Population's preferences by editing Wikipedia:12 worldwide languages

Yerali Gandica

Université catholique de Louvain, Louvain-la-Neuve, Belgium.

Although Wikipedia (WP) is available in 291 languages, the amount of content covered by different languages differs significantly. Even though there is no central authority who dictates which topics should be covered, one of the challenges for Wikipedia, in the last decade, has been to balance the coverage of content across its different languages. For this purpose, a recommendation system is applied by the Wikimedia Foundation in order to encourage Wikipedians to fill that gap.
However, contributing to Wikipedia means more than writing encyclopedic contents. Indeed, it allows communities to store cultural memories of events, to show reality through their own lens and to document their prominent people and places. In this sense, even though we understand the inconvenience of the imbalance between the information among several languages, we hypothesize that this genuine gap has some important implications. It represents legitimate preferences among individuals sharing the same language, which is a footprint of the whole groups' collective identity.
Our goal, in this communication, is to analyze the broad preferences by the population who is editing Wikipedia, depicted by categories over several worldwide languages. In addition, we are also investigating cultural language-based footprints, since it has a tendency to disappear by the globalization. Our analysis is, hence, limited to the first 10 years of the editions in each language, when no intervention to cover the gap between languages had yet been done. Our study covers twelve Wikipedias: the ones written in English (EN-WP), Spanish (ES-WP), French (FR-WP), Portuguese (PT-WP), Italian (IT-WP), Hungarian (HU-WP), German (DE-WP), Russian (RU-WP), Arabic (AR-WP), Japanese (JA-WP), Chinese (ZH-WP) and Vietnamese (VI-WP). Our selection has been done based on the interplay between a worldwide view and the WP sizes. Some limitations are present in our study, as the fact that to some extent, some WP languages have more global than local character, as for example the English one (EN-WP), which is worldwide edited. This language is only used for comparative purposes.
As an example, we start showing the number of edits in Fig. 1. In Y-axis is represented in different colors the proportion of edits in each category. In the upper part is expressed the total number of edits. Naturally, the EN-WP surpasses the other languages. FR-WP and ES-WP follow as the next more edited languages. The most edited categories are "Art" and "History", followed by "Nature" and "Politics" on a lower. All the languages follow roughly similar patterns, with some interesting particularities. For example, the category "Art" is predominantly edited in the DE-WP. While "History" dominates HU-WP and JA-WP. "Nature" appears more important for RU-WP, and "Politics" is the most edited category in VI-WP. We also report results about the number of editors and pages into categories, displayed by the several languages. Results are shown by several angles, and some extra measures complement the analysis.

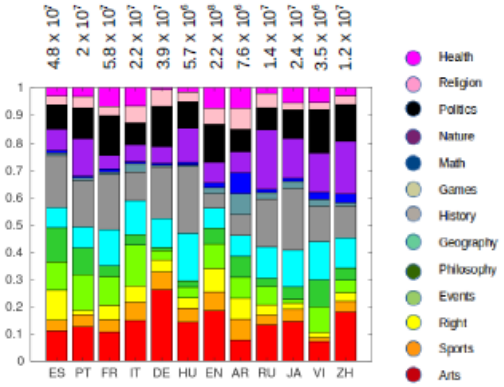

Figure 1: Distribution of edits for each language. A color is associated to each category. The coloring of each column gives the proportion of each category with respect to the total number of edits for the given language.