Wikipedia Culture Gap: Quantifying Content Imbalances Across 40 Language Editions

Author - Marc Miquel-Ribe and David Laniado  
Link to access the full paper -<https://www.frontiersin.org/articles/10.3389/fphy.2018.00054/full>  
Suggested by Abhishek Bhardwaj  
The online encyclopedia Wikipedia is the largest collaborative information repository, but there are content imbalances across different language editions. To investigate these imbalances, a computational method was developed to identify articles related to cultural context for 40 language editions, using geolocated articles, specific keywords and categories, and links between articles. Manual assessment found an average precision of 0.92 and an average recall of 0.95. Results show that a quarter of each language edition is dedicated to representing its cultural context, and this content is sustained over time. Cross-language coverage analysis reveals gaps and unique content, and the approach and findings can foster participation and inter-cultural enrichment of Wikipedias. The datasets produced are available for further research. The above research can be used to get an insight into the translational imbalances and also serve as a base to the future work we aim to accomplish.

The authors have expanded on this project by creating the [Wikipedia Diversity Observatory](https://meta.wikimedia.org/wiki/Wikipedia_Diversity_Observatory), and show that the coverage gaps exist in both directions, in other words English Wikipedia is not a superset of all other wikis.

Cross-lingual knowledge linking across wiki knowledge bases

Author - Zhichun Wang, Juanzi Li, Zhigang Wang and Jie Tang  
Link to access the full paper - <https://dl.acm.org/doi/abs/10.1145/2187836.2187899>  
Suggested by - Abhishek Bhardwaj  
Wikipedia has become one of the largest knowledge bases on the web with 513 million page views per day in January 2012. However, articles in different languages are very unbalanced, with English having 3.8 million articles while Chinese has less than half a million. This raises the question of how to link knowledge entries across different knowledge bases, which would greatly benefit many applications. This paper presents a linkage factor graph model for cross-lingual knowledge linking, defining features according to interesting observations. Experiments on the Wikipedia data set show a high precision of 85.8% with a recall of 88.1%, resulting in 202,141 new cross-lingual links between English Wikipedia and Baidu Baike. We can explore these cross lingual knowledge linking and observe and measure its impact on the translation imbalances.

I think the paper's relevance to research into translation imbalances is significant. The problem of translation imbalances affects the creation of cross-lingual links, which is essential in globalising knowledge sharing and improving information retrieval and machine translation. The paper identifies the challenges posed by translation imbalances and proposes possible solutions to the problem, such as finding language-independent features for mining cross-lingual knowledge links.  
Overall, the paper highlights the need to address translation imbalances in cross-lingual knowledge linking, which is crucial in improving multilingual access to information and promoting knowledge sharing across languages.

### Why the World Reads Wikipedia: Beyond English Speakers

Authors - Florian Lemmerich, Diego Saez-Trumper, Robert West, Leila Zia  
Link to access the full paper - <https://dl.acm.org/doi/abs/10.1145/3289600.3291021>  
Suggested by - Abhishek Bhardwaj  
Wikipedia is a primary multilingual knowledge source, read by millions of people worldwide daily. However, little is known about why users read different language editions. In a comparative study, a large-scale survey of Wikipedia readers across 14 language editions was combined with a log-based analysis of user activity. The study proceeds in three steps, analyzing survey results, matching responses to server logs, and characterizing behavioral patterns. The study found commonalities and differences among Wikipedia languages, distinctive patterns marking certain use cases, and certain use cases more common in countries with specific socio-economic characteristics. These findings advance understanding of reader motivations and behaviors across Wikipedia languages and have implications for Wikipedia editors and developers of Wikipedia and other Web technologies. The above paper can be used to analyze the cross lingual usage patterns and how are they used.

 In the task [#T331207](https://phabricator.wikimedia.org/T331207) there were a lot surveys in which my fellow contributors asked about the motivation of translation. I think the motivation to read and the motivation to translate are quite closely related (though this is only a hypothesis as of now). As if the motivation to read some specific type or genre of articles is high the motivation to translate them to other languages will also be higher. This paper gives a really good insight into the motivation of people to read a wiki article which I think can be extrapolated to the motivation of translators to translate the article.