# Geographic and linguistic normalization: towards a better understanding of the geolinguistic dynamics of knowledge

Han-Teng Liao
Oxford Internet Institute
University of Oxford
Oxford, United Kingdom
hanteng@gmail.com

Thomas Petzold
HMKW – University of Applied Science for Media,
Communication & Management
Berlin, Germany
t.petzold@hmkw.de

## ABSTRACT

This paper proposes a method of geo-linguistic normalization to advance the existing comparative analysis of open collaborative communities, with multilingual Wikipedia projects as the example. Such normalization requires data regarding the potential users and/or resources of a geolinguistic unit.

## Categories and Subject Descriptors

[**Human-centered computing**]: Collaborative and social computing systems and tools–*Wikis, Empirical studies in collaborative and social computing*

## Keywords

Geolinguistic analysis, Geographic normalization, Linguistic normalization, Methodological nationalism

## 1. INTRODUCTION

This paper examines the current practices and research in measuring the geolinguistic differences of Wikipedia development and argues for better geographic and linguistic normalization measurements to improve the understanding and policy of global development strategies.

User or editor engagement has been one of the main foci of attention for the global Wikimedia movement. As previously reviewed by [10], online user-generated encyclopedias can be seen as collaborative ecosystems that seek to generate and maintain an ongoing, mutually-reinforcing cycle of increased participation, content, and readership. Herein lie the issues of global inequality in terms of cognitive surplus (from which potential participation can draw upon), available and reliable knowledge sources (from which content can grow), and digitally literate users (from which readership can be developed). Such global inequality has manifested itself through geographic and linguistic distribution of information and communication resources [4, 5, 11]. Facing similar challenges, the Wikimedia Foundation, the hosting organization for all Wikipedia projects, has targeted the "Global South" regions of Brazil, India, and the Arabic language countries for engagement [13].

To address such inequality, some researchers have been focusing on its geolinguistic aspect as observed on the web with the aim to show where the uneven geographies or dependent relationships lie [6, 8]. In practice, the Wikimedia foundation has generated statistics reports on editors, user traffic and content, including "per capita" measurements such as the number of articles per 1,000 speakers[9], the number of editors per million speakers[14], etc. However, these measurements are often crude for detailed analysis or based on datasets that are difficult to manage. Thus, there is a need for sound methodologies and sensible data processes so that maintaining easily accessible comparisons are possible for both researchers and practitioners.

Crowston, Julien and Ortega[3] have proposed a measurement to compare how efficient a language version turns potential users into actual contributors. To account for potential users, they collected data from publicly available data sources regarding the numbers of language speakers, Internet population and people with tertiary education. To measure actual contributors, they used the rough definitions provided by the Wikimedia Foundation: very active Wikipedians (those with more than 100 monthly revisions); active Wikipedians (between 5 and 100 monthly revisions), and the rest. They found "a strong (but not perfect) correlation" between the total number of Wikipedia contributors on one side, and the Internet population, and total tertiary-educated population on the other. A clear implication of such correlation is that it allows cross-lingual comparison: which and how much more successfully language X has turned potential users into actual readers than average (or language Y).

We aim to further such efforts by proposing the concept of geo-linguistic normalization, which breaks the unit of analysis for languages (e.g. Arabic) down to geo-linguistic units (e.g. Egyptian Arabic, Saudi Arabia Arabic, etc.).

## 2. DEFINITION

We derive our concept of geolinguistic normalization from two sources.

First, geographic normalization, or simply data normalization, allows data to be compared using a sensible common denominator, thereby producing measurements of intensity or density, such as population density [1, 2]. Such normalization is particularly useful in "factoring out the size" in order to facilitate comparisons across unequal areas or populations [2]. In other words, the geographic normalization process means dividing a certain numeric attribute (e.g. GDP) by another (e.g. population), so as to derive another numeric attribute (e.g. GDP per capita), thereby minimizing the differences caused by the size of a geographic unit. Thus, it is similar to Crowston et al's work[3] in "factoring out the size"; the difference is that geographic normalization concerns about geography units.

*OpenSym '14* , Aug 27-29 2014, Berlin, Germany
ACM 978-1-4503-3016-9/14/08.
http://dx.doi.org/10.1145/2641580.2641623

Second, geolinguistic units are often expressed by more specific "language tags" defined in HTML and XML. A language tag often starts with a language code followed by a country code. For instance, the language tag "fr-CA" represents the geolinguistic unit of French as used in Canada. Additional information about each geolinguistic unit is compiled by the Unicode's Common Locale Data Repository (CLDR) Project. For instance, in its publicly available Language-Territory information document, the CLDR version 25 has listed the language population of French in Canada as 7,605,004[12]. Such information provides finer details than just language population of French in the world, thus opening up more analytical opportunities.

For instance, instead of merely comparing how Arabic-speakers are getting involved in open collaborative projects such as Wikipedia compared to Spanish-speakers, the detailed geolinguistic units allow comparison between, say, Egyptian Arabic and Saudi Arabia Arabic speakers, or that of Spanish Spanish and Mexican Spanish speakers. Often codified and used by browsers and major websites to provide different interfaces and content[7], such geolinguistic units can thus be used by analysts or designers to better know and thus support their users.

We therefore simply define geolinguistic normalization as data normalization based on a certain numeric size feature in each geolinguistic unit, such as the number of speakers, Internet users, etc. The usefulness of such an approach is explored by the following normalization of Wikipedia traffic data.

## 3. METHODS

To illustrate the usefulness of the proposal, we scraped (using Scrapy) and constructed time-series data from individual Wikipedia traffic statistics pages for both editing and viewing data. At the moment of study, only proportional numbers are released by the Wikimedia Foundation when it breaks down the editing or viewing traffic of a given language version across different regions. Figure  and Figure  show how the viewing and editing traffic for the Arabic Wikipedia can be broken down into respective countries. Egypt and Saudi Arabia are expected to dominate as the major countries of the Arabic speaking world.

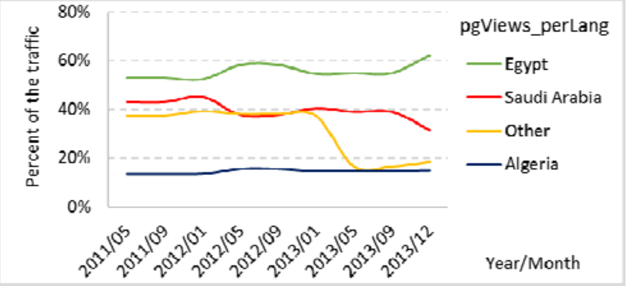

**Figure 1. Viewing traffic trend lines: Arabic Wikipedia**

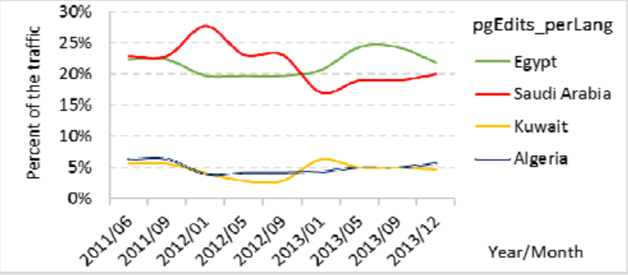

**Figure 2. Editing traffic trend lines: Arabic Wikipedia**

To normalize the above data points against geolinguistic units, or effectively "factoring out the size", one needs to select a size measure that serves as baseline for comparison. We used the number of speakers for each listed language across different countries, listed in "Language-Territory Information" compiled by the Unicode Consortium in CLDR version 25 [12].

We normalized the Wikipedia editing and viewing traffic against the speaker data and visualized the outcomes accordingly for comparison.

## 4. RESULTS

Figure 3 and Figure 4 show the normalized outcomes for Arabic Wikipedia. It becomes clear that once the size is factored out, smaller Arabic-speaking countries such as Kuwait, Bahrain, Qatar, and UAE become significant. These smaller countries also have higher penetration rates (above and around 80%), when compared with Saudi Arabia (54%) or Egypt (44.1%), according to the 2012 data provided by the International Telecommunications Union (ITU).

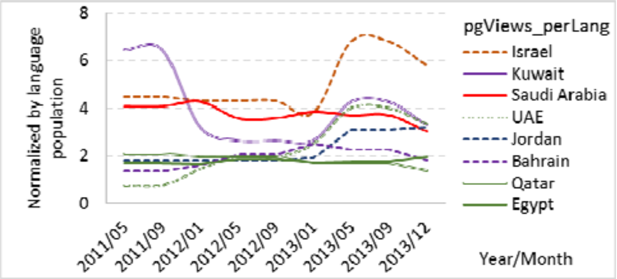

**Figure 3. Normalized viewing traffic trend lines: Arabic Wikipedia**

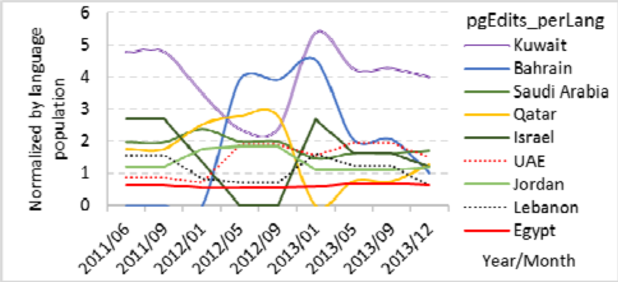

**Figure 4. Normalized editing traffic trend lines: Arabic Wikipedia**

Moreover, it is intriguing to find that Israel is among the top contributors of both editing and reading traffic. Given the political situation in the region, more meaningful interpretation of the findings may require researchers to further breakdown the traffic from Israel based on its even finer geographic and/or linguistic profile.

## 5. DISCUSSIONS

As geographic normalization needs to be justified by choosing a sensible common denominator for comparison and analysis, the proposed geolinguistic normalization points to the need of considering the geolinguistic unit as a sensible alternative unit of analysis for theories and research. In the case of Wikipedia, each language version can been seen as a collaborative project among a group of language users that may stretch across national and regional boundaries. Thus, in this scenario, using geolinguistic unit as a finer unit of analysis (cp. linguistic unit) can provide insights into user activities (both editing and viewing) across regions, while controlling the factor of country size. For instance, we only begin to recognize the relative importance of Israel, Kuwait, etc. in contributing to the editing and viewing traffic of Arabic Wikipedia after such normalization is executed.

Crowston et al [3] proposed a measurement for cross-lingual comparison and found strong correlation between the number of contributors in each language version and two population measurements (Internet population and total tertiary-educated population). We propose here a measurement for cross-regional comparison within each language version. Future research may be conducted to see if similar or other correlations can be identified between the editing (or viewing) traffic and similar population measurements across each geolinguistic unit. For instance, further comparison can be conducted by replacing the number of language speakers in each geolinguistic unit with the estimate number of corresponding Internet population or that of tertiary-educated population.

More research also needs to be conducted on sensible normalization of data or activities generated by online collaborative projects such as Wikipedia. By dividing the observed values against the data of size regarding the offline world, such normalization provides comparative insights not only between the online and offline, but also within the chosen unit of analysis (e.g. across languages and/or across countries).

While the proposed geolinguistic normalization has the potential to provide finer insights to unpack the relationship between online activities and offline conditions, such normalization also demands a finer set of datasets that go beyond just numbers aggregated at language level or country level. Researchers need more numeric size features for geolinguistic units, including their corresponding numbers of population, Internet population, educated population, available offline published resources (e.g. books), etc. As most publicly available datasets are aggregated at the unit of country or language, not at the level of geolinguistic unit, it becomes a challenge for researchers to conduct finer analysis using the geolinguistic grouping as basic unit of analysis.

Nevertheless, an increasing amount of online data contains language and geography information that can be used for analysis based on geolinguistic units. As in the case of our study of Wikipedia traffic report data, the traffic data of a language version is disaggregated into countries based on their geographic locations. In addition, the “language tags” defined by HTML and XML standards, used by modern browsers, and logged by modern web servers, usually contain the user’s linguistic preference at the level of geolinguistic units (such as Kurdish-speaking population in Turkey, identified by the Unicode CLDR as ku_Latn-TR). This is significant because it provides a step away from the pitfalls of using countries as the unit of analysis, or “methodological nationalism”[15].

Thus, while comparative analysis based on the geolinguistic units may face some data and methodological challenges, the benefits and potentials in harnessing and analysing web data at finer levels seem to outweigh the difficulties ahead. One way to tackle the research challenge is to build and maintain a geolinguistic database of “size” measurements, including population, literate population, Internet population, etc.

For the strategic development of open collaborative communities, the proposed geolinguistic normalization and database may better target a specific group of users. For instance, it is possible to identify, within a Wikipedia language version, which country has developed more human and/or content resources, and which country has more room to grow. In that regard, the finer-sized offline and online measurements at the level of geolinguistic units demand researchers and practitioners to pay attention to the differences of the socio-economic environments behind different “language tags” such as ar-EG, ar-QA, ar-SA, etc.

## 6. ACKNOWLEDGMENTS

We acknowledge the open source software tools called Scrapy for making the web mining tasks easier.

## 7. REFERENCES

[1] American Planning Association 2006. *Planning and Urban Design Standards*. John Wiley & Sons.

[2] Cote, P. Effective Cartography: Mapping with Quantitative Data. Harvard Graduate School of Design.

[3] Crowston, K. et al. 2013. Sustainability of Open Collaborative Communities: Analyzing Recruitment Efficiency. *Technology Innovation Management Review*. January: Open Source Sustainability (2013).

[4] Featherstone, M. and Venn, C. 2006. Problematizing Global Knowledge and the New Encyclopaedia Project. *Theory, Culture & Society*. 23, 2-3 (May 2006), 1 –20.

[5] Graham, M. et al. 2011. *Geographies of the World’s Knowledge*. Convoco!

[6] Graham, M. and Zook, M.A. 2013. Zero Geography: Augmented Realities and Uneven Geographies: Exploring the Geo-linguistic Contours of the Web. *Environment and Planning*. 45, 1 (2013), 77–99.

[7] Liao, H.-T. 2013. How does localization influence online visibility of user-generated encyclopedias? A case study on Chinese-language Search Engine Result Pages (SERPs). *Proceedings of the 9th International Symposium on Open Collaboration* (Hong Kong, Aug. 2013).

[8] Liao, H.-T. and Petzold, T. 2010. Analysing Geo-linguistic Dynamics of the World Wide Web: The Use of Cartograms and Network Analysis to Understand Linguistic Development in Wikipedia. *Cultural Science*. 3, 2 (2010).

[9] List of Wikipedias by speakers per article - Meta: *http://meta.wikimedia.org/wiki/List_of_Wikipedias_by_speakers_per_article*. Accessed: 2014-05-05.

[10] Okoli, C. et al. 2012. The people’s encyclopedia under the gaze of the sages: a systematic review of scholarly research on Wikipedia.

[11] Petzold, T. et al. 2012. A world map of knowledge in the making: Wikipedia’s inter-language linkage as a dependency explorer of global knowledge accumulation. *Leonardo: Art, Science and Technology*. 45, 3 (2012), 284–284.

[12] Unicode Consortium 2014. Language-Territory Information, CLDR Version 25.

[13] Wikimedia Meta 2012. Global Development. *Wikimedia Meta*.

[14] Wikipedia Statistics - Site map: *http://stats.wikimedia.org/EN/Sitemap.htm*. Accessed: 2014-05-05.

[15] Wimmer, A. and Schiller, N.G. 2002. Methodological nationalism and beyond: nation–state building, migration and the social sciences. *Global networks*. 2, 4 (2002), 301–334.