Words of Wisdom, Wars of Words

Harvard University - GOV 2001 – Final Project – Spring 2016

Edmond Awad and Sohan Dsouza

Abstract

Through history, language has been a highly utilized tool for humans to communicate in order to establish cooperation over shared resources and to avoid destructive conflicts. The importance of language manifested itself through the evolving and contemporary technology from printing to online forums and blogs. The recent availability of data allowed the possibility for quantifying the influence of languages and analyzing the dynamics that govern their relations. This study aims at contributing to this line of work, by expanding on an earlier study by Ronen et al to causally link the global centrality of a language to the global fame of its natives. We also use data on international interactions to discover a statistical link between shared languages and increased interaction – both positive and negative.

1 Introduction

The evolutionary success and dominance of humans over other species owes the faculty of language perhaps more than any other ability [7]: a self-constructed system of communication, for self-expression, exchange of ideas, for organizing themselves into purposeful groups and successfully coordinating the share of the available resources with others. This enabled humans to create complex social structures, define friend and foe, and gradually articulate hierarchies of society from the familial to the national. In turn, the central role played by language in social development has assured it a place as one among the defining characteristics of the societies that use it. Indeed, it can miscommunicate and be used to other just as well as it can communicate and be used to unite [13]. This makes the effect of language on culture and society a pertinent field of study for the understanding of history, and of a modern, globalized, connected world of speakers of different languages coming into more frequent contact than ever before.

Books have been the primary vehicle for the transmission of information for millennia. Although their form has changed over their existence – from clay tablets to papyrus to vellum to paper – and technological advances have rendered them increasingly accessible. And in the world of books, translation has been one of the main ways written information has crossed language barriers, from the medieval preservation in Arabic of ancient European works [6], to Chinese

copies of Indian Buddhist texts [3], to the European Renaissance flourishing of printed Bibles in local languages [15]. In each of these cases, certain languages, being the lingua franca or high languages of their time, became regional or even continental hubs of translation. In a globalized, connected world, such linguistic network centrality scales up to a planetary level.

In the digital age, social media and crowd-sourced platforms play an increasing role in the collection and dissemination of information. Among the former, Twitter has been among the leading social media platforms [1] – especially for public, open communication – for almost the decade it has existed. Among the latter, Wikipedia has dominated the knowledge reference space [2], rendering the traditional encyclopedia industry practically obsolete with its crowd-sourced and -managed mostly-open editing platform. Both platforms support and host content in multiple languages, providing opportunities for acquisition of data on linguistic demographics of contributors.

With their recent acquisition of such data from Twitter and Wikipedia, as well as a database of global translation networks from UNESCO, Ronen et al [14] used data on book translation among different languages, as well as data on multilingual co-expression among Twitter users and Wikipedia editors to map language centrality into three global networks. They confirmed the role of English as a global hub language, with less-central satellite hubs among a few other Indo-European languages and Chinese. Their research also employed measures of global fame based on cross-language Wikipedia presence and , establishing a strong correlation between

We are extending their work using causal inference to verify potential causality in the relationship between Ronen et al's measures of language centrality and fame. In addition, we use data on the frequency and tone of international interactions to examine the relationship between Ronen et al's global language networks and the networks of cooperation and conflict among countries in the world.

2 Data

Most of the data sets used in this study were used in [14]. This includes a collection of more than 2.2 million book translations compiled by UNESCO's Index *Translationum* project, the linguistic co-expressions expressed by the community of digitally engaged knowledge specialists that edit Wikipedia, and linguistic co-expressions expressed in Twitter. Each of these data sets was used to construct a global language network in which a link between two languages means that many translations occurred among them, many users have edited Wikipedia pages in both languages, and many users have tweeted in both languages, respectively.

To measure the global influence of the speakers of a language, two other data sets that estimate the number of famous people associated with each language were used. The first is about people who were born between 1800 and 1950 and have articles in at least 26 Wikipedia language editions. The second is based

on Human Accomplishment, a published volume listing individuals that have made significant contributions to the arts and sciences before 1950 [12]. Finally, another data set that includes the population and GDP of a language was also used.

We added to these a data set of international interactions (refer to Figure 1) from "GDELT - Global Data on Events, Location and Tone" [11, 9]. The data set is produced by the Google-curated GDELT Project, which daily processes news sources in broadcast, print and web media from around the world in over 100 languages. We extracted country-to-country events for the decade from 2006 - 2015, and used a per-country linguistic demographic representation look-up from the CIA World Factbook [4] to link directional language pairs to their respective quantities of material cooperation and material conflict.

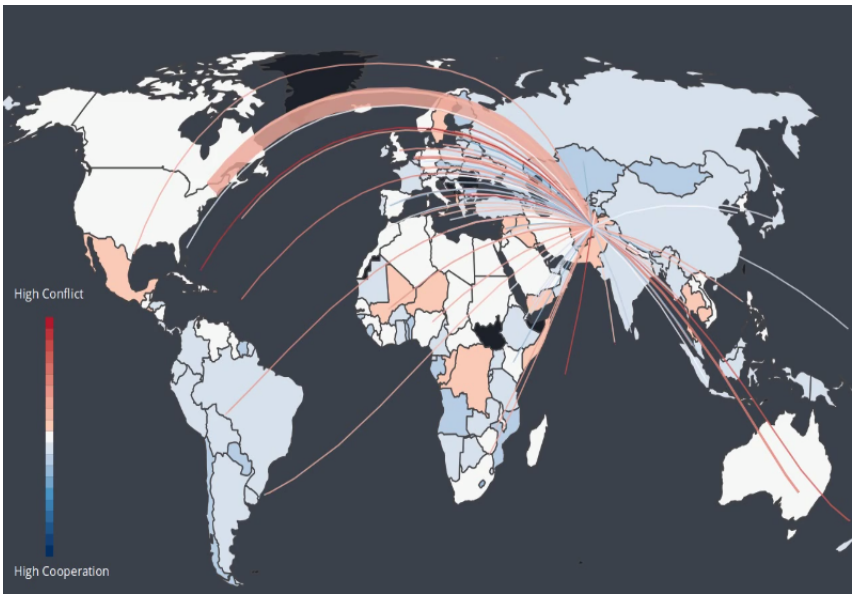


Figure 1: International relations map based on GDELT data, showing the interaction initiated from Afghanistan towards the rest of the world. The color of the links from Afghanistan to a country signals its attitude from that country: red being conflicting, and blue being cooperative. The color of each country signals its overall attitude towards the rest of the world.

3 Methods

We start by analysing the work done by [14]. Looking at the distribution of the dependent variables (i.e. number of famous people in both data sets), they seem to be right-skewed (and heavy-tailed). A possibility is that the dependent variables follow a log-normal distribution. As such, in their statistical models,

the authors assumed the dependent variables to follow a log-normal distribution. Accordingly, they transformed all dependent and independent variables using log_{10}, and applied classic linear regression (Gaussian, specifically) to the transformed data.

To test the plausibility of the models they used, we calculated the robust standard errors for all 14 models. They all seemed to closely match their respective standard errors from the original publication. Additionally, we tested the plausibility of the following underlying assumptions of the models: *Global Stat*, *Skewness*, *Kurtosis*, *Link Function*, and *Heteroscedasticity*.

For most of the models, all the above assumptions were found admissible. For others, only the *Global Stat* and *Link Function* were found to be not admissible. Since Skewness and Kurtosis are admissible for all models, the choice of the normal distribution (on log-transformed data) seems to be valid.

After that, we analyzed the possibility of any biases when studying causal inference [10]. The most relevant possible bias in our case is the extrapolation bias. Since the variable of interest, the eigenvector centrality of a language, is continuous, we converted it to a dichotomous variable by thresholding at 0.01 in all three data sets (Twitter, Wikipedia, and book translations). This value was appropriately chosen to have relatively similar-sized control and treatment sets. Values below that threshold were considered "not central" in the eigenvector sense, and the others are considered "central".

To make sure that this step did not make any effect on the previous analysis, we plot the in-sample fit charts for each data set before thresholding against its counterpart after thresholding, and measured the correlation. Further, we plot the in-sample fit chart for each pair of the three data sets after thresholding and compared them with their corresponding fit charts before thresholding. For conciseness, we don't include these charts in this report.

Using the modified data sets, we checked the imbalance between treated and control data sets using multivariate imbalance measure L1, and the local common support (LCS). The calculated values indicated a relatively high imbalance $(0.4 - 0.6)$ and a moderate percentage of local common support $(33 - 58\%)$.

This indicates high model dependence in the data. Calculating the percentage of points in the convex hull [10], a more conservative method for testing for imbalance showed that 0% of the counterfactual points are within the data points.

This called for the use of matching in order to decreases the imbalance [8]. We tested the following matching techniques and their variations: propensity score matching, Mahalanobis matching, Mahalanobis matching with replacement, Mahalanobis matching with caliper, coarsened exact matching (CEM), and CEM with cuts. We got the best results using the propensity score matching for Twitter data set, and CEM for Wikipedia and book translation data sets in terms of imbalance decreases and percentage of local common support increase.

Checking the imbalance and local common support after matching, we noticed an improvement in both the imbalance $(0.17-0.38)$ and the LCS $(57-75\%)$ for the three data sets.

	ATT	Relative increase in probability of fame
Twitter EG Centrality	182	284%
Wikipedia EG Centrality	75	789%
Book trans. EG Centrality	102	581%

Table 1: Average treatment effect on the treated and the relative increase in probability of fame for a language using the three data sets.

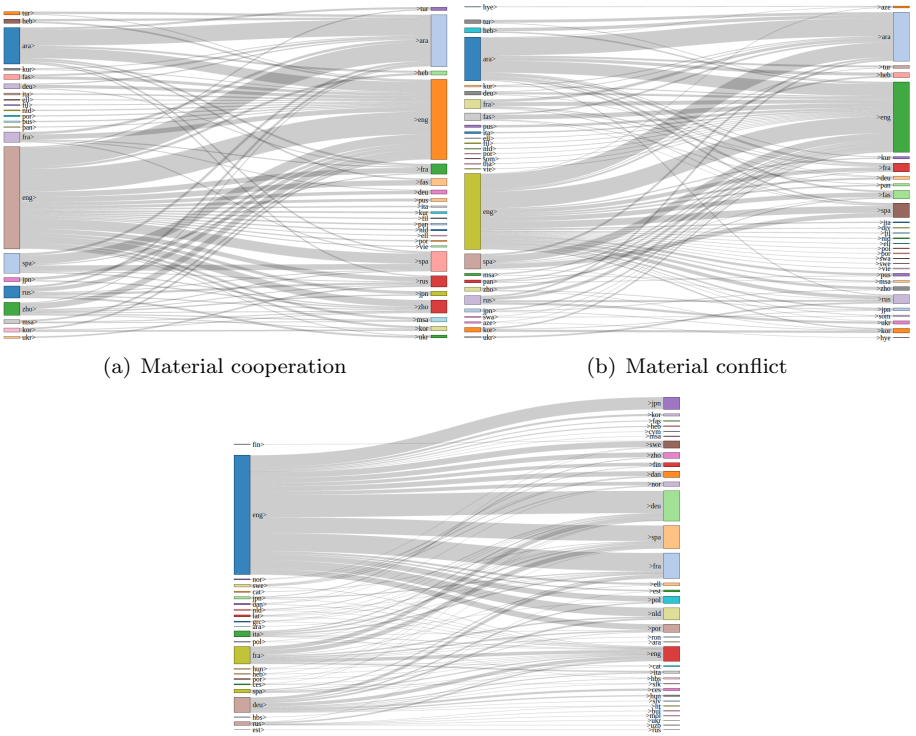
4 Results

After preprocessing the data set used in [14] as explained before, we were able to quantify the causal effect of the centrality of a language on the number of famous people that speak this language, for each of the three data sets. Table 1 shows the average treatment effect on the treated (i.e. when the language is central central) ATT. Additionally, we calculate the probability of being famous in a language by dividing by the number speakers, and use it to calculate the relative increase in this probability for a person when speaking a central language instead of a non-central language.

Next, we move to analyze the effect of a language centrality on the number of cooperative and conflicting interactions that it initiates. As it turns out, when controlling for the GRP per capita and the population of each language, the language centrality has no significant association with the the cooperative or conflicting interactions it initiates. We found that using matching though decreases the imbalance and and increases the percentage of the local common support, the resulting data provides inconsistent results regarding the causal effect of language centrality and amount of interactions produced. Theoretically, controlling for other co-variates (GDP and population in our case) after matching should not have any effect given that the treatment variable and the covariates are unrelated. Unfortunately, this is not the case. While not including the covariates after matching shows a significant effect of the treatment, controlling for those variable shows no effect.

In fact, understanding the international relations among two languages would be better understood given other relations between this pair of languages. As such, analyzing the effect of the number of bilingual online users or number of translated books among a pair of languages would make a better choice than the language centrality (refer to Figure 2 for an overall impression). Additionally, this number has better intuitive explanation (e.g. number of translated books between two languages) instead of a centrality measure which can only be understood on a high level (e.g. high/low centrality) without any specific meaning of the actual value itself (e.g. what does it really mean to have an eigenvector centrality of 0.1?).

Motivated by this, we turned to analyzing the correlation and causation of links weight between a pair of languages in the three data sets on the amount of positive (cooperation), and negative (conflict) interactions between them.



(a) Material cooperation

(b) Material conflict

(c) Book translations

Figure 2: Three Sankey plots illustrating relative weight of links among pairs of languages from left to right, scaled by (a) number of material cooperation actions, (b) number of material conflict actions, and (c) number of book translations. One can notice the high similarity between the number of cooperation and conflict actions among pairs of countries. A less similar pattern can be also seen in the number of book translations among pairs of countries.

The first observation is that for a pair of languages, as the number of users who speak both languages (more specifically, tweet or edit in Wikipedia in both languages) increases, the amount of positive and negative interactions both increase. Figure 3 shows this correlation without controlling for other variables.

Unfortunately, as we control for the GDP and population of each of the two languages and the interaction between their populations and their GDPs, only the number of common Twitter users correlates with the amount of both positive and negative interactions. Figure 4 shows the effect of including/excluding of the Twitter data in the model. Though, this variable does not provide a high semi-partial correlation (i.e. increase in R^2), but an F-test of the two models provide $[F(1,162) = 4.54, p = 0.035]$ and $[F(1,162) = 3.76, p = 0.05]$ for the cooperative and conflicting interactions respectively.

To confirm the validity of the models, we produced the robust standard errors

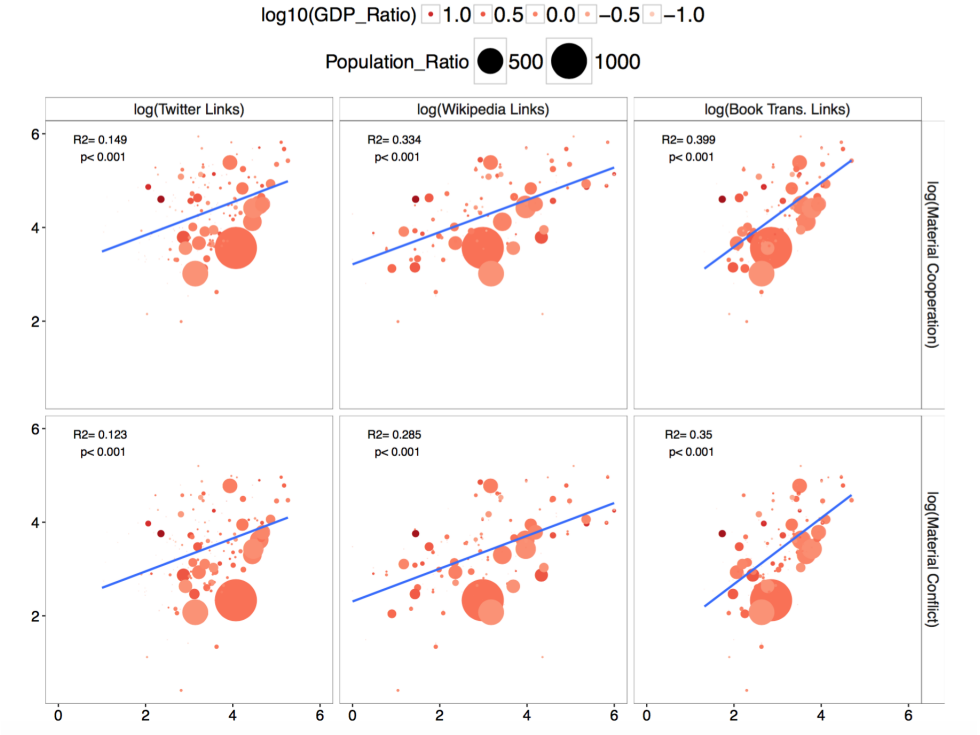


Figure 3: The number of common users/publications of a pair of languages and the amount of interaction among pair of languages. Top row shows the amount of cooperation as a function of common users/publications. The lower row is for the amount of conflict. The figure shows that the correlation is higher for Wikipedia bilingual users and book translations than for the number of bilingual Twitter users.

which turned out to be very close to the standard errors. After applying matching on the data, the effect of number of common Twitter users becomes slightly stronger. The ATT and the relative increase of interactions (RII) are $[ATT = 4421, RII = 92\%]$ for cooperative interactions and $[ATT = 4381, RII = 56\%]$ for conflicting interaction.

5 Discussion

The results of this work are merely an attempt at analyzing the causal effect of connections between languages on their influence and spread. A central language increases visibility of its speakers to the rest of the world, thus increasing their chances of being publicly recognized.

Moreover, two languages being mutually commonly spoken raises the like-

	Amount of material cooperation actions taken between two languages, based on actions between countries.	
	(1)	(2)
$\log_{10}$(Population1)	0.373***	0.318***
	(0.104)	(0.106)
$\log_{10}$(GDP1)	-11.064***	-10.540***
	(2.903)	(2.883)
$\log_{10}$(Population2)	0.348***	0.297***
	(0.090)	(0.092)
$\log_{10}$(GDP2)	-11.368***	-10.835***
	(2.840)	(2.821)
Twitter		0.077**
		(0.036)
$\log_{10}$(Population1)*$\log_{10}$(Population2)	0.154***	0.148***
	(0.043)	(0.043)
$\log_{10}$(GDP1)*$\log_{10}$(GDP2)	2.670***	2.525***
	(0.654)	(0.651)
(intercept)	48.493***	46.616***
	(12.544)	(12.441)
Observations	170	170
R^2	0.708	0.716
Adjusted R^2	0.697	0.703

***,**,* significant at 0.1%, 1% and 5% levels, respectively. Standard errors in parentheses.

(a) Material cooperation

	Amount of material conflict actions taken between two languages, based on actions between countries.	
	(1)	(2)
$\log_{10}$(Population1)	0.408***	0.354***
	(0.113)	(0.115)
$\log_{10}$(GDP1)	-10.295***	-9.777***
	(3.154)	(3.139)
$\log_{10}$(Population2)	0.382***	0.332***
	(0.098)	(0.100)
$\log_{10}$(GDP2)	-10.485***	-9.957***
	(3.085)	(3.072)
Twitter		0.076*
		(0.039)
$\log_{10}$(Population1)*$\log_{10}$(Population2)	0.134***	0.128***
	(0.047)	(0.046)
$\log_{10}$(GDP1)*$\log_{10}$(GDP2)	2.466***	2.323***
	(0.711)	(0.709)
(intercept)	45.126***	43.266***
	(13.626)	(13.547)
Observations	170	170
R^2	0.667	0.675
Adjusted R^2	0.655	0.661

***,**,* significant at 0.1%, 1% and 5% levels, respectively. Standard errors in parentheses.

(b) Material conflict

Figure 4: Regression table explaining the amount of interaction as a function of the number of bilingual Twitter users. Note that when controlling for population and GDP per capita for each language, and the interaction of populations in the two languages and the interaction of GDP for both languages, the number of bilingual Twitter users still correlates with the amount of interaction.

lihood of more interactions, positive or negative. This is sensible given that having common means to interact increases the possibility to both forge positive relations e.g. trade and agreements, and have more issues to fight over e.g. resources. As a result, speaking a language that is commonly learned as a second language opens the door for more international interactions with other countries, but as a proverbial double-edged sword.

Future work will focus on more well-defined innate characteristics of languages such as complexity and flexibility, and their effect of the international relations. Specifically, we intend to examine the relationship between languages and cooperation/conflict from a linguistics perspective, grouping languages by grammatical or semantic features, using an approach similar to that used by [5].

References

[1] Alexa.com. Alexa - top sites by category: Computers/internet/on the web/online communities/social networking, 2016 (accessed May 05, 2016).

[2] Alexa.com. Alexa top 500 global sites, 2016 (accessed May 05, 2016).

[3] Daniel Boucher. Gndhr and the early chinese buddhist translations reconsidered: The case of the saddharmapuarkastra. *Journal of the American Oriental Society*, 118(4):471–506, 1998.

[4] Central Intelligence Agency. The world factbook, Apr 2016.

[5] M Keith Chen. The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. *American Economic Review*, 103(2):690–731, 2013.

[6] Matthew E Falagas, Effie A Zarkadoulia, and George Samonis. Arab science in the golden age (750–1258 ce) and today. *The FASEB Journal*, 20(10):1581–1586, 2006.

[7] Marc D Hauser, Noam Chomsky, and W Tecumseh Fitch. The faculty of language: what is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579, 2002.

[8] Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236, 2007.

[9] Kalev H. Leetaru. The gdelt project, Apr 2013.

[10] Gary King and Langche Zeng. The dangers of extreme counterfactuals. *Political Analysis*, 14(2):131–159, 2006.

[11] Kalev Leetaru and Philip A Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. In *of: Paper presented at the ISA Annual Convention*, volume 2, page 4, 2013.

[12] Charles Murray. *Human accomplishment: The pursuit of excellence in the arts and sciences, 800 BC to 1950*. Harper Collins, 2003.

[13] Peter H Nelde. Language contact means language conflict. *Journal of Multilingual & Multicultural Development*, 8(1-2):33–42, 1987.

[14] Shahar Ronen, Bruno Gonçalves, Kevin Z. Hu, Alessandro Vespignani, Steven Pinker, and César A. Hidalgo. Links that speak: The global language network and its association with global fame. *Proceedings of the National Academy of Sciences*, 111(52):E5616–E5622, 2014.

[15] Jonathan Sheehan. *The Enlightenment Bible: translation, scholarship, culture*. Princeton University Press, 2005.