Mapping Wikipedia's Geolinguistic Contours

Martin Dittus and Mark Graham

Abstract

Wikipedia is one of the predominant ways in which internet users obtain knowledge about the world. It is also one of the most important mirrors, or augmentations, of the world: it contains representations of all manner of places. However, Wikipedia's knowledge of the world is characterised by a linguistic inequality. Although it is written in a growing number of languages, some languages are overrepresented and contribute significantly more to Wikipedia's body of knowledge than others. This deeply affects how the world is represented on Wikipedia, and by whom: it has been shown that for many countries in the Global South, there are more articles written in English than in their respective native languages. As a result, a significant number of people are being excluded from the collective process of knowledge production, solely on the basis of their native language. Who writes these representations of local places, and for which audiences? We present early findings from the first study of Wikipedia's geolinguistic contours. We investigate to what extent local languages are involved in the process of creating local representations. In a large-scale quantitative analysis across the almost 300 language versions of Wikipedia, we identify regions of the world where local languages such as Armenian, Catalan or Malay are dominant sources of representation for local places, and we contrast these findings with instances where representations are significantly shaped by foreign languages. Where do, and do not, we see significant amounts of local content available in local languages? Where are the most detailed local representations largely written in foreign languages, intended for foreign audiences? And what factors can explain this?

Introduction

Our world is ever more augmented by digital information. Streets, buildings, regions, monuments, and events exist not just as their material articulations, but also as their digital representations (Graham, Zook, & Boulton 2013). This paper focuses on Wikipedia, one of the predominant platforms for such digital representations of the world (Graham, Straumann, & Hogan 2015). Through a set of

DOI 10.14361/dcs-2019-0109

linguistic mappings, we explore how digital information on the platform may be selectively available to particular communities but not others, purely based on the languages spoken by these communities. If we want to better understand the extent to which digital augmentations are relevant, available and accessible to local populations, we need to ask questions about the geolinguistic contours of the web: the extent to which digital content about places in the world is captured in particular languages but not others.

Although the knowledge production process by Wikipedia's global community is in principle open to anyone, prior work has shown that Wikipedia's global participation geography is highly unequally distributed. Many regions of the world are largely excluded from the process, in part because internet connectivity is a necessary prerequisite, and billions of people remain disconnected (Graham et al. 2014; Graham, Straumann, & Hogan 2015). Although Wikipedia strives to collect all of human knowledge, in practice only a subset of the world's population is participating in the process, which raises the question to what extent Wikipedia's geographic knowledge of the world is often written by outsiders.

Informed by such concerns, the Wikipedia community has introduced the concept of *knowledge equity*,[1] the idea that everybody should have the opportunity to participate in the creation of knowledge, and that communities around the world should have the capacity to make decisions about how they are being digitally represented. The concept of knowledge equity asks that local representations are produced in a form that is accessible to local communities, most importantly by using languages that are read and written by the local population. We ask, how often is this currently taking place?

Recent research findings support the expectation that the involvement of local communities can be instrumental when trying to capture local concerns. In a study of Wikipedia's global language culture, it was found that much of the content in Wikipedia's language editions is dedicated to represent the corresponding cultural context unique to the respective language, content that was often not encountered in other language editions (Miquel-Ribé & Laniado 2018).

Past work has shown that the presence of an international community of people literate in a given language can be a further driver for certain knowledge production practices on Wikipedia. At a basic level, shared language culture among Wikipedians is associated with a shared interest in particular thematic topics (Karimi et al. 2015). More importantly, it was found that shared language may be an even stronger driver behind shared thematic interests than the geographic distance between participating communities (Samoilenko et al. 2016). In other words, it has been shown that a shared language culture can bring together geographically disparate contributor communities.

Informed by these observations, in this paper we relate local-language cultures to broader global language communities. We ask to what extent the size of a global

1 https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2017/Direction.

language community may inform whether the dominant language for Wikipedia representations is a local or a foreign one. Specifically, we focus on the question of local-language representation on Wikipedia, asking how often Wikipedia articles about local places are written in local languages. We present initial and ongoing empirical work that is designed to offer a broad overview of the geolinguistic patterns of the Internet. Does Wikipedia's highly unequal geography of participation also mean that its representations of certain parts of the world are inherently inequitable, as local communities are not contributing local-language representations of their own places? To the extent that this is the case, can the presence of contributors from other countries who are literate in the local language help foster the development of richer local-language representations?

Our paper addresses the following research questions: First, where do, and don't, we see significant amounts of local content available in local languages? Second, where are local representations largely created for foreign audiences? Finally, what factors can help explain these relationships?

Sources of Empirical Data

We study Wikipedia's language geography through a lens of global language cultures, with countries as unit of analysis. We first identify a set of local languages for every country, these are languages that either have official status at national or regional level or are in use by at least 30 percent of the population. We then link these to a set of wiki languages, these are the languages in which there is geographic content available on Wikipedia. The resulting data covers more than 70 global languages. The details are presented in the following paragraphs.

One methodological challenge is the pluriversality of language culture: in any given place, languages always coexist with other languages, which leads to complex intersections of use and of capacity among those literate in certain languages. Countries can have multiple official languages, people may read and write multiple languages and places can be written about in multiple languages. To address this, in our analyses, we rely on the concept of a dominant language: the local language with the highest population literacy rate (the dominant local language), or with the largest number of articles (the dominant wiki language). We identify such dominant languages for every country and use them as the basis of our analyses.

As a consequence of this methodological choice, our analysis places an emphasis on the comparison of larger language groups. While this is appropriate for an analysis at global scale, research aiming to study regional effects would require an approach that also incorporates information about less widely spoken languages.

Local-Language Geography

We rely on a data set of local languages to assess whether Wikipedia content is accessible to a local population. We begin by asking if articles are written in languages that are understood by a large subset of the local population. To this purpose, we use a data set of territory-language-information published by the Unicode Common Locale Data Repository[2] (CLDR). This CLDR survey captures the languages that are commonly spoken, read, and written in every country of the world. The project's aim is to "provide approximate figures for the literate, functional population for each language in each territory: that is, the population that is able to read and write each language, and is comfortable enough to use it with computers."[3]

We use the CLDR survey as a basis to identify local languages for every country. The CLDR identifies those languages that have official status at national or regional level in a country, which is an important starting point for such a list. A limitation of this survey is that it does not always correctly identify all official languages, particularly in countries where historically, colonial languages have displaced local languages. This is the case for Kuanyama in Namibia and the Mossi language in Burkina Faso, among others. To address this limitation, we regard as local languages any languages that either have official status at national or regional level, or that are in use by at least 30 percent of the population, and use this as the basis for our analysis. The map in Figure 1 shows the resulting number of local languages per country.[4]

For every country, we then identify a dominant local language: the local language of a country that the largest share of the population is literate in. To this purpose, we determine the population percentages for every local language according to the CLDR survey and identify the most widely used local language. The population share of the dominant local languages is shown on the map in Figure 2. A large majority of the respective country's population tend to be literate in this dominant language. More than 50 percent of the population can speak, read and write this language in 151 countries (90 percent of all countries included in the study.)

According to the CLDR survey, in some African and South Asian countries, the majority of the population are not literate in the dominant local language. In the three countries with the lowest literacy rates, only between 26 percent and 30 percent of the population are literate in the dominant language, all of them on

2 http://cldr.unicode.org (Territory-Language Information, version 34beta).

3 https://unicode.org/cldr/charts/latest/supplemental/territory_language_information.html.

4 Certain countries are excluded from the study because necessary empirical data were not available. This is described in more detail in the section "Data Used for Analysis" on page 153.

the African continent: Chad, Guinea and Mozambique. In these three cases, the dominant local language is a former colonial language: French or Portuguese. In Mozambique for example, where only 27 percent of the population are literate in the sole official language Portuguese, an even smaller share is literate in other local languages such as Makhuwa, Ndau and Tsonga. In other words, these low population rates for the dominant local language are the result of a highly multilingual language culture, as well as a colonial history where an outside language was introduced for administrative purposes, and where only an elite was literate in it.

On the other hand, in some countries, we encounter a highly multilingual language culture: 48 countries (28 percent) have three or more official languages and 11 (7 percent) have five or more. In such multilingual countries, Wikipedia contributor effort is split or even multiplied across multiple language communities, resulting in a scenario where the same local knowledge needs to be represented multiple times, once per local language. Does this put these populations at a disadvantage when it comes to their capacity to produce local-language representations on Wikipedia?

Fig. 1: Number of local languages per country, including official languages at national and regional levels, as well as those in use by at least 30 percent of the population, for countries included in the study

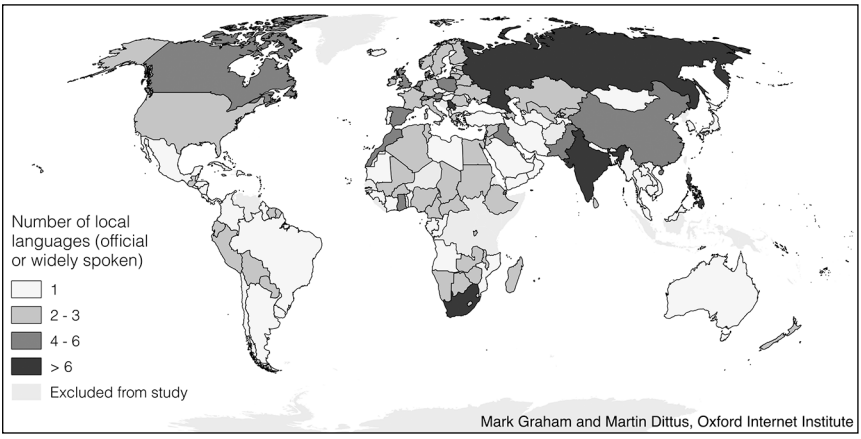
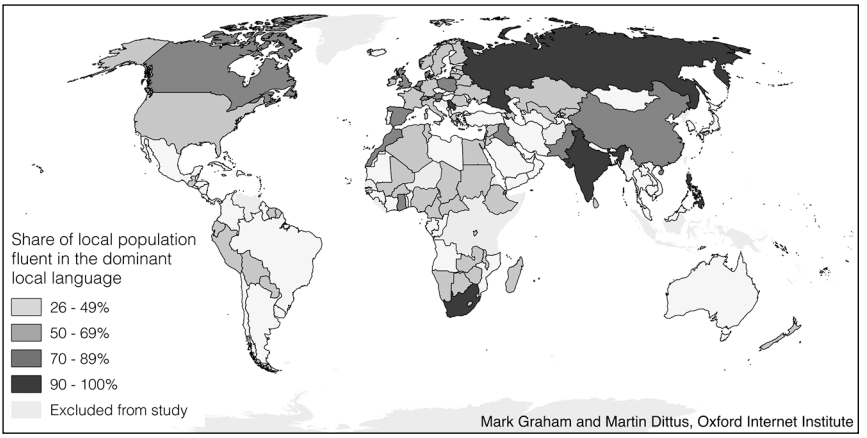


Fig. 2: Share of local population literate in the dominant local language. Only for countries included in the study



Wiki-Language Geography

We identify all articles referring to places (e. g. articles about buildings, battles, city districts, monuments, or events) across Wikipedia's almost 300 language editions. We establish whether an article is about a place based on geotags, a widely used annotation scheme in Wikipedia articles that provides geographic coordinates. These geotags are readily identified in Wikipedia's publicly available contribution history and can then each be spatially joined to every country.

To this purpose, we retrieve a list of geotagged articles for all 300 Wikipedia-language editions, as recorded in the contribution history.[5] We identify the primary geotag per article (some articles can have multiple geotags). If none is marked as such, we select the first geotag added to the page. We exclude articles with more than four geotags and consider these lists rather than articles about specific places. Each article is then mapped to a country based on the location of its primary geotag, using a Natural Earth boundary data set.[6] We find that 141 of 297 wikis have at least one geotag and can be considered for inclusion in our study.

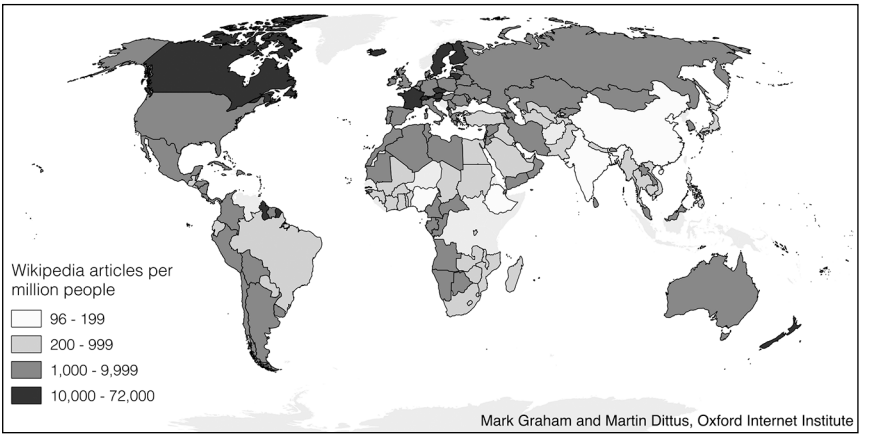
The map in Figure 3 shows the overall number of Wikipedia articles of local places, aggregated per country, across all of Wikipedia's languages (normalised by population). The map illustrates the significant global inequality of these local representations: in relative terms, countries in the Global North tend to have an order of magnitude more articles written about them than the Global South. (Note that the map uses an exponential scale.)

5 https://dumps.wikimedia.org (downloaded in February 2018).

6 http://www.naturalearthdata.com (Admin-0 country borders at 1:10 M scale, version 4.0).

For every country we identify the dominant wiki language; this is the language with the most Wikipedia articles for the given country, which may not always be a local language. Special care is taken to specifically identify content that is produced by humans. In certain Wikipedia-language editions, automated scripts have contributed a large amount of geographically referenced content that amounts to little more than a placeholder page. This is particularly an issue for wikis with small language communities, as it creates an appearance of significant knowledge production activity without actually involving human effort. To address this, we filter out wikis which were largely created by automated bots. We identify such wikis based on the Wikipedia concept of language depth,[7] an empirical measure that captures the degree to which a wiki has been produced by a large community in an ongoing process, rather than a small number of contributors who created articles that were then never updated. We remove all wikis with a language depth below 10, a low threshold which filters out wiki languages that are heavily shaped by automated scripts, such as Cebuano and Waray.

Fig. 3: The number of geotagged articles about local places in any language supported by Wikipedia, normalised by population (articles per million people). Only for countries included in the study



Data Used for Analysis

We join these two data sets of local-language geography and Wikipedia-language geography, matching every Wikipedia-language edition to its respective local language, based on the set of local languages provided by the CLDR survey. We exclude some languages and countries from our analysis of the Wikipedia-lan-

7 https://en.wikipedia.org/wiki/List_of_Wikipedias#Detailed_list (snapshot taken in May 2018).

guage geography to address a shortcoming of the available empirical data. Some wikis follow geotagging conventions that are not captured in the official geotag data dumps. We exclude such wikis from our study, as it is not possible to determine their language geography with our methods. We further exclude all countries from our analyses where such a language is the dominant local language. This excludes some Eastern European languages such as Bulgaria, Croatia and Slovakia, but also several African countries such as Niger, Senegal and Somalia and smaller nations such as the Faroe Islands. Finally, we exclude countries for which certain explanatory measures are not available: population counts, broadband connectivity and education rates.

In total, 169 countries are included in the study. 73 languages are dominant local languages in at least one country; the most common dominant local languages are shown in Table 1a. In turn, 35 Wikipedia-language editions are the dominant wiki language in at least one country; the most common ones are shown in Table 1b. All 73 dominant local languages have a Wikipedia-language edition; however these local wikis are not necessarily the most prolific providers of local content: in 102 countries, the dominant wiki language is *not* the dominant local language (60 percent of all countries included in the study).

Table 1a: The 20 most common dominant local languages, by number of countries

Dominant Local Language	**# Countries**
English	34
Arabic	18
Spanish	18
French	13
Portuguese	7
German	4
Dutch	3
Traditional Chinese	2
Italian	2
Malay	2
Romanian	2
Greek	2
Russian	2
Czech	1
Danish	1
Japanese	1

Table 1b: The 20 most common dominant wiki languages, by number of countries

Dominant Wiki Language	**# Countries**
English	98
French	9
German	8
Spanish	7
Catalan	4
Russian	4
Italian	3
Serbian	3
Dutch	2
Greek	2
Arabic	2
Serbo-Croatian	2
Swedish	2
Romanian	2
Slovenian	1
Bavarian	1

Dominant Local Language	# Countries
Afrikaans	1
Icelandic	1
Armenian	1
Hungarian	1
(Others)	53

Dominant Wiki Language	# Countries
Belarusian	1
Ukrainian	1
Cebuano	1
Kurdish	1
(Others)	15

Local-Language Representations

For each country, we determine how many Wikipedia articles about local places are available in the dominant local language. We normalise these measures by population, which allows us to compare the national volume of local-language content across countries, independent of their relative sizes.

What factors can explain the distribution of local content?

We seek to better understand where we see significant amounts of local content available in local languages. Which parts of the world are augmented by dense clouds of local content? Which parts of the world are layered with almost no local content? And what factors may explain this distribution of local content?

The global distribution of content volumes is shown in the map in Figure 4. As before, the map makes apparent the high density of representations in countries in the Global North, which tend to have an order of magnitude more articles written about them than the Global South. (Note that the map is using an exponential scale.) The highest-ranking countries are shown in Table 2.

Fig. 4: Number of articles in the dominant local language, normalised by population (articles per million people). Only for countries included in the study

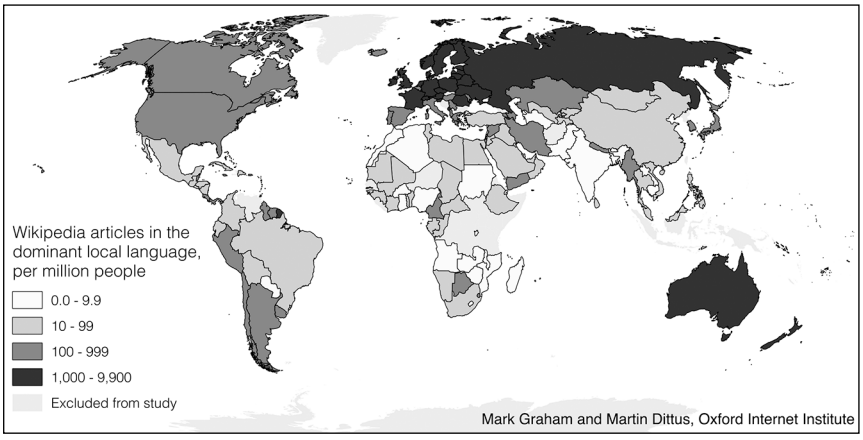


Table 2: Countries with the largest number of articles in the dominant local language per capita (in articles per million people)

Country	Continent	Articles per million people
Lithuania	Europe	9,836.2
Sweden	Europe	9,499.8
Estonia	Europe	7,115.3
Norway	Europe	3,918.8
Latvia	Europe	3,785.2
Slovenia	Europe	3,620.2
Czechia	Europe	3,388.7
Andorra	Europe	3,183.3
Poland	Europe	3,134.9
Belarus	Europe	2,901.3
San Marino	Europe	2,425.1
Austria	Europe	2,337.3
Montenegro	Europe	2,070.8
Finland	Europe	2,045.4
Germany	Europe	1,985.6
Netherlands	Europe	1,901.1
Denmark	Europe	1,819.9
Liechtenstein	Europe	1,661.3
Switzerland	Europe	1,518.3
France	Europe	1,479.8

For an explanatory analysis, we relate local content volumes to potential explanatory factors in a regression model. We seek to explain the number of local-language articles of local places, normalised by population. As explanatory features, we employ national indicator measures which may explain the presence and absence of local contribution capacity.

The full set of explanatory valuables is shown in Table 3. Since the independent variable is expressed relative to the size of the respective country, we also rely on relative measures for all explanatory variables. They span a wide range of concerns: the degree of local-language diversity, population density, GDP as a general indicator of economic development, broadband connectivity and cost, educational attainment and adult literacy. Data for educational attainment and adult literacy is not available for many high-income countries such as the United States and the United Kingdom, since these surveys are largely focused on capturing the status of low- and middle-income countries. To address these gaps, we assign the

global average rates to all countries with missing measurements. This ensures that our model covers a wider range of countries; however, as a consequence, the model now underestimates global inequalities in education and literacy rates, which reduces the likelihood of finding an effect relating to these measures.

We use this model in an ordinary least-squares linear regression, and the regression coefficients are reported in Table 4. We find that broadband per capita is the only significant measure in this model. Maybe surprisingly, measures of local-language diversity are not statistically significant (the number of local languages and the population share literate in the dominant local language). Similarly, educational attainment and literacy rates are not statistically significant.

This result confirms prior research in the literature which found that connectivity is a necessary prerequisite for Wikipedia participation (Graham et al. 2014; Graham, Straumann, & Hogan 2015). The map in Figure 5 shows the global distribution of broadband connectivity, indicating that the Global North is significantly more well-connected; Africa and South Asia is particularly low in connectivity rates.

Table 3: Features for a regression model to explain the relative amount of local-language representations of local places

Variable	Description	Source
local_languages	The number of local languages that are designated official languages at national or regional level, or which are read and written by at least 30 percent of the population.	CLDR
dll_pop_share	The share of the local population literate in the dominant local language.	CLDR
pop_density	People per square km of land area.	World Bank*
gdp_pcap	GDP per capita.	World Bank
broadband_p100	Number of broadband connections per 100 people.	World Bank
broadband_cost_GNI	Cost of broadband relative to gross national income.	ITU**
primary_completion	Enrolment rate in the last year of primary school education.	World Bank
adult_literacy	Adult literacy rate.	World Bank

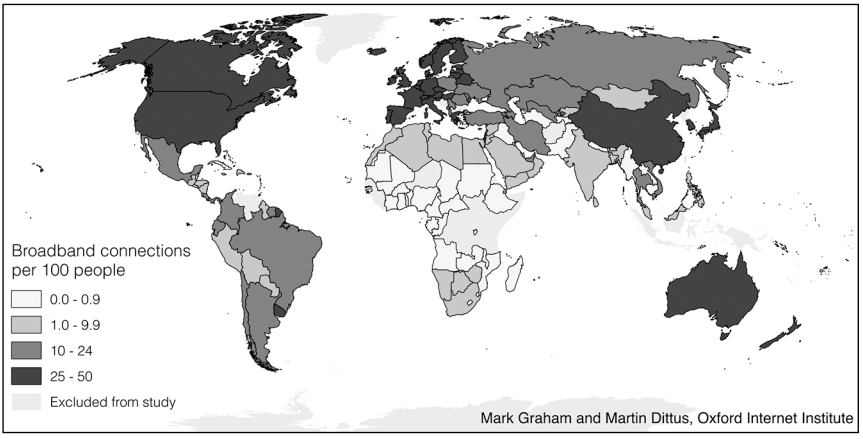
* https://data.worldbank.org (all data from 2017, or most recent year available).

** https://www.itu.int/ (ICT prices for 2017).

Table 4: Linear regression model to explain the number of articles in the dominant local language, in articles per millions of people. Model fit: adjusted R^2=0.250

Variable	Coefficient	Significant?
local_languages	(11.43)	–
dll_pop_share	(-7.83)	–
pop_density	(-0.06)	–
gdp_pcap	(0.00)	–
broadband_p100	**57.37**	$p < 0.001$
broadband_cost_GNI	(0.19)	–
primary_completion	(7.87)	–
adult_literacy	(-1.03)	–
(Intercept)	(-67.91)	–

Fig. 5: broadband connections per 100 people. Only for countries included in the study



Local-Language Equity

We define local-language equity as the degree to which content about local places is accessible to people literate in local languages. For the present study, we operationalise this concept by means of a simple comparison: is the dominant wiki of a country written in a local language? In other words, are the most detailed representations of this country written in a local language understood by the population, or in a foreign one?

Dominance of Foreign-Language Content

We seek to better understand where local representations are largely created for foreign audiences. For every country, we determine whether the dominant wiki language is a local language, as opposed to a foreign one. The map in Figure 6 shows the results of this comparison. The map reveals a striking pattern: most content for countries in Africa, Central and South America and many South Asian countries is written in foreign languages. Many people in the global south thus lack an ability to access the bulk of local information on Wikipedia about places that they inhabit.

Even in cases where the dominant wiki language is a local one, it is not always the dominant local language. In India, the dominant local language is Hindi, which according to the CLDR has a population literacy rate of about 40 percent, and yet three times more articles about India are written in English which only has a literacy rate of 20 percent (11,000 compared to 39,000 articles). In Spain, only 20 percent of the population are literate in Catalan, but the Catalan wiki has almost twice the number of geotagged articles about Spain than the Castilian wiki (60,000 compared to 36,000 articles). Finally, in Madagascar, most of the population is literate in the national language Malagasy (90 percent), yet less than a dozen Wikipedia articles about the island nation are written in this language. The dominant wiki language for Madagascar is English (1,500 articles), which after a 2010 referendum is no longer considered an official language.

When the most detailed representations of a place are being written in a foreign language, this could be considered a form of displacement. Table 5a and 5b list the most common such displacements: the local languages that are most commonly being displaced by a foreign language and the foreign languages that are most commonly being displaced by a local language. The map in Figure 7 indicates the global locations of such displacement for the four most frequently displaced languages: Arabic, French, Spanish and Portuguese. The maps reveal that in many of these cases, the dominant local language is a former colonial language. The map in Figure 8 indicates the global locations where English is the dominant wiki language, including countries where displacement took place. Taken together, the maps suggest that in many places, former colonial languages are being displaced by English which offers more detailed representations of local places. English, in other words, is beginning to dominate how the world is represented on Wikipedia.

These displacement patterns generally suggest a relationship to the relative sizes of local-language communities. Many of these displacements are in countries where the dominant language only accounts for a relatively small share of the population (refer to the map in Figure 2). In such cases, it may become possible for a dominant global language with a larger language community to dominate local representations. Prominent counterexamples to this pattern are Canada and Switzerland, two countries with strong language fragmentation, where the

dominant wiki language is a local one. It may matter that both their dominant local languages, English and (Swiss) German, respectively, also have significant communities of people outside these countries who are literate in these languages.

Fig. 6: Are the most detailed representations of a place written in a local language?

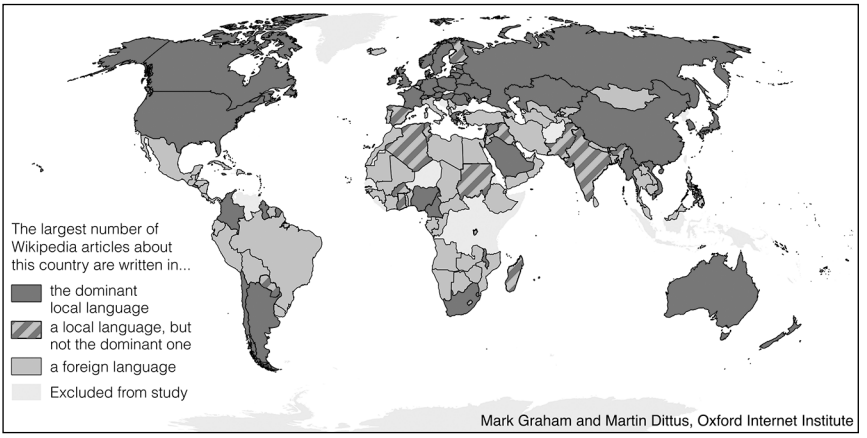


Table 5a: The dominant local languages that are most commonly displaced by a foreign language, by number of countries, limited to the top 20

Dominant Local Language	**# Countries**
Arabic	13
Spanish	12
French	11
Portuguese	7
English	5
Malay	2
Dutch	2
Traditional Chinese	1
Persian	1
Greek	1
Dzongkha	1
Italian	1
Divehi	1
German	1

Table 5b: The dominant wiki languages that are most commonly displacing a dominant local language, by number of countries

Dominant Wiki Language	**# Countries**
English	60
German	5
French	4
Serbian	3
Serbo-Croatian	2
Russian	2
Italian	2
Catalan	2
Hebrew	1
Spanish	1
Bavarian	1

Dominant Local Language	# Countries
Bangla	1
Bemba	1
Azerbaijani	1
Moroccan Arabic	1
Afrikaans	1
Icelandic	1
(Others)	18

Fig. 7: Locations of the four most commonly displaced languages

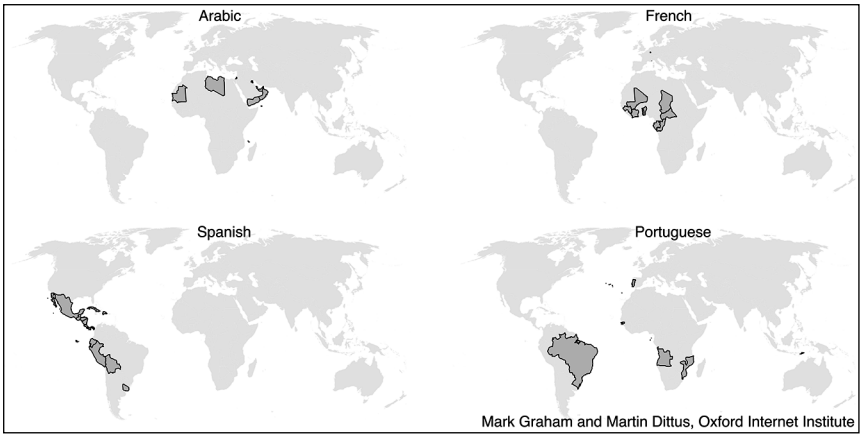
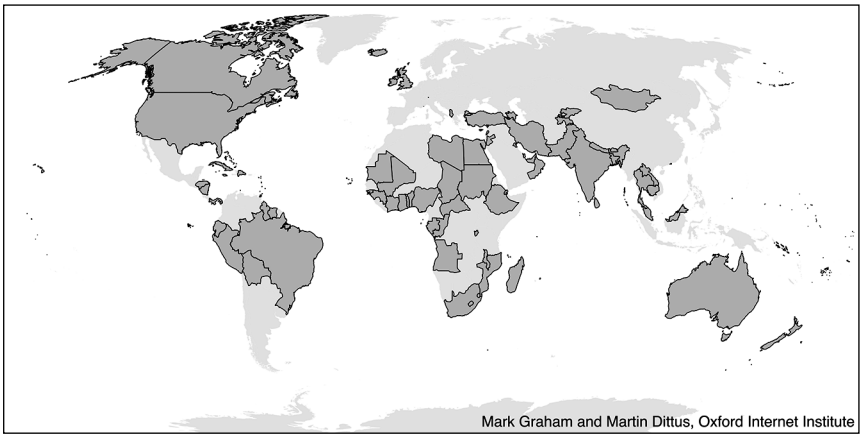


Fig. 8: Countries where English is the dominant wiki language, accounting for the largest number of articles about local places



What factors can explain foreign-language dominance?

We seek to better understand what factors may explain this particular distribution of foreign-language dominance. Why do foreign-language representations dominate in some countries and not others? Might it have to do with the relative popularity of a language – the local and global size of a language community? Is local-language fragmentation in certain countries a contributing factor? Or might it have to do with other factors relating to foreign interest, such as global tourism flows?

To answer these questions, we employ a logistic regression model to explain a binary outcome: is each country's dominant wiki written in a local language? For this analysis, we collect national factors relating to the respective sizes of local and global language communities and other factors of foreign interest in the respective country. These explanatory variables are listed in Table 6.

The resulting regression model is shown in Table 7. Only two variables are statistically significant: the number of people outside the country who are literate in the dominant local language and of people outside the country who are literate in the dominant wiki language. In other words, what matters is the size of the global language communities of the local and foreign languages. Local languages that have a large global language community are more likely also providing the most detailed Wikipedia representations of this country. Conversely, foreign languages that are more widespread have a competing tendency to dominate local representations. When the dominant local language does not have a large international community of people literate in it, a foreign-language community may dominate local representations.

Table 6: Features for a logistic regression model to explain whether each country's dominant wiki is written in a local language

Variable	Description	Source
local_languages	The number of local languages that are designated official languages at national or regional level, or which are read and written by at least 30 percent of the population.	CLDR
domestic_dll_pop	Domestic population literate in the dominant local language, in millions.	(derived)
remote_dll_pop	Global (non-domestic) population literate in the dominant local language, in millions.	(derived)
domestic_dwl_pop	Domestic population literate in the dominant wiki language, in millions.	(derived)

Variable	Description	Source
remote_dwl_pop	Global (non-domestic) population literate in the dominant wiki language, in millions.	(derived)
tourism_arrivals	Annual tourist arrivals, in millions.	World Bank

Table 7: Logistic regression model with confidence intervals to explain whether the dominant wiki is written in a local language. Model fit: pseudo-R^2=0.129

Variable	Coefficient	Significant?
local_languages	(0.000000)	–
domestic_dll_pop	(0.000001)	–
remote_dll_pop	**0.000014**	$p < 0.001$
domestic_dwl_pop	(0.000001)	–
remote_dwl_pop	**-0.000007**	$p < 0.02$
tourism_arrivals	(0.000000)	–
(Intercept)	(0.000000)	–

Conclusion

We find that local-language equity is largely dependent on factors that affect the capacity to produce representations in the dominant local language. Most importantly, the presence of broadband connectivity is associated with an increase in local-language content production. Additionally, the presence of an international community of people literate in a given languages can be an important driver for local-language representations. Where these factors are absent, local representations are more likely to be written in non-local languages.

In other words, the results reveal a simple majority rule of representations, and consequently inequity as a default outcome: in the absence of structured support,[8] dominant global languages provide the most detailed local representations of the world, possibly at the expense of local perspectives.

The overall dominance of English is striking. The map in Figure 8 illustrates the significant extent to which it is dominating representations of the world on Wikipedia. This indicates a tendency towards homogenisation: the presence of a single dominant language, possibly at the cost of local-language representations.

8 Structured support may include the presence of special-interest groups who train and coordinate Wikipedia contributors in the creation of articles covering underrepresented topics, as well as the provision of funds and other support to engage in such activities.

In many countries in the Global South, former colonial languages like Spanish, French and Portuguese are being displaced by English as a dominant representation language. Although the colonial language may often still be dominant in the country, it typically has a lower literacy rate than the primary languages of the Global North (refer to Figure 2). This is particularly the case in many African countries, where multiple traditional languages are in use by different segments of the population, resulting in a high degree of language fragmentation. As a result of all these factors, language communities in such places may find themselves to be a relative minority, compared to the dominating presence of the global English-language community.

References

Graham, M./Hogan B./Straumann R. K./Medhat A. (2014): "Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty." Annals of the Association of American Geographers 104 (4), pp. 746–764. https://doi.org/10.1080/00045608.2014.910087.

Graham, M./Straumann, R. K./Hogan B. (2015): "Digital Divisions of Labor and Informational Magnetism: Mapping Participation in Wikipedia." Annals of the Association of American Geographers 105(6), pp. 1158–1178. https://doi.org/10.1080/00045608.2015.1072791.

Graham, M./Zook M./Boulton A. (2013): "Augmented Reality in Urban Places: Contested Content and the Duplicity of Code." Transactions of the Institute of British Geographers 38(3), pp. 464–479.

Karimi, F./Bohlin L./Samoilenko A./Rosvall M./Lancichinetti A. (2015): "Mapping Bilateral Information Interests Using the Activity of Wikipedia Editors." Palgrave Communications 1(1). https://doi.org/10.1057/palcomms.2015.41.

Miquel-Ribé, M./Laniado D (2018): "Wikipedia Culture Gap: Quantifying Content Imbalances Across 40 Language Editions." Frontiers in Digital Humanities 5, p. 12.

Samoilenko, A./Karimi F./Edler D./Kunegis J./Strohmaier M. (2016): "Linguistic Neighbourhoods: Explaining Cultural Borders on Wikipedia through Multilingual Co-Editing Activity." EPJ Data Science 5(1). https://doi.org/10.1140/epjds/s13688-016-0070-8.