Wiki Workshop (12th edition) – May 21-22, 2025

Translation Imbalances Between Wikipedia Language Editions

Adam Wight
Wikimedia Germany

Simulo
independent / Uni Bremen

Kavitha Appakayala
Wikimedia Foundation

Abhishek Bhardwaj and **Nathaly Toledo**
Outreachy & Wikimedia

Abstract

Wikipedia consists of 342 separate language editions, and users are able to translate articles between these languages. We hypothesized that the number of translations from each language might be proportional to the number of articles available in that language, but we found only weak correlations. Instead, translations are being made from certain languages such as English at a rate totally out of proportion to relative wiki size and activity.

Introduction

Wikipedia articles are written by editors active in each language based on secondary sources from the same language when possible. However, a special workflow exists for translating articles directly between different language editions.

The Content Translation tool released in 2014[1] simplifies the translation workflow, and has been used to translate over 2 million[2] articles between Wikipedia language editions. The tool offers visual editing, machine translation, and conversion between site-specific building blocks.

This tool leaves a record of each published translation and we hypothesized that the number of translations between languages would be related to the number of active editors on the target wiki.

Methods

We did basic statistical analysis on public and private Content Translation and Wikipedia databases, comparing aggregate translation counts with summary statistics for each language edition.

This paper introduces a new measurement called "translation hegemony", defined as the total number of outgoing translations from a single language A to every other language, divided by the total number of incoming translations from any language into A. A translation hegemony greater than 1 means that more translations are being made from that language than into it.

Results

Out of the total 2,144,411 translations made using the Content Translation tool, 1,572,685 or 73.3% originated with an English article. 36,309 or 1.7% of translations are made into English. This gives the English edition a translation hegemony of 41.4 . The next-highest translation hegemony is a three-way tie: Russian, German, and Bosnian all have a translation hegemony of roughly 3.

English is an outlier in many ways, so let's examine two more comparable editions: Spanish and German Wikipedias have 2,005,319 vs. 2,981,879 articles respectively (or 1 : 1.49), and a similar proportion of active editors (people making 5 or more content edits in the most recent month) at 13,730 vs. 18,557 (or 1 : 1.35). However, German has a translation hegemony of 3.0 (38,486 translations out and 12,693 translations in) while Spanish has a translation hegemony of 0.57 (87,483 translations out and 154,514 translations in).

Comparing translation hegemony with overall wiki size (Figure 1) we find a correlation of $r = 0.62$. Removing English as an outlier of an order of magnitude higher translation hegemony, the remaining languages show only a weak correlation of $r = 0.41$.

We also checked whether the number of active editors could be better correlated with translation hegemony but found similar results (Figure 2): we found $r = 0.96$ but once English is removed the correlation drops to $r = 0.39$.

Public data sources and notebooks are shared on the research project page.[3]

Discussion

Digging into specific languages which show disproportional translation flows, we see large flows between regionally- and historically-connected languages such as English to Spanish (United States-Mexican border), Castilian Spanish to Catalonian, and Russian to Ukrainian. This hints at a mechanism such as informational magnetism[4] or patterns of editor multilingualism.

[1] (Laxström et al., 2015), `https://www.mediawiki.org/wiki/Content_translation`

[2] `https://en.wikipedia.org/wiki/Special:ContentTranslationStats`

[3] `https://meta.wikimedia.org/wiki/Research:Content_Translation_language_imbalances`

[4] (Mark Graham and Hogan, 2015)

Another possibility is that the wiki size is less of a factor than its topical completeness, and translators often find their desired source topic in the English edition. However, work by the Wikipedia Diversity Observatory suggests that overlaps between language editions are small, and contradicts the idea that English Wikipedia is a superset of other wikis.[5] In other words, there should still be plenty of opportunities for translating important articles such as cultural context content defined in the paper.

A strong factor working against translations into English is that the Wikipedia community on that language edition has chosen to disallow machine translation into English after a 2016 discussion about the low quality of automatic translation at the time.[6] Machine-assisted translation has proven to be very popular with the users of Content Translation, and is integrated into the tool for hundreds of other language pairs.[7]

The topical biases in translation offers another potential avenue for research. The previous Content Translation suggestion algorithm defaulted to showing a list of popular articles (by page views) from the source language. Following these suggestions can be assumed to reproduce Anglophone culture in the target wiki—popularity on the source wiki is unrelated to cultural context on the target wiki. Today, Content Translation has updated the suggestion algorithm to include a more tailored “for you” list which tries to find articles related to the editor’s previous editing activity.

An experimental Content Translation workflow called “section translation” might reduce extrinsic topical biases because the article to be translated is one that the user is already reading, not one chosen by an algorithm.

Conclusions

No single, obvious explanation emerged from our research to date.

References

[Laxström et al.2015] Niklas Laxström, Pau Giner, and Santhosh Thottingal. 2015. Content translation: Computer-assisted translation tool for wikipedia articles. *CoRR*, abs/1506.01914.

[Mark Graham and Hogan2015] Ralph K. Straumann Mark Graham and Bernie Hogan. 2015. Digital divisions of labor and informational magnetism: Mapping participation in wikipedia. *Annals of the Association of American Geographers*, 105(6):1158–1178.

[Miquel-Ribé and Laniado2021] Marc Miquel-Ribé and David Laniado. 2021. The wikipedia diversity observatory: helping communities to bridge content gaps through interactive interfaces. *Journal of internet services and applications*, 12(1):10.

[Tiedemann and Thottingal2020] Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt–building open translation services for the world. In *Annual Conference of the European Association for Machine Translation*, pages 479–480. European Association for Machine Translation.

[5] (Miquel-Ribé and Laniado, 2021)

[6] https://en.wikipedia.org/wiki/Wikipedia:Content_translation_tool#English_Wikipedia_restrictions

[7] (Tiedemann and Thottingal, 2020)

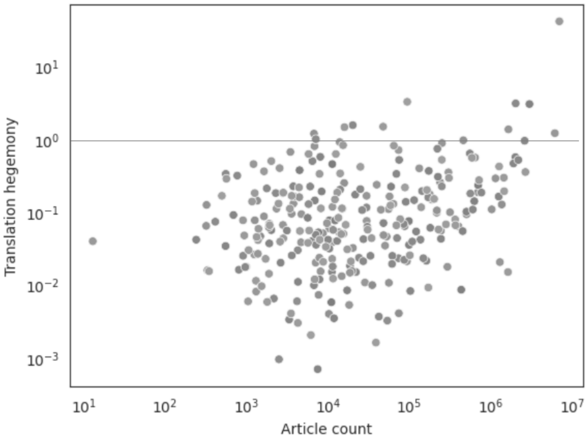

Figure 1: Wiki article count vs. translation hegemony

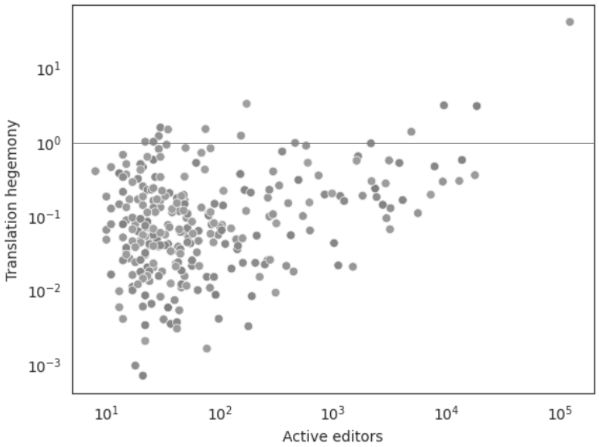

Figure 2: Wiki active editor count vs. translation hegemony

Legend: In the figures above, each dot is a language edition of Wikipedia. The horizontal lines at 10^0 indicate a translation hegemony of 1, or an equal flow of outgoing and incoming translations.