Urban dictionary slang and its influence across the world

Francesco Ignazio Re,Hugo Julien

April 11, 2021

1 Context and motivation

1.1 Introduction

The world of today is fully connected. Digitalization and technological advancement tore down physical boundaries and established continuous flows of information among people belonging to very different cultures and societal contexts. These processes are continuing to grow, with the on going expansion of access to the Internet, even though their implications were firstly analyzed in the second half of last century. Specifically, in the sixties, McLuhan was the first one to coin the term a "global village" (McLuhan, 1962), referring to the shift in dynamics that this progress was bringing on a political, economical and, most of all, on a cultural level. One of the major implications of this on-going phenomenon is that different cultures are now able to influence each other without necessarily physically coexisting in nearby places. The advent of social media seems to have pushed this trend forward, allowing for cultural exchange and communication to happen right in the palm of our hands. In this sense, an interesting point to focus on would regard the understanding of how cultural influences have evolved during the last 20 years. Indeed, a task like this is difficult to tackle. Culture itself is a multi-factor concept that is hard to measure, and possibly even harder to interpret. However, we decide to focus on a single factor that we think may be an expression of culture: language, and in particular, slang.

1.2 Use of slang words

By slang we intend any word or short phrase used in informal registers and typically restricted to specific social groups and environments. These expressions are generally formed and developed along with the history of the country where they are from and, thanks to their colloquial use, they are often a good reflection of the culture they represent (Zhou & Fan, 2013). Furthermore, slang expressions are widely used in movies, plays and on-line, where their use is more easily trackable.

1.3 A case of cultural influence: Americanization

The United States is a country whose culture has been very influential towards the outside world. Specifically, after the end of WWII, the United States provided the major point of reference for many Europeans (and non) companies, consumers and cultures (Kipping, 2002). Since the advent of 2000s, US culture has had still a lot of influence thanks to the presence of big tech companies, pop music culture, the movie industry, social media etc. (on many socials, in fact, a great share of the most influential users are American). Of all these information channels, many of them pass through language, and in many cases, also through slang.

1.4 Project question

Linking together the above points, our research question would then be on the study of American slang influence in the last 15 years, to see if it has been observed a significant increase towards American slang imitation, that has not been witnessed instead with slang belonging to other countries.

2 Project Plan

After laying out the research question, we proceed by identifying the project pipeline. We need a source of world slang expressions, along with the geographical regions where they were used and the year when they were in vogue. We then need a source to get a measure of influence for each word by country, in the time span when the word was most used. Once we obtain this data, we would be able to test for correlation between the years when words were in vogue and their influence values. At this step, we have not yet defined 'influence value', which will be explained in later sections.

3 Data Retrieval

We have two data retrieval steps to execute: we need first to retrieve the slang words and classify them by country, and then we need to retrieve a measure of how much each word is used in countries other than its own.

3.1 A slang word dictionary: Urban Dictionary

Urban dictionary is an online dictionary for slang expressions[1] released in 1999 and still in use today. It contains more than 5 million expressions, and as of now most of them not are not necessarily slang, but rather any word, event or phrase belonging to English or other languages. Insertion of words is allowed to any registered account. This ensures the site to always be up to date with new slang words, but it also implies that there may be many words defined arbitrarily according to the user perception of them. What we will try to do in the pre-processing section is to try to leave out all the words that we think do not pertain our study. In order to retrieve this data, we proceed in the following way:
We first use `urbscrape` [2], a python script that allows scraping and storing every single word and definition from the website. Running the script produces a database, that we now made publicly available, with all words and definition up to March 21 2021. After filtering out the words we are not interested in, we use the python library `urrlib` [3] to retrieve information such as the time when the word was inserted and the likes that the word received on the website.

3.2 Popularity search with Google Trends

Platforms such as Google can be easily mined as sources of raw data about human social interactions. Specifically, Google Trends is a platform that allows to look up a term and show its normalized search value in different areas of the world. The normalization is done with the respect to the highest value among all the selected regions, over the selected time. This implies that, in our study, we will not be taking directly into account the magnitude of a word search, that is the overall volume of searches that the word elicited. Hence, every country value tells us 'how many searches occurred in the country in comparison to the country with the highest value'. Given a word w, given its searched value V_c^w for country c, given C the set of all countries and given c_w the country the word w originated in, then we can define an influence score s as follows:

$$s(w) = \sum_{\{c \in C \setminus \{c_w\}\}} V_c^w.$$

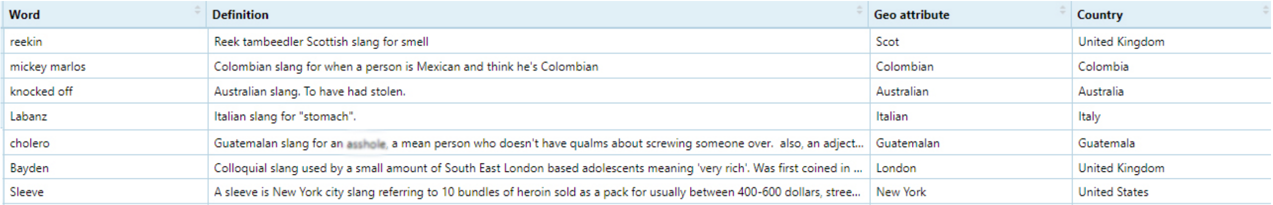
$s(w)$ for word w is the sum of the search values of all countries but the one the word originated in. In most cases, the original country is also the one with the maximum search value of 100. In this case, we can interpret $s(w)$ as the spread that the word had in other countries compared to the country of origin. In the case the country of origin has not value 100, the interpretation changes and may become problematic, as the comparison is done not with the respect to the origin country. We hence decide to annotate the country that had the maximum value for each word, so we can decide how to treat words that had this country differ from their origin one. The Google trend score was calculated during a span of two years, starting from one year before the word was entered on Urban Dictionary, till one year after it was entered.

4 Data Processing

We first need to process the urban dictionary database. We do that with R and the dplyr library. From the SQL database we extract a table made of word names and definitions. We filter only the words containing the string `slang` in their definition. Then, after retrieving all the existing English demonyms associated with the physical place they refer to, we filter all the words containing a demonym in their definition. A demonym is an adjective for the inhabitants of a place. Hence, a possible pair (demonym, city) may be i.g. (Bostonian, Boston). We then also retrieve world cities and regions with their corresponding country. In this way we can redirect every word to the country it belongs. We continue with the following steps:

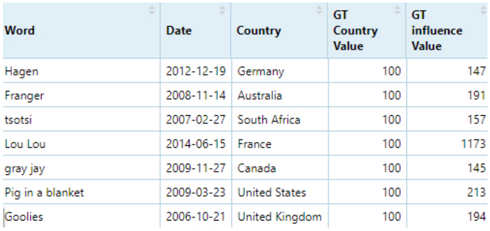
- We aggregate by countries and keep all the countries with more than 100 words.
- We consider that many slang are common words used with a different meaning in specific contexts. This happens specifically in English, that has a huge variety of word meanings according to the country the word is spoken in. For this reason, we decide to filter out all the most common English words.
- We decide to keep only the words that had their maximum value in the country of origin. This helps for interpretation purposes, as suggested in the previous section, and is valid with the assumption that a slang word is always looked up more in the country of origin. Moreover, we noticed that words having more searches in a country different from their own may may be related to meanings different from the slang one. For example, the British slur `full length` has a rather creative meaning if compared to what the rest of the world uses the term for; the country who searched this term the most, that is Trinidad and Tobago, may have referred more likely to 'full length movies' or similar, rather than to the specific British slang.

We still need to make a final small note; that is, urban dictionary is an American website. Hence, when defining a slang word, if no geographical attribute is mentioned, then the user may imply the term to be American. Hence, among all the words that are defined as slang but do not have a demonym, we keep the ones having US as the most influential country on Google Trend and we deem them as American.



Word	Definition	Geo attribute	Country
reekin	Reek tambeedler Scottish slang for smell	Scot	United Kingdom
mickey marlos	Colombian slang for when a person is Mexican and think he's Colombian	Colombian	Colombia
knocked off	Australian slang. To have had stolen.	Australian	Australia
Labanz	Italian slang for "stomach".	Italian	Italy
cholero	Guatemalan slang for an [redacted], a mean person who doesn't have qualms about screwing someone over. also, an adject...	Guatemalan	Guatemala
Bayden	Colloquial slang used by a small amount of South East London based adolescents meaning 'very rich'. Was first coined in ...	London	United Kingdom
Sleeve	A sleeve is New York city slang referring to 10 bundles of heroin sold as a pack for usually between 400-600 dollars, stree...	New York	United States

Figure 1: Dataset sample after processing urban dictionary terms. We have the column word, definition, geo attribute - containing either denomyms or geographical places, and then country.



Word	Date	Country	GT Country Value	GT influence Value
Hagen	2012-12-19	Germany	100	147
Franger	2008-11-14	Australia	100	191
tsotsi	2007-02-27	South Africa	100	157
Lou Lou	2014-06-15	France	100	1173
gray jay	2009-11-27	Canada	100	145
Pig in a blanket	2009-03-23	United States	100	213
Goolies	2006-10-21	United Kingdom	100	194

Figure 2: Dataset sample after retrieving Google Trends scores. We have the column Word, Date, GT Country Value - containing the google trend value of the country of origin, and then GT Influence Value, containing for each word the score defined in section 3.2.

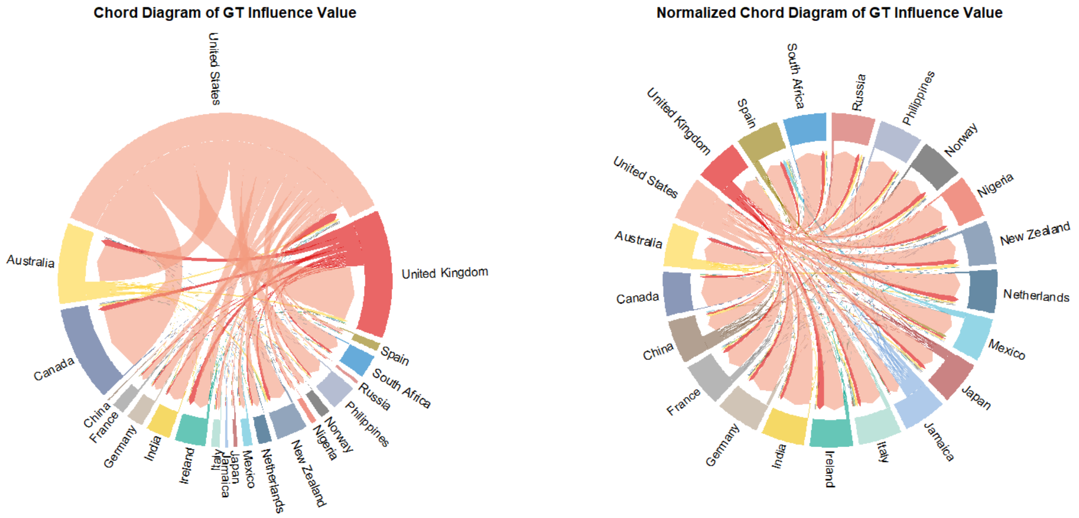
5 Data Analysis

After retrieving the data and processing it for our purposes, we proceed by carrying out the statistical analysis.

5.1 Exploratory Analysis

We start off with the exploratory analysis, with a word corpus of circa 4800 words. In this section we want to analyze the influence flow in terms of google search between countries. In other words, not considering the individual words, we want to understand for each country how much its words are searched and by whom. In order to achieve this, we operate on the adjacency matrix used to create the table in Figure 2.

The matrix has as vertices the countries, namely the ones selected in the Data Processing section, and as weighted edges the overall `GT Influence Value` from each country to the others. We build a Chord Diagram, that we show in Figure 3.

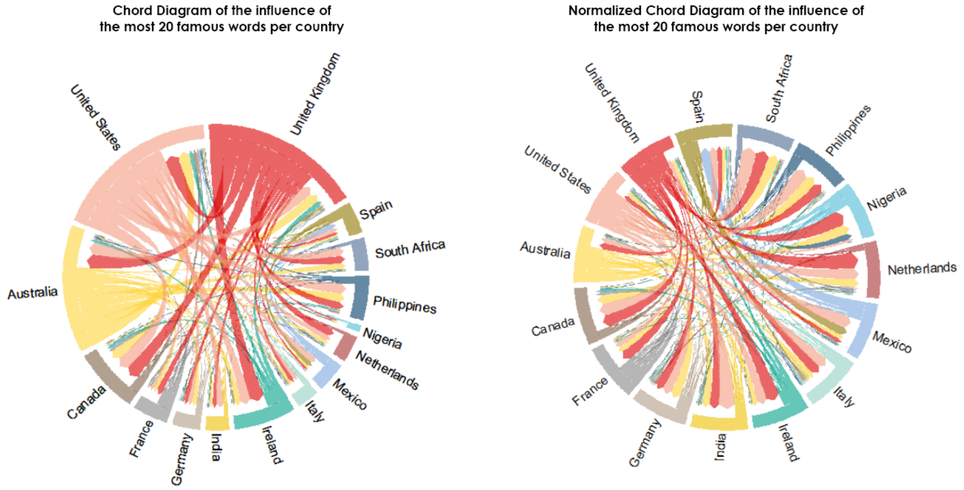


Chord Diagram of GT Influence Value

Normalized Chord Diagram of GT Influence Value

Figure 3: In the left figure we have the non-normalized `GT Influence Value` of each country being redirected towards the countries that conducted the search. In the right figure, for each country, the incoming search values and the outgoing ones sum to the same quantity. In this way, we see that countries such as France, Germany, Italy, search more for foreign slang than they are searched. Jamaica seems to search less than it is searched. This does not mean that overall countries search Jamaican slang more than French, as the figure is normalized.

Overall, we notice how the US is the country with the most searched words. However, it is also the country with more words in the dictionary. For this reason, we decide to further restrict the exploratory analysis to the 20 most famous words of each country. In this way, given the same number of words, we can observe which country has witnessed more external google searches. Results are shown in Figure 4.



Chord Diagram of the influence of the most 20 famous words per country

Normalized Chord Diagram of the influence of the most 20 famous words per country

Figure 4: The interpretation of the graphs is the same as the one provided in Figure 3. We notice how the words that are searched the most abroad belong, as expected, to English speaking countries. In the left graph we see that most of the countries receive more influence, in terms of terms they search on Google, than the one they transmit.

5.2 Hypothesis Testing

The first hypothesis that we test is: *There has been a significant increase towards American slang imitation, that has not been witnessed instead with other countries slang.* If we make the possibly bold assumption that every word that is searched is then also very likely to be imitated, we can then consider high searches for a term synonym of high imitation. Hence, we proceed by calculating the correlation between `GT Influence Value` and `Date`. We then carry out a permutation test for each country. We expect countries with more observations to produce distributions more concentrated around 0. Countries with few observations, even though we will be able to calculate the correlation for every single labelling in the permutation test, will be cursed by high variance, and we expect their results to be less certain. With that said, we present the results in the plot below.

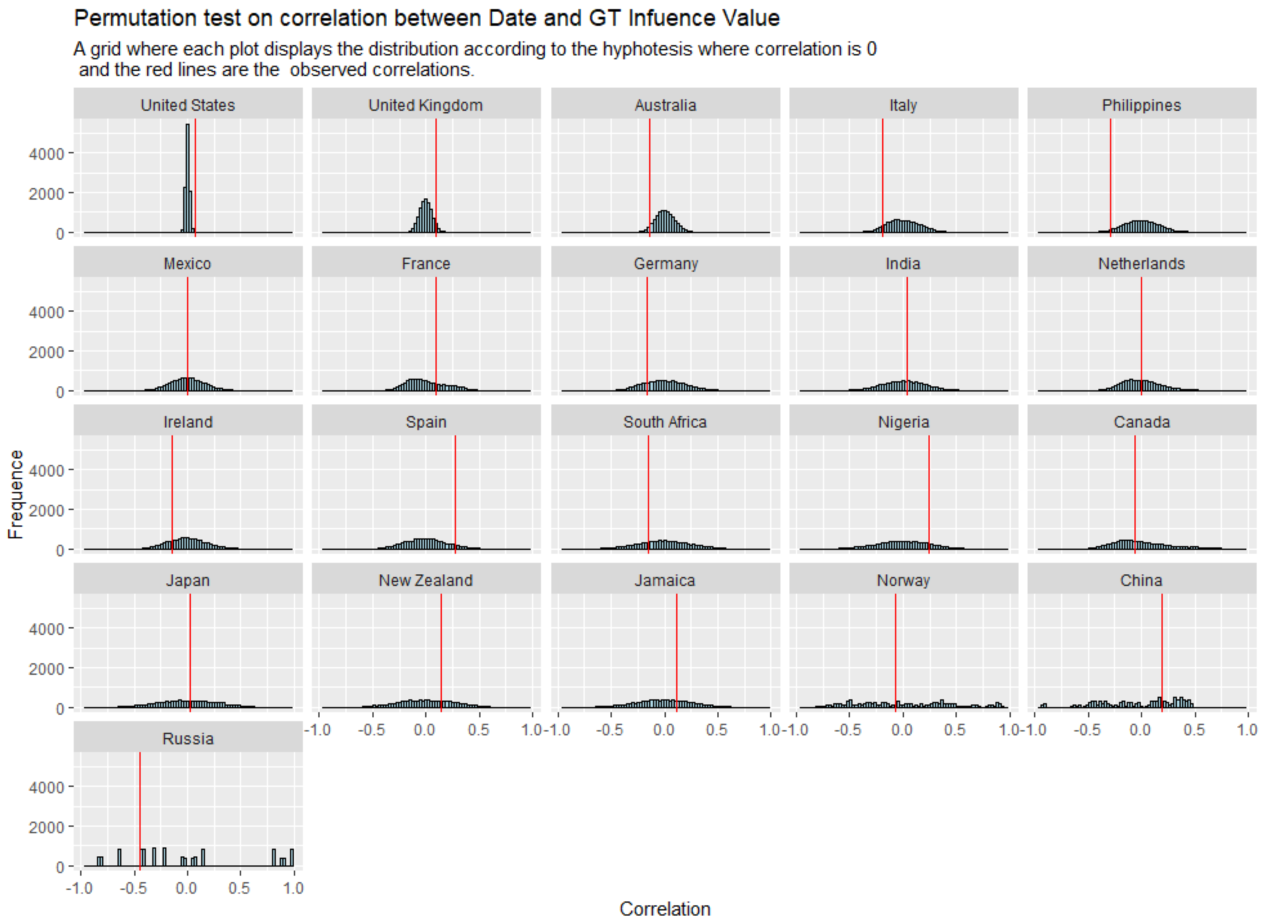


Figure 5: Permutation test per country. Countries are ordered from the one with most words to the one with less words.

Here below the p-values for the hypothesis that the correlation has increased for each country.

Table 1: Table showing observed correlation and pvalue for each country.

	United States	United Kingdom	Australia	Italy	Philippines	Mexico	France
Correlation	0.056	0.09	-0.14	-0.194	-0.295	-0.005	0.098
pval >	0.0002	0.06	0.95	0.90	0.97	0.501	0.282
	Germany	India	Netherlands	Ireland	Spain	South Africa	Nigeria
Correlation	-0.166	0.0328	-0.006	-0.14	0.274	-0.148	0.245
pval >	0.80	0.44	0.48	0.77	0.077	0.72	0.16
	Canada	Japan	New Zealand	Jamaica	Norway	China	Russia
Correlation	-0.064	0.022	0.145	0.11	-0.073	0.185	-0.44
pval >	0.529	0.47	0.30	0.33	0.53	0.39	0.83

We notice that the slightly positive correlation that the United States experiences appears to be the only significant one at level 0.05. For all the other countries, we cannot reject the null hypothesis, as all of them are not significant considering both the unilateral (>) rejection region and the bilateral one, hence the correlation we observe could be due to a random effect. Focusing at last on the United States case, we carry a final analysis to better investigate the relationship. We group the data in words submitted on Urban Dictionary before and after the year 2011 (01.01.2011). The groups have approximately 2400 and 1400 values. The average searches for each group are 93 and 121. The means appear different and their difference results statistically different from 0 at level 0.05 - which may not surprising giving that the variance of the mean estimator shrinks significantly with large n. We cannot assess whether this difference is given due to a cultural trend or to Urban Dictionary related dynamics i.g. becoming less popular and having only very famous words defined for later years.

5.3 Conclusions

To conclude, we noticed that the US was the country with overall more words google-searched from abroad. Specifically, almost every country seemed to have the United States as the primary source of influence. With that said, it is hard to draw conclusions in this sense as Urban Dictionary is a source that is more representative of American jargon than it is of other countries slang. In fact, we then focused on solely the 20 most popular words for each country, and meanwhile the US was still popular among other countries, also Australia and the United Kingdom seemed to be as much. Also, on the normalized values, these said countries seemed to be the only ones to send out more influence than the one they received. Hence, we could argue that are generally the English speaking countries the ones whose slang are searched on Google the most. Indeed, this conclusion would be plausible, given the inter-nationality of the English language. Given that these slang words are reflective of the cultures they originated in, we could argue that English speaking countries have a privileged highway in transmitting their culture abroad through slang. On the hypothesis testing, the observed correlations were not significantly different from 0 for most of the countries. That is, in the years from 2004 to 2021, it seemed that the way words have been searched abroad has not changed. However, we remind that this does not mean that the magnitude of words' search has stayed the same: the search values on which we based our study are in fact relative quantities. Hence, this result could pave the way for the hypothesis that the magnitude of searches has grown uniformly in all countries during the last 15 years. In this sense, older and younger words may appear with the same score on Google Trend because the volume of their searches has grown everywhere proportionally. The only country whose correlation was significantly greater than 0 is the United States. However, the observed correlation of 0.056 presents only a very weak linear relation between the two variables, and it's hard to draw conclusions. Moreover, even if there was a linear correlation, still this would not imply a causal dynamic underneath. Finally, regarding the significance of the t-test, this may relate to a slight increase of searches towards American slang, but as pointed in the previous section, the causes underneath may as well be others.

5.4 Critique

The following points present the main weaknesses about the work and their resolution could be a focus in future work. First of all, we selected the slang words from Urban Dictionary selecting word definitions containing the word `slang`. We hence ruled out all the slang that did not contain the word `slang` in it. From 5 million words, we ended up with roughly 40 000. An example is the slang `OK Becky`, that had been quite popular in the first decade of the 2000, but that was not considered in the analysis. Secondly, in the definition of a word, adjectives like `Spanish`, that refers both to the language (spoken in many countries) and the country, may have caused disambiguation. We grouped all the terms that had `Spanish` in it as "belonging to Spain", even though some of them may have meant Spanish as the adjective referring to the language rather than the place - that could have been in Latin America, for example. This may explain the lack of presence of countries like Argentina, Venezuela, etc. As a third point, the boldest assumption that we make is possibly stating that the imitation of a word is a direct effect of the look up of a word on Google. Indeed, the use of other on-line platform such as Twitter could be very useful in detecting the actual use of these words. On a statistical level, the concurrent tests are more than 20, hence multiple testing adjustments could have been carried out, given the observed alpha level being 0.05. For future work, it would be also be interesting to include the magnitude of searches in the analysis and to calculate if there is some sort of homophily among countries that are physically close to each other, how we can see for example between the US and Canada om Figure 3.

Notes

[1] www.urbandictionary.com

[2] https://github.com/samuelstevens/ubscrape

[3] https://docs.python.org/3/library/urllib.html

References

Kipping, . T., M. (2002). *The 'americanisation' of european companies, consumers and cultures: contents, processes and outcomes.* Lille: Publications de l'Institut de recherches historiques du Septentrion. Retrieved from http://books.openedition.org/irhis/1935

McLuhan, M. (1962). *The gutenberg galaxy: The making of typographic man.* University of Toronto Press. Retrieved from https://books.google.ch/books?id=y4C644zHCWgC

Zhou, Y., & Fan, Y. (2013). A sociolinguistic study of american slang. *Theory and practice in language studies*, *3*(12), 2209.