Gunita: Visualizing the evolution of the Philippines' languages with a historical Philippine text corpora

Amiel Bornales
De La Salle University-Manila
2401 Taft Ave, Manila
Philippines
amiel_bornales@dlsu.edu.ph

Jonn Yuvallos
De La Salle University-Manila
2401 Taft Ave, Manila
Philippines
jonn_yuvallos@dlsu.edu.ph

Courtney Ngo
De La Salle University-Manila
2401 Taft Ave, Manila
Philippines
courtney.ngo@dlsu.edu.ph

Abstract

While there are many culturomic studies in other countries, only a few studies focus on the culturomics unique to the Philippines. This study developed a Philippine news sources scraper and used a pre-existing Tagalog corpora containing books and poems across 100 years to build a continuously growing corpus. This study introduces Gunita, a web application that allows users to visualize how an n-gram is used over time and know which article, book, or poem the n-gram is used in to shed light on how Filipinos communicate through written text.

1 Introduction

Culture is an ever-evolving aspect of society. It provides each person their own cultural identities and heritages, and is an essential area of human society that is ripe with research. This field of research is known for its cultural studies, and it aims to explore the connections between gender, race, class, etc. and their effects on culture (Barker, 2003). While cultural studies are able to draw powerful conclusions on human culture, these conclusions are based off of a collection of carefully chosen works that represent only a minority of the media available at the time (Michel et al., 2011). Analysis of cultural trends has always been hampered by the lack of suitable data, and so, to help further research along, a corpus containing 5,195,769 digitized books was created and analyzed (Michel et al., 2011). This led to the creation of the field of culturomics, which is a form of computational lexicology that studies human behavior and cultural trends through the quantitative analysis of digitized texts.

2 Related Works

The analysis of language and words is an important aspect of culturomics, and large text corpora has been a necessity for language analysis for many of these studies. However, these researches differ in how they handle their data. For example, Michel et al. (2011) used a large corpus to investigate cultural trends between 1800 and 2000, such as insights into lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. Given that their scope is extremely broad, they focused more on the general state of things, and only go into specifics for important figures or events. For example, it was shown that Tiananmen (written in Chinese characters) was barely mentioned while the English term "Tiananmen" was mentioned increasingly in this period. This shows that the Chinese term of "Tiananmen" was not heavily used in China after the massacre due to censorship, while the English term of "Tiananmen" was mentioned frequently in

Western media during the aftermath. This research will adopt the same focus and analyze the data based on news and historical events.

According to Leetaru (2011), it is possible to analyze cultural trends even further with culturomics because it seeks to provide insights and encourage explorations on broad cultural trends. This is done by analyzing vast digital corpora computationally because the extremely large amount of textual data required for a proper corpus is impossible to be manually read by a human (Michel et al., 2011). Once the data has been processed, the data is reflective of the time the data came from, which gives us a snapshot of how the information environment was back then (Leetaru, 2011).

Ilao et al. (2011) performed culturomic analysis on online tabloid articles and was able to provide insights on various aspects of Filipino culture, like the Filipino Christmas tradition, the Filipino outlook on extreme calamities, and linguistic trends. With that in mind, this study also aims to study the cultural evolution of the Philippines, with a strong emphasis on the evolution of a few of the languages in the Philippines. These languages have changed significantly over the years, in terms of spelling, vocabulary, orthography, and grammar. In the case of Filipino, this is due to a variety of reasons, such as transitioning to a modern spelling system from Spanish-influenced spelling system, the release of the Grammar of the National Language, and the language wars that occurred when the Philippines was deciding on its national language (Ilao and Guevara, 2012). However, it is difficult to accurately trace the history of the other languages in the Philippines such as Cebuano and Ilocano due to the lack of available historical resources.

3 Methodology

The structure of the data input flow for this study is illustrated in Figure 1. It is divided into three stages: data collection, data processing, and data storage. After the data has been stored, the Gunita system can be used to gain insights on specific n-grams using the three features: n-gram usage visualization (line and frequency chart), wildcard search (simplified regular expressions), and co-occurrence search.

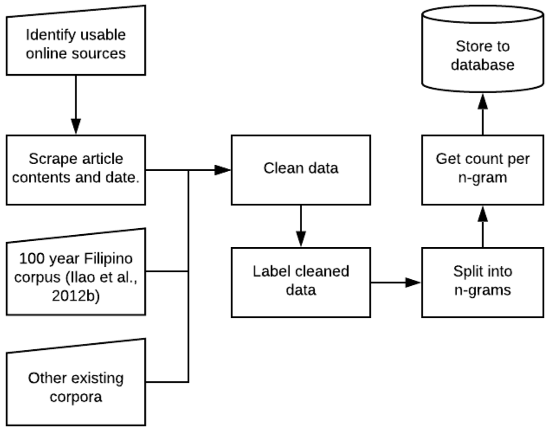

Identify usable online sources
Scrape article contents and date.
100 year Filipino corpus (Ilao et al., 2012b)
Other existing corpora
Clean data
Label cleaned data
Split into n-grams
Get count per n-gram
Store to database

Figure 1. The data input flow this project comes from a Philippine news source scraper and from the text corpus of Ilao and Guevara (2012). The raw HTML data from the scraper is stripped off HTML tags and other characters, and the whole article is labelled a language using FastText. The result is split into n-grams, underwent n-gram counting, and stored into the database.

3.1 Data Collection

The data will come from two sources. The first source for the data is scraped from online news articles and websites, and the second source came from the historical Tagalog corpus by Ilao and Guevara (2012).

Before any scraping is done on a target source website, it must have a directory containing all articles, and the article must have a publishing date. If a website passes this criteria, Python's request library is used to send requests to the web server’s article directory and receive the HTML response. Then, the article links from the raw HTML data is retrieved with the help of a Python library called BeautifulSoup, which allows the querying of the article links using an HTML parser. Lastly, all the article links are saved in CSV file, with each target source having their own CSV file to avoid any confusion and accidental mixing of data. Once all the links have been collected, the crawler visits each article link and scrapes all its article text content. The text contents are also added as a separate column in its corresponding CSV file.

3.2 Language Labeling

To label each article by its language, the language classifier from FastText was used. However, since the model classifies 176 different languages, there are cases where the article that was written in one of the four given languages

(Tagalog, Cebuano, Ilocano, English) is labelled as some other language. To fix this, the researchers extracts the language probabilities returned by the model. A sample result would be ['english' : .90, 'tagalog' : .40, 'cebuano': .20], which means that the probability of the text being english is 90%, tagalog 40%, and cebuano 20%. The article will then be classified as the language that has the highest probability given that it is either Tagalog, Cebuano, Ilocano, or English.

3.3 Data Processing

From the raw gathered article data, BeautifulSoup's built-in HTML parser was used to remove the HTML tags, and Python's built-in regular expression (RegEx) library was used to remove unneccessary special symbols. The RegEx used performs two operations: the first is to remove any periods or commas that are not used in numbers, and second is to remove any quotation marks so there would be no issues when exporting the cleaned data into a CSV file. To determine whether the period or comma is used in a number, the RegEx library checks if its adjacent characters are numbers, for example “1,234” or “1.23m” will retain its period and commas, however the string “cat.dog” will have the period removed since it does not meet the requirement of having 2 adjacent numbers. The usage of capture groups in RegEx will be used in retaining the characters adjacent to the removed period or comma.

After cleaning the data, the words will be separated into n-grams, where n will range from 1 to 3. To create and count the n-grams, the CountVectorizer from the Python library sklearn was used.

3.4 Data Storage

Once the data has been cleaned it is stored into the database. The entity relationship diagram of the database is shown in Figure 2.

Database Tables: The ngram_occurrences table contains an auto incremental ngram_occ_id as the primary key, ngram_id as the foreign key referencing the ngram in the ngram master list, n_number as the size of the n-gram (1, 2, or 3), ngram as the n-gram itself, date as the date the n-gram was published, lng_src is the language source of the n-gram (1 for Cebuano, 2 for Tagalog, and 3 for Ilocano), date as the date the data was taken from, n_number as the total amount of occurrences the ngram has on that date, source_id and source_link are foreign keys that reference the source of the ngram.

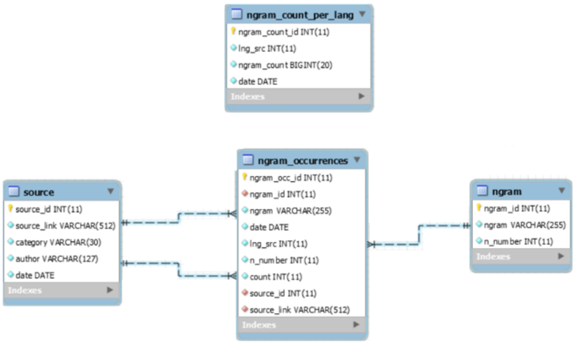


Figure 2. The database has a table for the source URIs, n-grams, n-gram counts for each language, and the n-gram occurrences which includes the URI the word was mentioned, the date the article was published, and how many times the word was mentioned in that article.

The database is also indexed on multiple columns, the index is ngram_occurrences_idx_ngram_src_number_date. This multi-column index contains the ngram, lng_src, n_number, and date as its indexes. This multi-column index was added to the database because a query can only make use of one index at a time, so instead of having multiple single column indexes that can only check one column, a multi-column index was used instead. Indexes are important because they make queries more efficient and run faster. These indexes improve query performance by providing the database the position of the relevant data in the table, making the database jump ahead to that position without having to scan all the other rows in the table while looking for relevant data. This will allow faster querying for the ngram, lng_src, n_number, and date columns since MySQL will know the locations of the data inside the table.

The ngram table contains an auto incremental ngram_id as the primary key, ngram as the ngram, and n_ number being the size of the n-gram (1, 2, or 3). None of these fields can be nulled as they are all used for RegEx querying. The ngram table is indexed on n_number to speed up the queries.

The details (title, author, date, and link) of all the scraped articles are also stored in another table. This table contains the source_id as the auto incremental primary key. The link must be present because this is used to ensure that the same article isn't scraped again. The category of the source must also be present to let the user know what kind of medium the data is from. The author must be present to let the user know who wrote the piece. Lastly, the date must be present because the source cannot appear in the visualization graphs without it.

Lastly, the monthly statistics of the corpora is stored as well. The ngram_count_per_lang contains an auto incremental ngram_count_id as the primary key, lng_src signifies which language the data came from, ngram_count represents the sum of all ngram counts for that month, and the date serves as the date where the data came from. This table is indexed on the multi-column index of date_lng_src_index, this index utilizes the columns date, and lng_src to speed up the queries.

3.5 Data Visualization

For the visualizations, the researchers used Chart.js, a data visualization library in JavaScript.

Line Charts: Line charts are used to visualize the change of a value over time. In the example shown in Figure 3, the chart shows the size of the English lexicon over time from Michel et al. (2011). Line charts will be used to visualize the word count over time.

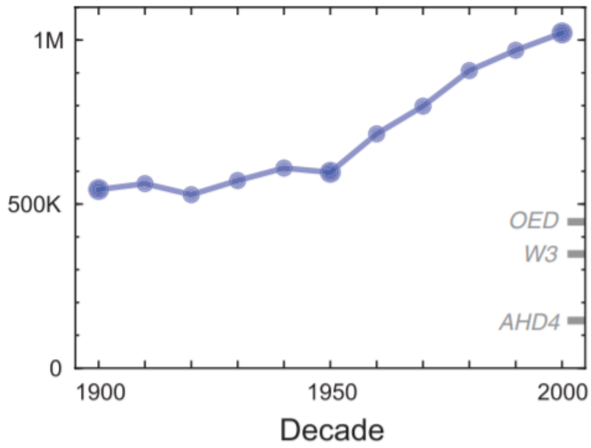


Figure 3. Line chart of the English lexicon and the coverage of the words in three dictionaries. Image taken from Michel et al. (2011).

Frequency Charts: Frequency charts is a variety of a line chart that uses a frequency of a value over time instead of value over time. The chart in Figure 4 shows the usage frequency of the word "slavery" over time, the frequency value is computed by the word count for that year divided by the total amount of words in the corpus for that year (Michel et al., 2011). This chart will be used to find the usage of the word over time, as well as compare it with the usage frequency of other words.

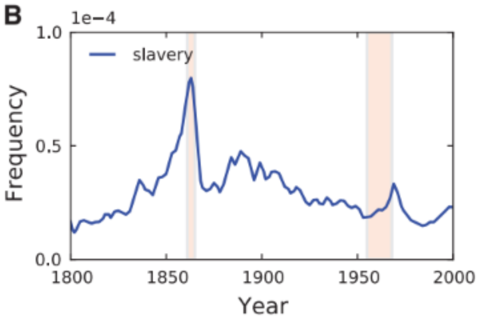


Figure 4. Frequency chart showing the word frequency of the word 'slavery' with red marks highlighting the events of the Civil War and the Civil Rights Movement. Image taken from Michel et al. (2011).

3.6 User Testing

For the user testing, individuals who have backgrounds in linguistics or language analysis were selected, as most of the features in the system are tailored for their use. The testing procedure is split into 4 parts: a background interview of the user, an introduction to the system, a demonstration on how to use the system, and an exit interview.

The background interview consisted of questions regarding their experience in the field, as well as inquiries on how they would normally do analysis on certain words or languages, as well as questions on what technologies or existing systems they use to aid in their analysis.

The system was then introduced to them, and they were given the chance to experiment with the features as well as accomplish certain tasks that are given to them. After the demo phase, a final interview is conducted to gather feedback and suggestions regarding the system.

4 Gunita Features and Interesting Searches

4.1 N-gram Search

This feature allows the user to query for an n-gram from the database and see the n-gram's usage over time through a line chart or a frequency chart.

Pasko vs Christmas: Figure 5 shows the yearly Christmas trend in the Philippines. Instead of having "Christmas" and "Pasko" occur only in December, it can be seen that the Christmas season starts around September and peaks in December and quickly falling off in January. This graph agrees with the findings of Ilao et al. (2011) regarding the Filipino Christmas season. The users can interact with the chart to know the exact number of n-gram occurrences in the timeline. When they click on a specific time period in the timeline, they can also see all the sources where the n-gram was mention along with a link to the news source if necessary. This feature provides the user context on why the n-gram gained or lose popularity.

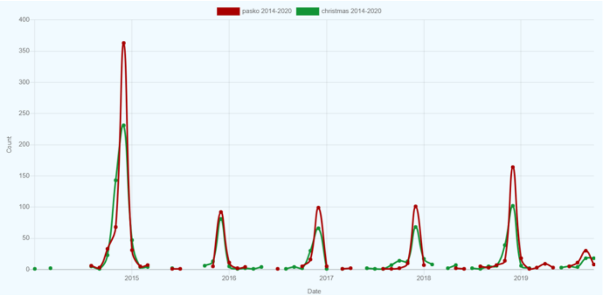


Figure 5. Graph showing the occurrences of "Pasko" (in red) vs "Christmas" (in green) in the collected Philippine news text.

Pilipinas vs Filipinas: Figure 6 shows that "Pilipinas" is more widely used compared to "Filipinas". Despite efforts from the Komisyon sa Wikang Filipino (2013) pushing for the use of the official term "Filipinas", it can be seen that "Pilipinas" is the more preferred version of the word, however the data is primarily from news articles and so, it only shows the more preferred version in news articles. It is also important to note the the recent usage of the term "Filipinas" are present because of news articles reporting on the Komisyon sa Wikang Filipino (2013) proposal to change the Philippine's official Filipino term from "Pilipinas" to "Filipinas".

West Philippine Sea vs South China Sea: Figure 7 shows that the usage of "South China Sea" is generally more prevalent before 2019, while the usage of "West Philippine Sea" is used more after 2019. This could be due to rising political tensions over this region.

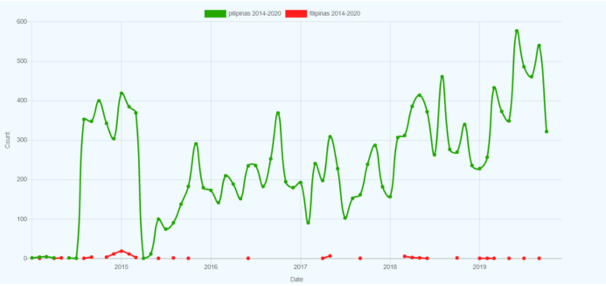


Figure 6. Graph showing the occurrences of "Pilipinas"(in green) vs "Filipinas" (in red). The graph here uses scraped Philippine news text and the text corpus collected by Ilao and Guevara (2012).

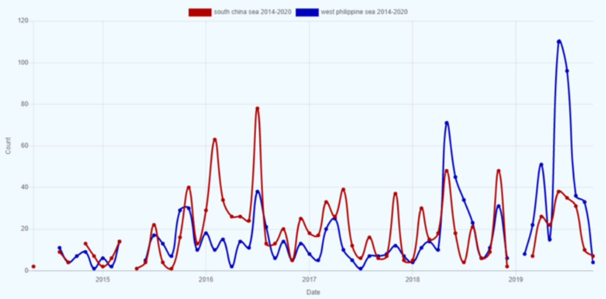
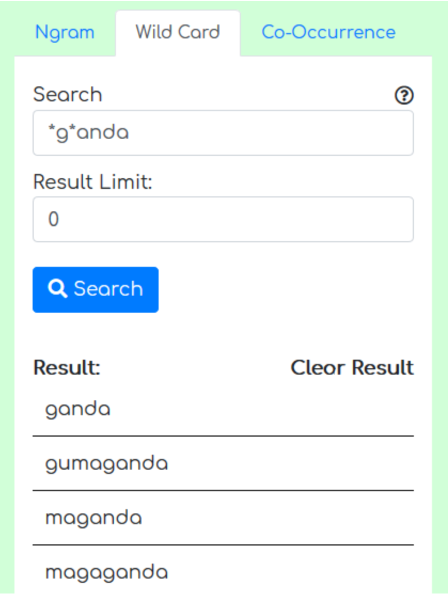


Figure 7. Graph showing the occurrences of "West Philippine Sea" (in blue) vs "South China Sea"(in red) in the collected Philippines new sources.

4.2 Wild card search

The wild card search function allows users to search for the derivations of root words. With the use of a simplified version of regular expressions, users can use a combination of 3 symbols: "*", "+", and "?". The "*" symbol can be used to represent 0 or more characters, the +" symbol can be used to represent 0 or 1 characters, and the "?" can be used to represent any single character.

In the figure below, the word "ganda" is used to serve as an example. The result limit is set to 0 to signify that the user wants to list all n-grams that match the Wild Card search term.



Search

*g*anda

Result Limit:

0

Result: Clear Result

ganda

gumaganda

maganda

magaganda

Figure 8. Wild card results for "*g*anda" showing the derivations of the word "ganda".

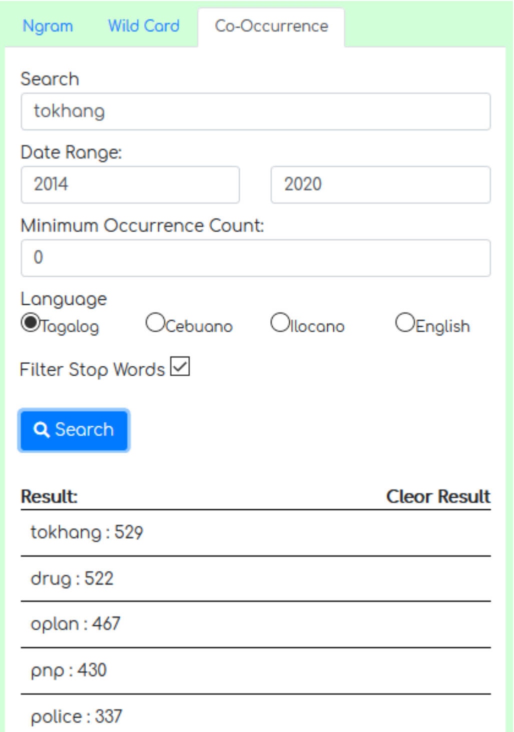
Figure 8 shows that the input *g*anda was able to capture many derivative forms of the root word ganda, although there are some words that aren't derivative forms of "ganda", it is still acceptable. This feature is important since the Filipino language has many variations where pre-, in-, and post-fixes are place in placed in different parts of the derivation.

4.3 Co-occurrence Search

The co-occurrence search allows the user to search for the words that most commonly occur with the given n-gram.

The n-gram "tokhang" will be used to serve as an example. The minimum co-occurrence count can be set to limit the results; it is currently set to 0 to list all co-occurrences regardless of how many times the co-occurrence appeared in the corpus. The user can also select the source language and has the option to filter stop words as well.

The co-occurrence results in Figure 9 show that the word "tokhang" is very commonly used along with the "pnp", the Philippine National Police. While there are many words that aren't related to the word "tokhang" but are instead just extremely common words in the Filipino languages. This search shows that the word "tokhang" is commonly used to refer to the Philippine National Police war on drugs.



Search

tokhang

Date Range:

2014 2020

Minimum Occurrence Count:

0

Language

● Tagalog ○ Cebuano ○ Ilocano ○ English

Filter Stop Words ☑

Result: Clear Result

tokhang : 529

drug : 522

oplan : 467

pnp : 430

police : 337

Figure 9. The co-occurrence search results for "tokhang" show that "drug", "oplan", "pnp", and "police" are commonly used with the word "tokhang".

5 User Testing

Three experts from the fields of language education, Philippine studiess, and Filipino language tried Gunita and checked what features they can use for their respective research works.

The general feedback from all users is that they had a positive experience with the system since they liked the majority of the features that were shown to them. The most common feedback was that the visualizations were nice and was also very insightful as it highlighted the increase and decrease of the usage of a certain word over time. They also liked the feature of viewing the sources of the points in the graph, as it provided the cause for the increase or decrease as well as a way for the user to look deeper as to why the increase or decrease happened.

Although the users were initially surprised at the large amount of data that Gunita contains, a large portion of it was still lacking especially on dates before 2014 and the scarcity became more evident as the users searched for more specific n-grams. The users also suggested that the system should cover more categories of textual data such as short stories and blog posts and should not be limited only to news articles. They mentioned that including these resources does not only provide more data, but also gives more forms of writing since news articles normally use a formal way of writing as opposed to textual data from social media sites and blog stories. The users also suggested a Parts Of Speech (POS) Tagger on the n-grams to get rid of specific parts of speeches in the co-occurrence results, as well as provide a more specific visualization to an n-gram depending on the context of how it was used. The usage of the RegEx search also confused the users as it required them to know how to use RegEx patterns.

The users all agreed that the system can provide a lot of research potential, and that it would help them in their own research.

6 Conclusion and Future Works

The general objective of visualizing the evolution of written texts in the Philippines was achieved by developing a system that generates the said visualization given certain search parameters. The requirements of the system were determined by reviewing related systems that currently exist, Gunita was then tested by experts to verify the requirements as well as gather feedback on what features could be added or improved upon on the system. After reviewing multiple systems and finishing the development of our three main features (n-gram search, wild card search, co-occurrence search) the general consensus among the experts interviewed was that the system had many applicable uses in the field of linguistic research.

However, there is much room for improvement. The wild card search can be quite limited due to the simple implementation of RegEx. A balance between user accessibility for users unfamiliar with RegEx and thorough searching with complex RegEx must be found.

The system is also extremely biased to recent news articles from 2014-2019, and is not very accurate for mediums of literature outside of news articles since the tone of news articles are formal. For example, trying to visualize and gain insights on particular slang and informal words may not work due to data sparsity. The system would benefit from more sources of data such as WattPad and Twitter.

The system could implement Parts Of Speech (POS) tagging in the future as this would provide better context when visualizing the usage of some n-grams as well as provide better results when searching for related n-grams with Wild Card or Co-occurrence search.

The languages can also be labelled better. In its current state, the system labels whole articles according to its predominant language, so if a Filipino article quotes an English sentence, that English sentence will be classified as Filipino. A solution to this problem would be to implement sentence-level language labelling, this can be done by splitting the contents of the article on the punctuation marks, and then labelling the language per sentence instead of per article.

Acknowledgments

The authors want to thank Dr. Joel Ilao for their generosity in sharing their historical Filipino corpus, and Dr. Charibeth Cheng for their valuable insights on this project. The authors also acknowledge Dr. Shirley Dita, Dr. Rhoderick Nuncio, and Dr. Raquel Sison-Buban for lending their expertise to the researchers on how to improve Gunita.

References

Komisyon sa Wikang Filipino. (2013). Pinagtitibay ang kapasiyahan ng pagbabalik ng gamit “filipinas” habang pinipig ang paggamit ng “pilipinas”.

Michel, J.B., Shen, Y., Aiden, A., Veres, A., Gray, M., Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., & others (2011). Quantitative analysis of culture using millions of digitized books. science, 331(6014), 176–182.

Leetaru, K. (2011). Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. First Monday, 16(9).

Ilao, J., Guevara, R., Llenaresas, V., Narvaez, E., & Peregrino, J. (2011). Bantay-Wika: towards a better understanding of the dynamics of Filipino culture and linguistic change. In Proceedings of the 9th Workshop on Asian Language Resources (pp. 10–17).

Ilao, J., & Guevara, R. (2012). Investigating spelling variants and conventionalization rates in the Philippine national language’s system of orthography using a Philippine historical text corpus. Proc. of O-COCOSDA.

Barker, C. (2003). Cultural studies: Theory and practice. Sage.