-truth geographic distribution of each language, thus equally representing language users from around the world. This is important because it ensures that speakers of under-resourced language varieties (i.e., Indian English or Algerian French) are represented, both in the corpora themselves but also in derivative resources like word embeddings.

**Keywords:** corpus building, geo-referenced corpus, web as corpus, language mapping, under-resourced language varieties

## 1. Under-Represented Populations

Where does digital language data come from and how well does it match up with the geographic distribution of human populations? This is an important question because NLP now depends on large text corpora (Baroni et al., 2009; Goldhahn et al., 2012; Majliš and Žabokrtský, 2012; Benko, 2014) that are derived from digital sources like web pages and social media. This paper raises two more cognate questions. First, are there populations that existing corpora fail to represent? Second, is there a significant difference between regional varieties of languages, so that, e.g., a model trained on American English would work poorly on Nigerian English (Davies and Fuchs, 2015; Cook and Brinton, 2017)? If the answer to both of these questions is *yes*, it follows that NLP needs to switch to geographically-balanced datasets. This paper derives such geographically-balanced gigaword corpora from the 423 billion word Corpus of Global Language Use (Dunn, 2020). The resulting family of corpora, GeoWAC, is evaluated against the CoNLL 2017 Shared Task data (Ginter et al., 2017) to determine the differences between otherwise comparable balanced and unbalanced corpora.

In Section 2 we review recent work showing that there are strong geographic effects present in digital language sources. Then in Section 3 we describe the collection and cleaning methods used to produce the GeoWAC corpus family. In Section 4 we develop a population-based sampling method that is used to adjust the amount of data drawn from each country. Section 5 describes the geographic distribution of the corpora that this sampling method produces. Section 6 evaluates the GeoWAC corpora against a baseline dataset that is not collected with geographic information, using both a corpus similarity evaluation and a word embedding evaluation. The primary contribution of this paper is to systematically remove the geographic biases that are present in existing gigaword corpora and to evaluate the difference between biased and unbiased datasets. These corpora are available in full under the GNU GPL license.[1]

## 2. Is Geographic Variation Significant?

An initial justification for the development of geographically-balanced gigaword corpora is based on a simple question: is there a significant difference between national varieties of languages (i.e., American English vs. Indian English or Cuban Spanish vs. European Spanish)? The more *linguistic variation* there is across national varieties, the more important it is to provide a balanced representation when training models in NLP.

First, we know from a long tradition of small-scale linguistic studies of non-digital language use that there is significant variation across geographic varieties (Erich, 2010). Do digital data sources continue to reflect this same variation or do digital registers constitute a non-geographic context in which geography is irrelevant? Recent work has shown that there is a significant correlation between manually-collected linguistic variants from the BBC Voices project (Upton, 2013) and automatically-collected linguistic variants from geo-referenced Twitter usage in the UK (Grieve et al., 2019). Another study has shown that there are strong correlations between web corpora from Canada and traditional Canadian dialect features (Cook and Brinton, 2017). While much of this type of work is based on English in inner-circle countries, there is nevertheless a consistent link between traditional studies of geographic variation and studies based on digital language data.

Second, we know from recent work on computational dialect modeling that both lexical and syntactic representations support the identification of national varieties with a high degree of accuracy, both for English (Dunn, 2019c) and other international languages (Dunn, 2019b). Furthermore, models of English dialects often group together inner-circle varieties separately from outer-circle varieties (Dunn, 2019b; Szmrecsanyi et al., 2019). This shows that there is a hierarchical organization of geographic variation, as predicted by the World Englishes paradigm (Kachru, 1982). Finally, work on grammar induction has shown that a construction grammar trained on biased CoNLL datasets has a significantly better fit on data from inner-circle countries than data from outer-circle countries (Dunn, 2019a).

[1] `www.earthlings.io/geowac_map.html`

**Variables**

$Population$ = The number of people living in a country (from UN estimates)
$Pct\_Internet\_Access$ = Relative internet availability in a country (by percent of population)
$Digital\_Population$ = $Population * Pct\_Internet\_Access$ (The digital population of a country)
$Pct\_of\_Country$ = A specific language's share of a country's digital language production
$Digital\_Lang\_Population$ = $Digital\_Population * Pct\_of\_Country$ (A language's share of the digital population)
$Threshold$ = Number of words desired in the corpus

**Algorithm**

```
while $N\_Words\_Total < Threshold$:
  for $country$ in corpus:
    $Target\_Pct = Digital\_Lang\_Population \div Global\_Total(Digital\_Lang\_Population)$
    $Current\_Pct = N\_Words\_Country \div N\_Words\_Total$
      if $Current\_Pct - Target\_Pct == GlobalMax$ :
        if $N\_Words\_Country > 1,000,000$ :
          $N\_Word\_Country = N\_Words\_Country - 1,000$
```

Table 1: Population-Based Sampling Algorithm.

These two pieces of evidence, the correspondence between traditional and digital dialect studies as well as the ability to accurately distinguish between geographically-defined language varieties, indicate that there is a significant amount of geographic variation in digital datasets. The implication, then, is that a corpus of American English or Swiss French does not provide an adequate sample of Indian English or Algerian French. Given the technical importance of large background corpora for training models in NLP and the social importance of increasing the demographic representation of models in NLP, the goal of GeoWAC is to provide data for under-resourced language varieties.

## 3. Collecting Geo-Referenced Documents

The data for this paper comes from the Common Crawl,[2] as processed in the Corpus of Global Language Use (henceforth, CGLU). This project includes the Common Crawl data from March 2014 until June 2019, a total of 147 billion web pages. This 423 billion word dataset has previously been visualized to show the underlying geographic biases of both web data and Twitter data.[3] The contribution of this paper is to produce and evaluate usable balanced corpora out of this much larger and imbalanced dataset.

The original corpus is sorted by language using the idNet language identification software,[4] assigning each web page to a single language. Here this dataset is further cleaned by splitting web pages by paragraph tags, deduplicating by paragraph, and checking language identification using the CLD2 and CLD3 language identification models.[5] [6]

This method produces gigaword corpora for 48 languages (with English divided into separate inner-circle, outer-circle, and expanding-circle corpora). In order to balance the geographic distribution of the corpora, ground-truth demographic data from individual countries is used to down-sample the amount of data per country until the corpus matches population demographics as closely as possible.

## 4. Population-based Sampling

How do we determine the proportion of each language's corpus that should come from a specific country? This section describes a population-based sampling method that considers three pieces of information: first, UN estimates of the population of each country (United Nations, 2017); second, the number of people in each country with internet access (United Nations, 2011); and third, the percentage of digital language use from each country that belongs to a specific language (Dunn and Adams, 2019).

Ideally, the number of words in each corpus would be proportionate to the number of people in each country who use that language in digital contexts. In other words, if Ireland accounts for 5% of the English-speaking digital population of the world, then Irish English should account for 5% of the corpus of English.

The sampling algorithm, shown in Table 1, first calculates the geographic distribution of the digital population for each language. The variable $Digital\_Population$ represents the number of internet users in each country. The variable $Digital\_Lang\_Population$ is thus the relative share of the digital population that is allocated to a specific language like English or French or Russian. This is calculated against the entire CGLU corpus (Dunn, 2020) because census-based language use estimates are not available for many countries and because there is a possible disconnect between digital and non-digital language use.

Given the digital language population for each country, we use the global sum (i.e., the world population of Spanish users) to determine the relative proportion of the overall corpus (i.e., for Spanish) that each country should contribute. This provides the target sampling: the percentage of words from a country in the Spanish corpus should match that country's percentage of global Spanish speakers. The only exception is that no country is down-sampled below 1 million words, allowing the corpora to maintain a broad geographic distribution.

This ideal case is not always possible because in cases like South Asia and Sub-Saharan Africa the actual population is under-represented (Dunn and Adams, 2019). Such imbalances are reduced, but not eliminated, by basing the sam-

[2] https://www.commoncrawl.org
[3] https://www.earthlings.io
[4] https://github.com/jonathandunn/idNet
[5] https://github.com/GregBowyer/cld2-cffi
[6] https://github.com/bsolomon1124/pycld3

| ISO3 Code | Language Name | Total Words | Share of Largest Country |
|---|---|---|---|
| ara | Arabic | 618 mil | 16.56% |
| aze | Azerbaijani | 204 mil | 97.33% |
| bel | Belarusian | 71 mil | 95.76% |
| bul | Bulgarian | 906 mil | 98.06% |
| cat | Catalan | 103 mil | 65.18% |
| ces | Czech | 1,117 mil | 86.61% |
| dan | Danish | 1,054 mil | 94.40% |
| deu | German | 1,108 mil | 18.33% |
| ell | Greek | 1,053 mil | 98.69% |
| eng | English (Inner) | 2,042 mil | 24.98% |
| eng | English (Outer) | 1,909 mil | 39.04% |
| eng | English (Expanding) | 2,100 mil | 1.48% |
| est | Estonian | 492 mil | 98.67% |
| fas | Farsi | 1,124 mil | 89.65% |
| fin | Finnish | 1,100 mil | 96.98% |
| fra | French | 2,085 mil | 16.92% |
| gle | Irish | 21 mil | 96.96% |
| hbs | Serbo-Croatian | 1,036 mil | 50.10% |
| hin | Hindi | 231 mil | 94.22% |
| hun | Hungarian | 1,100 mil | 89.98% |
| ind | Indonesian | 438 mil | 44.01% |
| isl | Icelandic | 180 mil | 98.89% |
| ita | Italian | 1,097 mil | 77.63% |
| jpn | Japanese | 1,099 mil | 25.09% |
| kat | Georgian | 137 mil | 99.07% |
| kaz | Kazakh | 95 mil | 96.85% |
| kor | Korean | 294 mil | 88.80% |
| lav | Latvian | 282 mil | 97.96% |
| lit | Lithuanian | 787 mil | 99.13% |
| mkd | Macedonian | 119 mil | 98.06% |
| mon | Mongolian | 121 mil | 99.24% |
| nld | Dutch | 1,136 mil | 44.93% |
| nor | Norwegian | 1,191 mil | 94.72% |
| pol | Polish | 1,107 mil | 86.87% |
| por | Portuguese | 1,779 mil | 67.41% |
| ron | Romanian | 1,062 mil | 83.11% |
| rus | Russian | 2,128 mil | 8.82% |
| slk | Slovak | 1,109 mil | 94.89% |
| slv | Slovenian | 481 mil | 98.25% |
| spa | Spanish | 2,051 mil | 9.32% |
| sqi | Albanian | 126 mil | 76.48% |
| swe | Swedish | 1,099 mil | 86.45% |
| tam | Tamil | 87 mil | 77.45% |
| tgl | Tagalog | 28 mil | 81.81% |
| tur | Turkish | 142 mil | 24.63% |
| ukr | Ukrainian | 517 mil | 88.77% |
| urd | Urdu | 46 mil | 77.53% |
| uzb | Uzbek | 39 mil | 98.10% |
| vie | Vietnamese | 1,100 mil | 90.31% |
| zho | Chinese | 2,099 mil | 54.59% |

Table 2: GeoWAC Corpus Family.

pling on digital populations (after adjusting for rates of internet access) rather than on actual populations.

In practice, there are four classes of languages in the GeoWAC corpus family. First, some languages (like Hindi or Georgian) do not reach the billion word threshold; these languages are not down-sampled at all. Rather, each corpus reports the geographic distribution of the data that is available. While this approach does not create geographically-balanced corpora, it does move a step forward in explicitly showing the biases of the corpora.

Second, some languages (like Greek and Bulgarian) reach the billion word threshold but are mainly used in one or two countries. These languages are balanced by population, but given the geographic concentration of Bulgarian speakers, for example, the GeoWAC Bulgarian corpus is unlikely to be significantly different than a baseline Bulgarian corpus.

Third, some languages (like Chinese and French) are used by a large international community that spans many individual countries. These corpora are capped at 2 billion words each and fully balanced using the algorithm in Table 1. While population balancing is important for all languages, it is international languages like these that are most likely to be imbalanced in the first place.

Fourth, English in digital contexts is unusually prominent, even in countries that are not traditionally English-speaking. This would lead, given the population-based sampling algorithm, to an under-representation of prototypically English-speaking countries. Thus, the English corpora are divided into 2 billion word sets representing inner-circle, outer-circle, and expanding circle varieties (Kachru, 1982). This allows a balanced version with traditional countries (the US, the UK, Canada, Ireland, Australia, New Zealand) while also capturing the ever-expanding reach of English as an international language.

## 5. Corpus Descriptions

The fifty corpora in the GeoWAC corpus family are shown in Table 2 by language, with the number of words as well as the maximum contribution of a single country for each corpus. For example, the Vietnamese corpus is 1.1 billion words, 90% of which is from Vietnam; the Chinese corpus is 2 billion words, 54% of which is from China. This gives an initial rough estimate of the geographic distribution of each corpus: the smaller the maximum value, the more dispersed the corpus is across countries.

First, as discussed above, we have smaller corpora which are unsampled because they are under a billion words in total; for these, the distribution is reported but not balanced. There are 24 of these unbalanced language corpora; 14 of these languages are drawn mainly (above 90%) from a single country. Four of the unbalanced corpora, however, are widely distributed: Arabic, Turkish, Indonesian, and Catalan. The point is that, even without population-based sampling, these four corpora are known to represent international populations. The full distribution of each language is also available as an external resource.[7]

[7] https://github.com/jonathandunn/earthLings/tree/master/GeoWAC

Second, we have gigaword corpora for 18 languages. Of these, four have a wide distribution (no more than 50% from a single country): German, Dutch, Japanese, and Serbo-Croatian. A further eight of these corpora are moderately distributed, with between 70% and 90% from a single country: Italian, Romanian, Swedish, Czech, Polish, Farsi, Hungarian, and Vietnamese. Each of these languages provides a case in which population-based sampling can remove geographic biases.

Third, we have eight gigaword corpora with 2 billion words each. These are specifically for widely dispersed languages: English (in three sub-sets), French, Portuguese, Russian, Spanish, and Chinese. These corpora are the most striking case in which population-based sampling is important for reducing geographic bias. For example, the largest country in the Russian corpus contributes only 8.82% of the total and the largest country in the corpus of expanding-circle English contributes only 1.48%. This dispersion means that the corpus does not represent only a single, restricted population. For example, the Russian corpus contains at least 100 million words each from Russia, Ukraine, Kazakhstan, Belarus, Kyrgyzstan, and Latvia; and over 60 million words each from Estonia, Uzbekistan, Ecuador, Azerbaijan, Moldova, and the United States.

To what degree do geographically concentrated languages benefit from population-based sampling? A good case study is German, which has approximately 200 million words each from Germany, Austria, and Switzerland. Thus, the main contribution to the corpus is from inner-circle countries. But there are also significant amounts of data (between 28 and 48 million words) from the United States, Sweden, Luxembourg, and Palau (once a part of German New Guinea). Similarly, the Serbo-Croatian corpus (or HBS, a cover term) is made up of large chunks from Croatia, from Serbia, and from Bosnia and Herzegovina. Thus, a geo-referenced approach is also important in cases where national boundaries fail to represent language populations. A similar situation is represented by Slovak (mostly from Slovakia but also from Czechia) and Czech (mostly from Czechia but also from Slovakia), in which geographic information can be triangulated against linguistic information.

## 5.1. Distribution of Major Languages

The languages which benefit most from geographic-balancing are those which are used across a large number of individual countries. In this section we take a closer look at inner-circle and outer-circle English, French, and Spanish. The goal is to examine the distribution of each corpus and, focusing on Spanish, to compare the balanced and unbalanced alternatives. The full distribution for each language is available as an external resource.[8]

The distribution for inner-circle Englishes is shown in Table 3 and for outer-circle Englishes in Table 4. In both cases the inventory of countries is pre-defined, but the relative share allocated to each country is determined by the sampling algorithm. For inner-circle Englishes, the corpus is equally divided between the four central varieties: the US, Canada, the UK, and Ireland. The representation of New Zealand and, especially, Australia, is significantly smaller. For outer-circle Englishes, the corpus is almost 40% from India, reflecting that country's high population.

| Country | N. Words | Pct. Corpus |
|---|---:|---:|
| Ireland | 510 mil | 24.98% |
| Canada | 510 mil | 24.98% |
| United States | 500 mil | 24.48% |
| United Kingdom | 464 mil | 22.74% |
| New Zealand | 46 mil | 2.29% |
| Australia | 10 mil | 0.53% |
| **TOTAL** | **2,042 mil** | **100.00%** |

Table 3: Distribution of Inner-Circle English Corpus.

| Country | N. Words | Pct. Corpus |
|---|---:|---:|
| India | 745 mil | 39.04% |
| Singapore | 358 mil | 18.80% |
| Philippines | 171 mil | 9.01% |
| Hong Kong | 164 mil | 8.63% |
| Nigeria | 159 mil | 8.34% |
| Pakistan | 144 mil | 7.58% |
| Malaysia | 131 mil | 6.89% |
| **TOTAL** | **1,909 mil** | **100.00%** |

Table 4: Distribution of Outer-Circle English Corpus.

French, in Table 5, is not divided into pre-defined inner-circle and outer-circle varieties; although French has a similar colonial history, such a division is not as common as for English. And yet there is a natural division between highly represented inner-circle varieties (France, Canada, Belgium, Switzerland) and a mix of less-represented outer-circle (Morocco, Senegal) and expanding-circle varieties (the US). The colonial history of France is represented by both African (Morocco, Algeria, Senegal) and South Pacific varieties (Viet Nam, French Polynesia).

| Country | N. Words | Pct. Corpus |
|---|---:|---:|
| France | 352 mil | 16.92% |
| Canada | 352 mil | 16.89% |
| Belgium | 348 mil | 16.71% |
| Switzerland | 284 mil | 13.65% |
| Morocco | 97 mil | 4.67% |
| Colombia | 95 mil | 4.58% |
| Luxembourg | 66 mil | 3.20% |
| United States | 39 mil | 1.91% |
| Italy | 29 mil | 1.40% |
| Senegal | 24 mil | 1.20% |
| Russia | 21 mil | 1.05% |
| Gabon | 19 mil | 0.91% |
| Algeria | 18 mil | 0.89% |
| Spain | 17 mil | 0.86% |
| Viet Nam | 17 mil | 0.85% |
| French Polynesia | 17 mil | 0.84% |
| Central African Rep. | 15 mil | 0.75% |
| Tunisia | 14 mil | 0.68% |
| **TOTAL** | **2,085 mil** | **100.00%** |

Table 5: Distribution of French Corpus.

[8] https://github.com/jonathandunn/earthLings/tree/master/GeoWAC

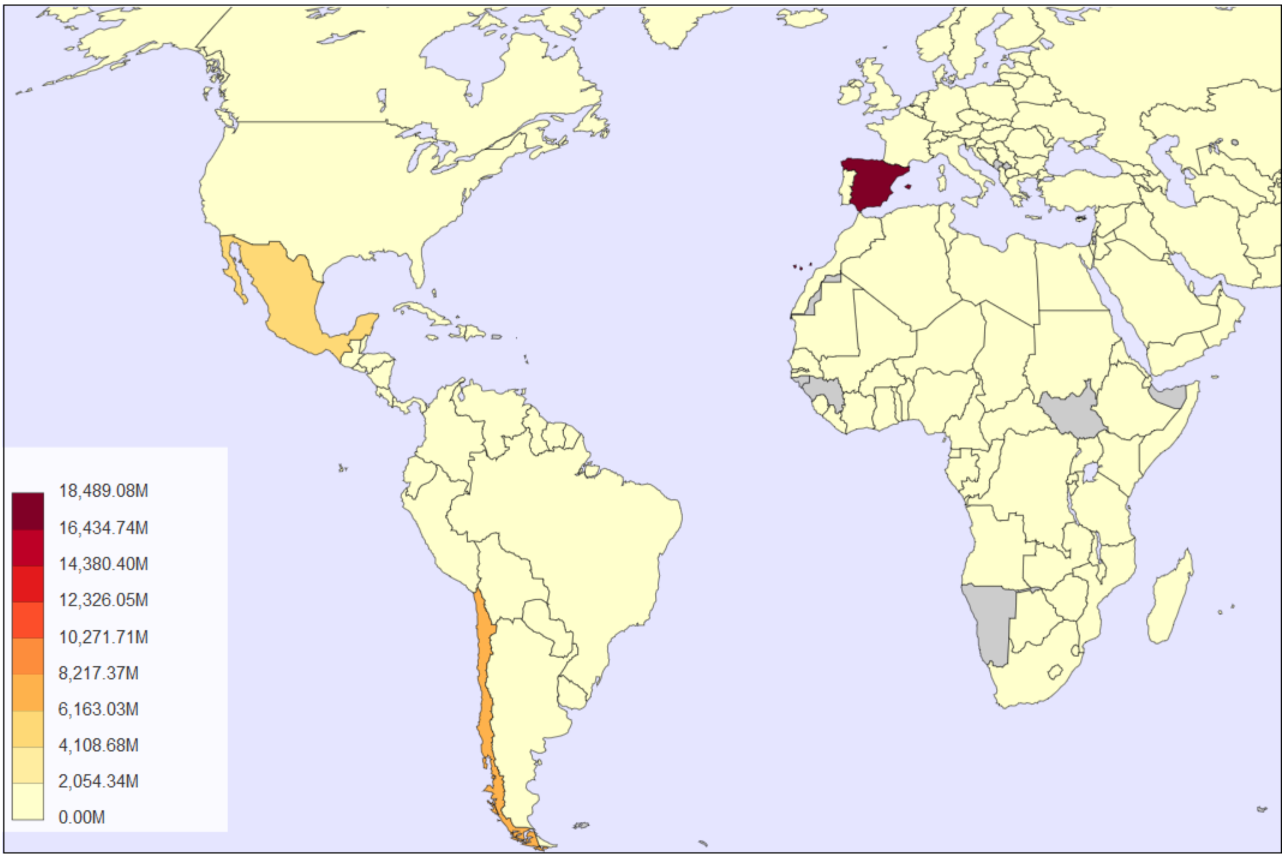

Figure 1: Distribution of Spanish without Population-based Sampling.

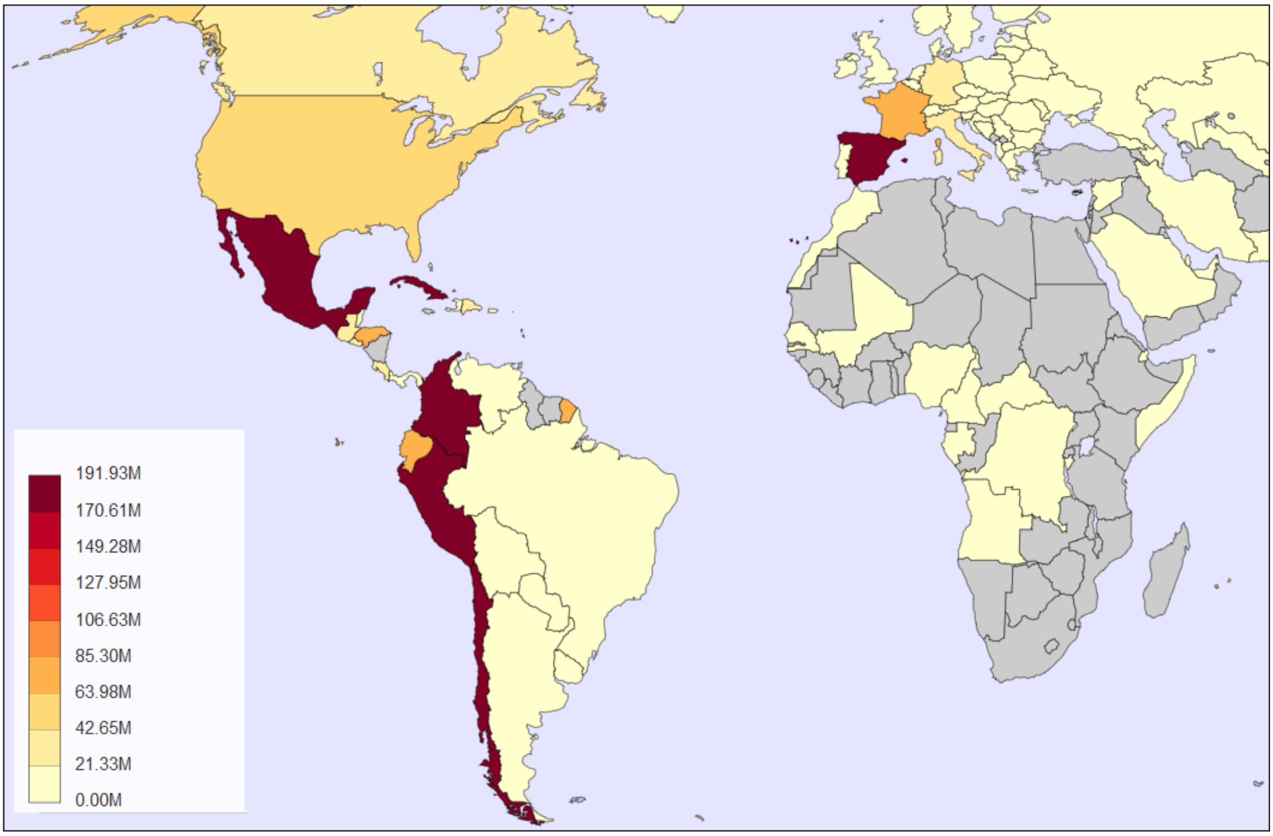

Figure 2: Distribution of Spanish with Population-based Sampling.

| **Country** | **N. Words** | **Pct. Corpus** |
|---|---:|---:|
| Spain | 191 mil | 9.32% |
| Cuba | 190 mil | 9.31% |
| Chile | 190 mil | 9.29% |
| Colombia | 189 mil | 9.24% |
| Mexico | 188 mil | 9.20% |
| Peru | 188 mil | 9.19% |
| Ecuador | 70 mil | 3.45% |
| Timor-Leste | 66 mil | 3.26% |
| Honduras | 66 mil | 3.24% |
| France | 65 mil | 3.19% |
| United States | 55 mil | 2.71% |
| Liechtenstein | 42 mil | 2.09% |
| Guatemala | 34 mil | 1.68% |
| Canada | 34 mil | 1.66% |
| Dominican Republic | 32 mil | 1.59% |
| Costa Rica | 26 mil | 1.28% |
| El Salvador | 25 mil | 1.22% |
| **TOTAL** | **2,051 mil** | **100.00%** |

Table 6: Distribution of Spanish Corpus.

The distribution of the Spanish corpus is shown in Table 6. Here again the corpus is naturally segmented into major varieties (over 100 million words per country: Spain, Cuba, Chile, Colombia, Mexico, Peru) and less major varieties (under 100 million words: Ecuador, Honduras, Guatemala, Costa Rica). This reflects the size of the digital population of Spanish users in each country.

The importance of population-based sampling can be visualized by comparing the maps of the distribution of Spanish without population-based sampling (in Figure 1) and with population-based sampling (in Figure 2). In the first case, the sheer amount of data from Spain over-shadows data from all countries except Mexico and Chile. Drawn at random, a corpus would almost entirely represent Spain. In the second case, however, population-based sampling has reduced this hegemony, with substantial representations from across South America and Central America.

### 5.2. Country-Specific Corpora

While the overall goal of GeoWAC is to provide unbiased corpora, there are certain applications for which large single country corpora are useful. For languages like Greek or Estonian, most of the corpus already comes from a single country. But for languages like Russian the amount of data for any given country is limited in the interest of representing many countries. In order to provide extra representation for major varieties of major languages, we also provide large country-specific corpora for languages which are widely distributed across countries in the main GeoWAC dataset: Arabic, Chinese, English, French, Portuguese, Spanish, and Russian. The summary of these country-specific corpora is shown in Table 7, with the language and country information together with the total number of words in each corpus.

This provides gigaword corpora for the central varieties of major languages: English in America, French in France, Portuguese in Portugal, Spanish in Spain, Russian in Russia. While the main family of corpora provide geographically balanced representations, these corpora provide geographically homogenous representations. As before, these corpora are available in full under the GNU GPL License.[9]

| **Language** | **Country** | **Total** |
|---|---|---:|
| **Name** | **Name** | **Words** |
| Arabic | United Arab Emirates | 102 mil |
| Chinese | China | 1,103 mil |
| Chinese | Taiwan | 282 mil |
| Chinese | United States | 138 mil |
| English | Canada | 1,018 mil |
| English | France | 706 mil |
| English | India | 979 mil |
| English | Ireland | 802 mil |
| English | Netherlands | 443 mil |
| English | Russia | 622 mil |
| English | Singapore | 383 mil |
| English | United Kingdom | 509 mil |
| English | United States | 1,072 mil |
| French | Belgium | 505 mil |
| French | Canada | 539 mil |
| French | France | 1,619 mil |
| French | Switzerland | 287 mil |
| Portuguese | Brazil | 237 mil |
| Portuguese | Portugal | 970 mil |
| Russian | Belarus | 712 mil |
| Russian | Kazakhstan | 574 mil |
| Russian | Kyrgyzstan | 127 mil |
| Russian | Latvia | 123 mil |
| Russian | Russia | 1,085 mil |
| Russian | Ukraine | 828 mil |
| Spanish | Chile | 845 mil |
| Spanish | Colombia | 355 mil |
| Spanish | Cuba | 235 mil |
| Spanish | Mexico | 746 mil |
| Spanish | Peru | 267 mil |
| Spanish | Spain | 1,014 mil |

Table 7: Inventory of Country-Specific Corpora.

## 6. Evaluation

This section evaluates the balanced GeoWAC corpora against the unbalanced but otherwise comparable CoNLL 2017 Shared Task data (Ginter et al., 2017). The question is whether it makes a difference to collect geo-referenced data and then balance the datasets using population demographics: are the corpora substantially different from corpora which are compiled without these requirements?

We break this into two questions. First, are the corpora themselves significantly different from the CoNLL baselines? In other words, if we view each set of corpora as producing ranks of unigram frequencies, do the texts themselves represent different classes? Second, are the products of the corpora different when collapsed into low-dimensional representations? In other words, if we train word embeddings from each corpora, are word similarity measures consistent across both datasets?

[9] https://www.earthlings.io/geowac_countries.html

| ISO3<br>Code | Language<br>Name | CoNLL<br>(Corpus Size) | GeoWAC<br>(Corpus Size) | Shared<br>Unigrams | Spearman's<br>$\rho$ |
|---|---|---|---|---|---|
| ara | Arabic | 1,083 mil | 592 mil | 562,897 | 0.79 |
| bul | Bulgarian | 277 mil | 826 mil | 327,559 | 0.76 |
| cat | Catalan | 701 mil | 106 mil | 114,937 | 0.73 |
| ces | Czech | 1,449 mil | 1,091 mil | 855,925 | 0.82 |
| dan | Danish | 1,203 mil | 1,031 mil | 606,384 | 0.74 |
| deu | German | 4,725 mil | 1,093 mil | 1,051,195 | 0.75 |
| ell | Greek | 542 mil | 939 mil | 522,709 | 0.71 |
| eng | English (Inner) | 7,272 mil | 2,025 mil | 428,515 | 0.71 |
| eng | English (Outer) | 7,272 mil | 1,896 mil | 463,212 | 0.61 |
| eng | English (Expanding) | 7,272 mil | 2,263 mil | 662,101 | 0.67 |
| est | Estonian | 234 mil | 472 mil | 503,558 | 0.72 |
| fas | Farsi | 927 mil | 1,126 mil | 345,194 | 0.76 |
| fin | Finnish | 753 mil | 1,062 mil | 1,177,237 | 0.71 |
| fra | French | 4,460 mil | 2,140 mil | 612,753 | 0.75 |
| gle | Irish | 19 mil | 21 mil | 37,682 | 0.66 |
| hbs | Serbo-Croatian | 460 mil | 875 mil | 510,963 | 0.78 |
| hin | Hindi | 132 mil | 407 mil | 26,936 | 0.62 |
| hun | Hungarian | 1,213 mil | 1,095 mil | 1,228,055 | 0.80 |
| ind | Indonesian | 4,491 mil | 429 mil | 244,744 | 0.73 |
| ita | Italian | 4,432 mil | 1,098 mil | 530,837 | 0.82 |
| jpn | Japanese | 3,850 mil | 1,099 mil | 550,749 | 0.58 |
| kor | Korean | 404 mil | 293 mil | 547,057 | 0.57 |
| nld | Dutch | 2,314 mil | 1,122 mil | 649,285 | 0.75 |
| nor | Norwegian | 997 mil | 990 mil | 525,370 | 0.69 |
| pol | Polish | 3,854 mil | 1,078 mil | 912,666 | 0.82 |
| por | Portuguese | 5,111 mil | 1,754 mil | 533,137 | 0.74 |
| ron | Romanian | 2,256 mil | 1,050 mil | 521,309 | 0.83 |
| rus | Russian | 2,287 mil | 2,077 mil | 1,109,541 | 0.76 |
| slk | Slovak | 551 mil | 1,078 mil | 662,990 | 0.79 |
| slv | Slovenian | 374 mil | 473 mil | 369,726 | 0.79 |
| spa | Spanish | 4,762 mil | 2,020 mil | 625,192 | 0.80 |
| swe | Swedish | 2,184 mil | 1,086 mil | 774,103 | 0.77 |
| tur | Turkish | 2,729 mil | 142 mil | 346,834 | 0.77 |
| ukr | Ukrainian | 355 mil | 501 mil | 433,176 | 0.72 |
| vie | Vietnamese | 3,835 mil | 1,088 mil | 138,960 | 0.65 |
| zho | Chinese | 1,304 mil | 2,028 mil | 193,347 | 0.31 |

Table 8: Similarities between CoNLL and GeoWAC corpora using Spearman's rank correlation coefficient.

### 6.1. Comparison with CoNLL Baseline Corpora

How similar are geographically balanced corpora with the unbalanced CoNLL baseline web corpora? To answer this question, we take the 36 languages (out of 48 total) which are represented in both datasets, as shown in Table 8. First, we show the size of each corpus by number of words. In most cases the CoNLL corpora have no ceiling while the GeoWAC corpora are capped at 1 or 2 billion words.

Work on corpus similarity (Kilgarriff, 2001; Fothergill et al., 2016) has shown that the $\chi^2$ measure, based on ranks of unigram frequencies, can accurately predict the overall similarity between two corpora in tightly-controlled experimental settings. This particular measure out-performs model-based approaches but is sensitive to mismatches in corpus size. Because the two datasets are not the same size per language, we instead use Spearman's rank correlation, which has been shown to work best in such cases (Kilgarriff, 2001; Fothergill et al., 2016).

The first step is to align the unigrams from both corpora: here, a word must occur at least ten times in both datasets before it is included in this measure. While earlier formulations limited the number of words, it has been shown that including more vocabulary items increases the accuracy of the measure (Fothergill et al., 2016). In Table 8 the number of shared vocabulary items for each language is also shown, ranging from 26,000 (Hindi) to 1,228,000 (Hungarian).

The higher the Spearman correlation, the more similar the GeoWAC and CoNLL corpora are for a given language. In cases where the similarity is relatively high (for example, Italian with 0.82) the process of balancing the corpora has had less impact on the data itself than in cases where the similarity is relatively low (for example, Vietnamese with 0.65). Even in these cases, however, there is substantial difference between the datasets, showing that population-based sampling produces corpora that differ significantly from unbalanced alternatives.

In addition to being balanced by population, the English corpora is divided into inner-circle, outer-circle, and expanding-circle sub-sets. Previous work has shown that baseline datasets better represent inner-circle varieties (Dunn, 2019a). We see a similar pattern here: the inner-circle sub-set is most similar to the CoNLL baseline (0.71); the outer-circle sub-set is the least similar (0.61). This provides yet another confirmation of the strong inner-circle bias of existing gigaword corpora.

The next question is whether the divergences between the two datasets is related to the relative geographic distribution of each language. First, the languages with less than 70% of the corpus coming from a single country (i.e., those with the widest geographic distribution) have the lowest average similarity with the baseline: 0.70. Second, those languages with a narrow distribution (between 70% and 90% coming from a single country) have the highest similarity: 0.75. The point is that all corpora are substantially different as a result of population-based sampling, but languages that are used across many countries are likely to be more different.

### 6.2. Comparison with CoNLL Embeddings

For a further evaluation of the sampling algorithm, we compare word embeddings that are trained on both sets of corpora. This provides more insight into the context of word-usage in each corpus family. For a baseline we use the CoNLL 2017 shared task word embeddings of dimension 100 that were created using the word2vec model (Mikolov et al., 2013; Ginter et al., 2017). We generate similar 100-dimensional word2vec embeddings for each language using the balanced GeoWAC corpora.

In order to compare the word embeddings we adopt a technique previously developed to compare the stability of word embeddings across multiple trainings (Hellrich et al., 2019). For each language, a set of 1,000 anchor terms, $A$, are selected from the CoNLL corpus based on the most common occurrences in that corpus. Then the $n$ most similar words ($msw$) are calculated using cosine distance for both the CoNLL and GeoWAC word embeddings. The Jaccard coefficient $j@n$ is then calculated for the two sets of most similar words:

$$j@n = \frac{1}{|A|} \sum_{a \in A} \frac{msw_{CoNLL}(a,n) \cap msw_{GeoWAC}(a,n)}{msw_{CoNLL}(a,n) \cup msw_{GeoWAC}(a,n)}$$

The results with $n = 10$ are shown in Table 9. These results show that the Jaccard coefficients are quite low. Although there is an inherent instability in word2vec-like word embeddings, as shown in (Hellrich et al., 2019), these Jaccard coefficients are lower than would be expected from normal instability. This indicates that the geographically-balanced corpora represent a substantially different set of examples of word usage than the CoNLL corpora.

At the same time, it is notable that the Jaccard coefficients across most languages fall inside a rather narrow band between 0.16 and 0.22. This means that for most languages, given the top-10 most similar words for the most frequent 1,000 words, the two data sets share an average of two words. This shows that population-based sampling is likely to have significant down-stream impacts on modeling tasks. Further, there is a significant correlation of 0.62 between the corpus similarity measures and the word embedding similarity measures across languages, showing that these two sets of evaluations are representing an underlying divergence between the balanced and unbalanced corpora.

| ISO3 Code | Language Name | Jaccard coefficient |
|---|---|---|
| ara | Arabic | 0.1318 |
| bul | Bulgarian | 0.1713 |
| cat | Catalan | 0.1525 |
| ces | Czech | 0.2005 |
| dan | Danish | 0.1596 |
| deu | German | 0.2140 |
| ell | Greek | 0.1131 |
| eng | English (Inner) | 0.1873 |
| eng | English (Outer) | 0.2139 |
| eng | English (Expanding) | 0.2169 |
| est | Estonian | 0.1883 |
| fas | Farsi | 0.2092 |
| fin | Finnish | 0.2020 |
| fra | French | 0.1670 |
| gle | Irish | 0.0880 |
| hin | Hindi | 0.0285 |
| hun | Hungarian | 0.2257 |
| ind | Indonesian | 0.2252 |
| ita | Italian | 0.1726 |
| jpn | Japanese | 0.0962 |
| kaz | Kazakh | 0.1442 |
| kor | Korean | 0.0717 |
| lav | Latvian | 0.1687 |
| nld | Dutch | 0.2200 |
| pol | Polish | 0.1835 |
| por | Portuguese | 0.1876 |
| ron | Romanian | 0.1902 |
| rus | Russian | 0.1508 |
| slk | Slovak | 0.1684 |
| slv | Slovenian | 0.1602 |
| spa | Spanish | 0.1908 |
| swe | Swedish | 0.1733 |
| tur | Turkish | 0.1674 |
| ukr | Ukranian | 0.1714 |
| vie | Vietnamese | 0.0254 |
| zho | Chinese | 0.0447 |

Table 9: Similarity of CoNLL and GeoWAC embeddings.

## 7. Conclusion

This paper has described and evaluated the GeoWAC corpus family which is designed to remove geographic bias from gigaword corpora in order to better represent population demographics in down-stream NLP tasks. The paper has shown (Section 5) that population-based sampling provides corpora that are more geographically distributed than unbalanced alternatives. Further, the paper has also shown (Section 6) that the baseline corpora over-represent inner-circle varieties of English and that, across all languages, there is a substantial divergence between balanced and unbalanced corpora in terms of both corpus similarity measures and word embedding comparison measures.

## 8. Bibliographical References

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Benko, V. (2014). Aranea Yet Another Family of (Comparable) Web Corpora. In *Proceedings of 17th International Conference Text, Speech and Dialogue.*, pages 257–264. Springer.

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

Davies, M. and Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide*, 36(1):1–28.

Dunn, J. and Adams, B. (2019). Mapping Languages and Demographics with Georeferenced Corpora. In *Proceedings of Geocomputation 2019*, pages 1–16.

Dunn, J. (2019a). Frequency vs. Association for Constraint Selection in Usage-Based Construction Grammar. In *Proceedings of the NAACL 2019 Workshop on Cognitive Modeling and Computational Linguistics*, pages 117–128. Association for Computational Linguistics.

Dunn, J. (2019b). Global Syntactic Variation in Seven Languages: Towards a Computational Dialectology. *Frontiers in Artificial Intelligence*.

Dunn, J. (2019c). 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.

Dunn, J. (2020). Mapping Languages: The Corpus of Global Language Use. *Language Resources and Evaluation*.

Erich, S. J. (2010). *Language and Space, Volume 1: Theories and Methods*. De Gruyter Mouton, Berlin, Boston.

Fothergill, R., Cook, P., and Baldwin, T. (2016). Evaluating a Topic Modelling Approach to Measuring Corpus Similarity. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 273–279. European Language Resources Association.

Ginter, F., Hajič, J., and Luotolahti, J. (2017). *CoNLL 2017 Shared Task - Automatically Annotated Raw Texts and Word Embeddings*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (FAL), Faculty of Mathematics and Physics, Charles University.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth Conference on Language Resources and Evaluation*, pages 759–765. European Language Resources Association.

Grieve, J., Montgomery, C., Nini, A., Murakami, A., and Guo, D. (2019). Mapping Lexical Dialect Variation in British English Using Twitter. *Frontiers in Artificial Intelligence*.

Hellrich, J., Kampe, B., and Hahn, U. (2019). The influence of down-sampling strategies on SVD word embedding stability. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 18–26.

Kachru, B. e. (1982). *The Other tongue: English across cultures.* University of Illinois Press, Urbana-Champaign.

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

Majliš, M. and Žabokrtský, Z. (2012). Language Richness of the Web. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 2927–2934. European Language Resources Association.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *ICLR Workshop*.

Szmrecsanyi, B., Grafmiller, J., and Rosseel, L. (2019). Variation-Based Distance and Similarity Modeling: A Case Study in World Englishes. *Frontiers in Artificial Intelligence*.

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

Upton, C. (2013). Blurred boundaries: the dialect word from the BBC. In C Upton et al., editors, *Analysing 21st Century British English: Conceptual and Methodological Aspects of the "Voices" Project*, pages 180–197. Routledge, London.