Population's preferences by editing Wikipedia: 12 worldwide languages

Yérali Gandica[1,2]

1 CeReFiM, University of Namur, Namur, Belgium.
2 Namur Center for Complex Systems - naXys, Université de Namur, Namur, Belgium.

Extended Abstract

One of the most important challenges for Wikipedia, in the last decade, has been to increase the coverage of content across different languages [1]. In this sense, a recommendation system is being applied by the Wikimedia Foundation in order to encourage Wikipedians to cover that gap, via collections of redlinks (hyperlinks from an existing article to a non-existing article that should be created) or tools such as "Not in the other language"[2].

We understand the inconvenience of the lacking of information among several languages and then the important effort in covering such as gap. However, this collectively 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. Language is a social construction, consequence of the human needs to express themselves. This social construction is part of the historical reservoir, shaping the cultural (and hence collective) identity.

Our goal, in this communication, is to analyze the broad preferences over the population who is editing Wikipedia, depicted by categories over several word-wide languages. It is rather common to encapsulate people into cultural boxes, depending on their original countries, where their childhood shaped their initial inner-cultural background. But, what can Wikipedia tell us about this preconceived cultural cliché that each society is used to be singled out? Are we able to uncover language-based patterns by means of the voluntary edition of Wikipedia? Those are the questions we will try to answer in this work. Finally, we are also interested in tracking this human footprint, as it has tendency to disappear as a consequence of the nowadays globalization.

Based on these questions, our analysis is hence confined to the first 10 years of the editions in each language, when not any intervention was yet done. Our study covers twelve Wikipedias. The ones written in English, Spanish, French, Portuguese, Italian, Hungarian, German, Russian, Arabic, Japanese, Chinese and Vietnamese. Our selection has been done based on the interplay between a worldwide view and the Wikipedias sizes. Some limitations are present in our study, as the fact that in more or less extend some WP languages have more global than local character, as for example the English one, which is worldwide edited. This language is used only for comparative purposes.

In Fig. 1 we show the number of editors for each language. Each color represents the number of editors in each category. In the inverted axes is illustrated the proportion of each category over the total number of editions for that language.

In the following we have classified authors in terms of their editing activity. For that endeavor we have taken the maximum number of editions in each language. Then, we defined as super-users the editors whose number of editions are over the 75 % of the top value. Medium users the ones with number of editions between the 75% and 25% and lower users to the rest. In Fig. 2 we show for each language the distribution of the editors' population, classified as previously defined. For each language is shown in the upper part the values for each distribution, while a zoom is shown for the same category, in the inverted axes. This result will be contrasted in the final version of the paper with the last section, where we will show quantitatively the unbalance between the editors' activity for each category and in each language, by means of gini-coefficients. Then, the conclusions are despited in terms of which categories are developed by more or less specialized

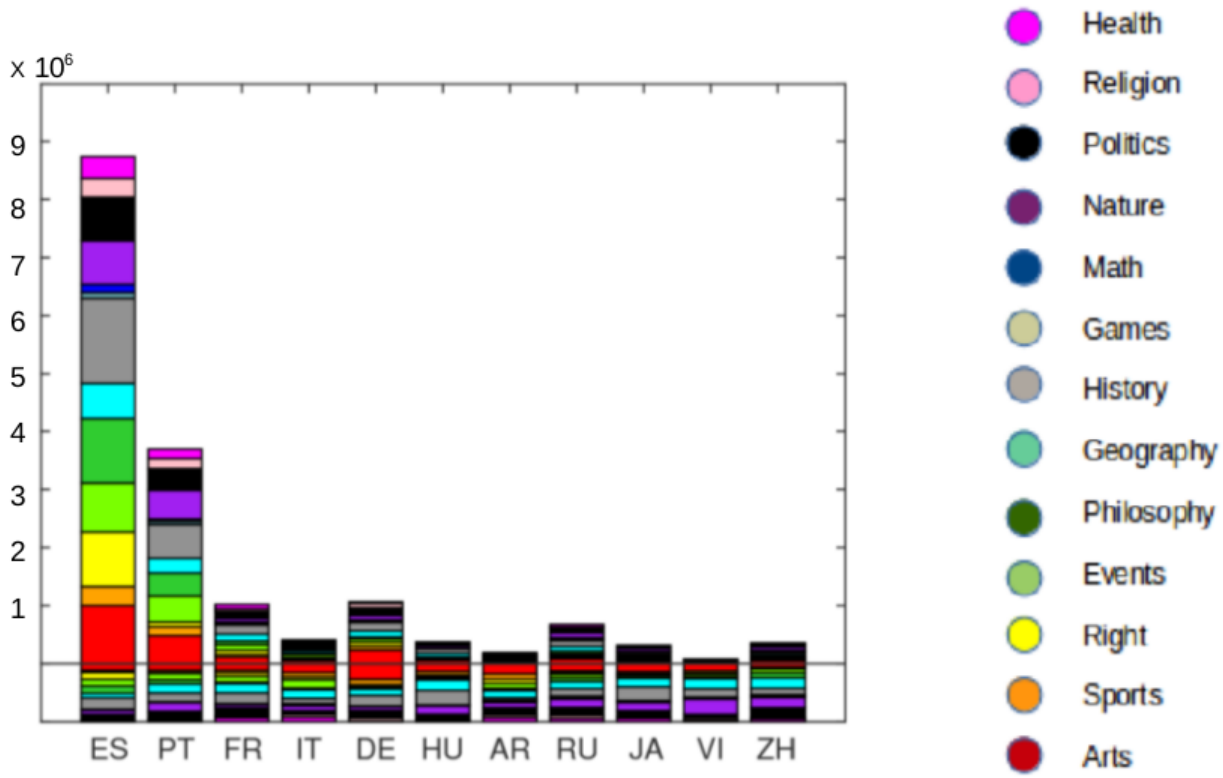

Figure 1: number of editors for each language. Each color represents the number of editors in each category. In the inverted axes is illustrated the proportion of each category over the total number of editions for that language

editors for each language.

References

[1] Ellery Wulczyn, Robert West, Leila Zia, and Jure Leskovec. 2016. Growing Wikipedia Across Languages via Recommendation. In Proceedings of the 25th International Conference on World Wide Web (WWW '16). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 975-985. DOI: https://doi.org/10.1145/2872427.2883077

[2] M. Manske. Not in the other language. https://tools. wmflabs.org/not-in-the-other-language/.

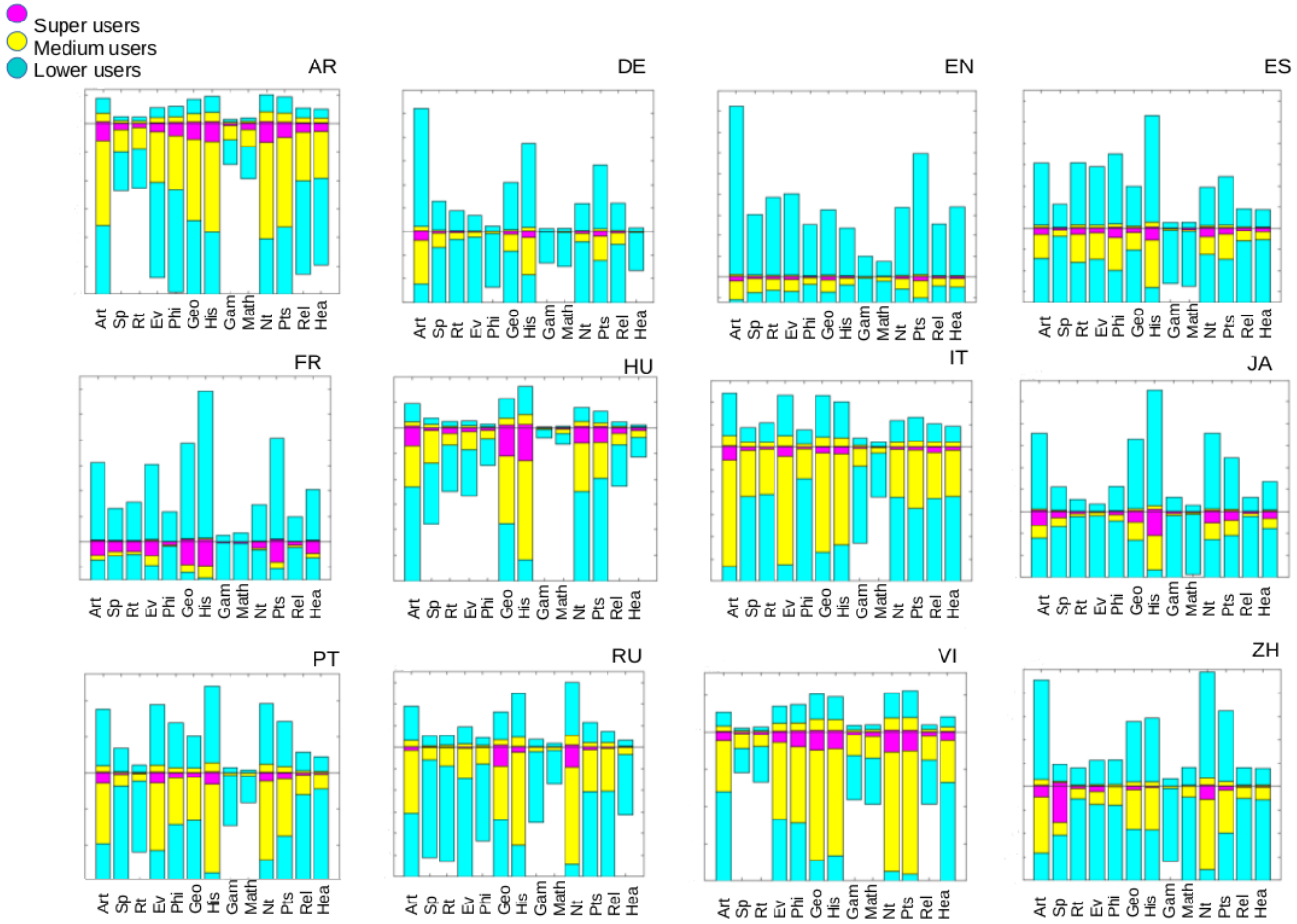

Figure 2: Editors classified in terms of their editing activity. A zoom for each language is shown in the inverted axes of each category.