Bantay-Wika: towards a better understanding of the dynamics of Filipino culture and linguistic change

Joel P. Ilao[*]
Rowena Cristina L. Guevara[†]
Digital Signal Processing Laboratory
University of the Philippines - Diliman
Tel: +632-981-8500 local 3370
[*]joel.ilao@up.edu.ph
[†]gev@eee.upd.edu.ph

Virgilio D. Llenaresas
Eilene Antoinette G. Narvaez
Jovy M. Peregrino
Sentro ng Wikang Filipino
University of the Philippines - Diliman
Tel: +632-981-8500 local 4583
upswfdiliman@gmail.com

Abstract

The Bantay-Wika *(Language Watch)* project was started in 1994 by the University of the Philippines (UP) - Sentro ng Wikang Filipino[1] (SWF) in order to track for long periods of time how the Philippine national language is being used and how it develops, particularly in the Philippine media. The first phase of this project, from 1994 to 2004, involved the manual collection and tallying of frequency counts for all the words in eleven major Philippine tablods. With increasing online presence of Philippine news organizations, the project was revived in March 2010, with UP-SWF partnering with UP - Digital Signal Processing (DSP) laboratory. The project objectives were also re-drafted to include the development of software that would automate the process of downloading of Filipino news articles. In this paper, we further detail the goals and the history and of the Bantay-Wika project, its accomplishments and plans for future work. The project ultimately endeavors to build a computational model for language development that can guide language policy makers in a multi-lingual country such as the Philippines, in drafting policies that can effectively promote the use and development of their national language.

1 Introduction

The Philippines is an archipelago of 7,107 islands with 171 living languages spoken by 94 million inhabitants[2], thus making it the 25th most linguistically diverse nation in the world among 224 nations in the 16th edition Ethnologue (2009) (Lewis, 2009). Based on the 2000 census conducted by the National Statistics Office (NSO), there are 14 major languages[3] spoken in the country, listed next in order of decreasing number of speakers: (1) Tagalog, (2) Cebuano Bisayan, (3) Ilokano, (4) Hiligaynon Bisayan, (5) Waray (Eastern Bisayan), (6) Kapampangan, (7) Northern Bicol, (8) Chavacano, (9) Pangasinense, (10) Southern Bicol, (11) Maranao, (12) Maguindanao, (13) Kinaray-a, and (14) Tausug (Roxas et al., 2009). With such a rich Philippine linguistic profile, it is not hard to understand how difficult it must have been when in 1935, then president Manuel L. Quezon pushed for the establishment and development of language to be commonly spoken by all of the inhabitants. In the years that followed, laws have been enacted to support such cause, leading to the establishment of a national language institute tasked to identify and select the basis of the national language from the list of Philippine native languages, and to further develop the national language into a modernized and intellectualized one. Table 1 lists a brief timeline showing some milestones related to the development of the Philippine national language.

It is worth noting that the issues concerning the Philippine national language have initially been controversial and divisive, especially in the so-called language wars of the 1960's, when Senate representatives who speak the Cebuano Bisayan language challenged the selection of Tagalog language as the basis of the national language (Gonzalez, 2003). During those times until now, efforts were made to differentiate the national language from its original Tagalog base, through conscious revisions of the grammar and orthography rules

[1] 'Sentro ng Wikang Filipino' means 'Filipino Language Center'

[2] Projected population of Filipinos for 2010 according to the National Statistics Office of the Republic of the Philippines (http:www.census.gov.ph)

[3] A major language is a language spoken by at least 1 million speakers.

Proceedings of the 9th Workshop on Asian Language Resources, pages 10–17,
Chiang Mai, Thailand, November 12 and 13, 2011.

Year	Milestone Description
1935	Concept of a language commonly spoken by all inhabitants was mandated by the constitution
1936	The Tagalog language was selected as the basis of the national language
1959	The National language was named as *Pilipino* through a Department of Education order
1987	The national language was officially named as *Filipino* through the constitution

Table 1: Philippine national language development milestones.

to incorporate elements from other Philippine languages (Santos, 1940; Surian ng Wikang Pambansa [SWP], 1976; SWP, 1987; Komisyon sa Wikang Filipino [KWF], 2001; UP Sentro ng Wikang Filipino [UPSWF], 2004; UPSWF, 2008; KWF, 2009), and by renaming the national language, first as *Pilipino* in 1959, then as *Filipino* in 1987[4] in order to garb it with more sense of national appeal. Almost a quarter-century after the idea of a national language was first floated, linguists and language planners are wondering how far the Filipino language has developed and where it is headed, knowing that in a multi-lingual society such as the Philippines, it is possible that a language may retrogress and die despite sincere efforts to develop and propagate its proper use to the rest of the population (Abrams and Strogatz, 2003). In this light, effective monitoring activities of the extent and quality of actual usage of the national language by its community of users is deemed important.

2 The Bantay-Wika Project

Cognizant of the importance of the national language in fostering national unity and identity, the University of the Philippines felt it is duty-bound to actively participate in the discussions on language development. Hence, the UP - Sentro ng Wikang Filipino was created in 1990. Realizing the value of quantitavely observing the dynamics of the Filipino language usage in ascertaining the trajectory of Filipino language development, the UP-SWF started the *Bantay-Wika project* in 1994. This project initially sought to observe the rates with which new words were being introduced in the Filipino working vocabulary, as evidenced in the tabloid media. Filipino-written articles from eleven tabloids[5] with wide circulation were transcribed on a weekly basis, and a database of word frequency counts was manually built.

2.1 Tracking of competing word forms

Ferguson (1968) noted that as a language develops, it undergoes a process of standardisation by means of conventionalization of its function and use within a given language community . Guided by this language developmental framework, the Bantay-Wika project also seeks to observe how the Filipino language conventionalizes, by comparing usage rates of competing word forms. One particular class of competing word forms actually tracked in the project are borrowed words from other languages, mainly from Spanish and American English languages, owing to the influence that Spain and USA have on the Philippine society as its long-time settlers. A cursory look into Filipino-written articles would show that many of the terms used in daily news reportage which have equivalent translations in either English or Spanish manifest in any of four possible ways: (1) direct borrowing, (2) calquing or loan-translation, (3) Filipinization of spelling, and (4) use of a native word. Table 2 shows examples of each of the four ways of improvising Filipino words with foreign equivalents.

Other classes of competing word forms are spelling variants that fall under the cases listed in Table 3. Towards the end of the first phase of Bantay-Wika (year 2001 to 2004), software that can process the gathered text corpora and produce summarized reports of word counts was developed. The reports are then manually inspected to identify spelling variants. This process resulted in over 957 words tracked, of which 85 pairs were identified as spelling variants.

2.2 Philippine Culturomic Analysis based on Filipino - written news articles

The topics that a group of people write about reflect their prevailing sentiments at any given time

[4] The official name of the Philippine national language is *Filipino*.

[5] The tabloids included in the project are *Abante Tonite, Abante, Bomba, Bulgar, Imbestigador, Kabayan, People's Balita, People's Taliba, Pilipino Star Ngayon, Remate,* and *Saksi Ngayon*

Case Description	**Example**
Removal of /u/ in /uw/	kuwento / kwento
Removal of /i/ or /y/ in /iy/	superstisiyon / supertisyon Dios / Diyos
Repeating consecutive similar vowels simplified to a single vowel	saan / san mayroon / mayron
Removal of h-	hospital / ospital
Removal of prefix i-	ipinapakita / pinapakita
Simplification from /ng/ to /n/	pangsarili / pansarili
Simplification from nakapag- to naka-	nakapagsaing / nakasaing
Removal of infix -i-	maitatanggi / matatanggi
/tion/ vs. /siyon/ vs. /syon/ vs. /sion/	intervention / intervensiyon / intervensyon / interbensiyon / interbensyon
/i/ vs. /e/	galing / galeng
/o/ vs. /u/	kumpanya / kompanya
/c/ vs /k/	Cristo / Kristo
/s/ vs /z/	Luson / Luzon
/p/ vs. /f/	Pilipinas / Filipinas
/b/ vs. /v/	unibersidad / universidad
Miscellaneous cases	Maynila / Manila manyika / manika manga / mga

Table 3: Cases of Spelling Variants seen in Filipino texts

Technique	**Description**	**Example**
Direct Borrowing	verbatim use of foreign word in the sentence	shooting trophy mesa
Calquing	word-for-word translation to Filipino	tigil-putukan (*ceasefire*)
Filipinization of spelling	respelling of terms according to the Filipino alphabet	girlfrend gerlprend (*girlfriend*)
native word	use of a word found in any of the Phil. languages	bahay (*house*) mata (*eye*)

Table 2: Different ways of improvising Filipino terms with foreign equivalents

frame. Hence, it is possible to study cultural trends based on quantitative analysis of large corpora of digitized text, a field of study called Culturomics. Using normalized frequency plots of specific search words and phrases, Michel et al. (2011) were able to investigate cultural trends in fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorhip, and historical epidemiology. Inspired by these recent developments in Culturomics, we wanted to gain more insights into the dynamics of topics that interest Filipinos, using news corpora curated from websites of two popular Philippine tabloids. Articles were downloaded using a web-crawler from the websites of Abante[6] and Abante-tonite[7] within a two-year window period from March 22, 2009 to May 16, 2011. Each downloaded html-file was then automatically date-stamped by parsing its URL for date information, and then preprocessed by removing markup tags and scripts.

[6] http://abante.com.ph

[7] http://www.abante-tonite.com

After which, the resulting text-files were each word-tokenized and content-modified by lining up sentences one after the other. Metadata for the unannotated Abante and Abante-Tonite text corpora include URL, publication date and download date information. Table 4 describes the sizes of the Abante and Abante-tonite news text corpora. Figure 1 shows a graph of the number of downloaded articles per day for the entire observation period. Note the gaps in the dates of downloaded data for both Abante and Abante-tonite, indicating that not all dates have available downloadable data from the corresponding websites within the 2-year analysis period.

Data from Filipino-written news articles covering a 2-year analysis period from March 2009 to May 2011 was used to support the Culturomic analysis presented in this paper. *Bantay-Wika*, however, is a current project of UP-DSP and UP-SWF, with ongoing data collection efforts from Philippine news websites. The results will be published after the project has been completed. Intellectual Property rights for the text corpora are owned by the University of the Philippines - Diliman and data wil be made available to researchers for free except for requesting parties intending to use them for commercial purposes.

Feature	**Abante-Tonite**	**Abante**
No. of Downloaded Articles	31864	29713
Number of Sentences	442.1K	478.7K
Number of Unique Sentences	413.9K	397.2K
Number of Words	10.6M	10.9M
Number of Unique Words	242K	245.6K

Table 4: Description of Corpora used in Culturomic Analysis

The normalized frequency value of a given search word for a particular date was obtained by dividing the total number of sentences wherein a search word was seen at least once, with the total number of sentences obtained for the given date. Normalized frequency plots for particular search words/phrases have revealed interesting insights into the Filipino culture, as seen in the succeeding graphs. Note that the search process used for all the search words is case-insensitive.

2.2.1 Investigating the actual duration of Filipino Christmas season

The Philippines, being a predominantly Catholic Christian nation, celebrates the longest Christmas season in the world, starting from the onset of September and continuing until the feast of the Sto. Niño celebrated every third Sunday of January of the next year. Counting the number of times the words ‘Christmas’ and ‘Pasko’[8] were mentioned in news media, whether by simple greetings or actual news reportage over a long time period can give us an idea of how intensely Filipinos celebrate Christmas. Figure 2 shows the normalized frequency plots for search words ‘Christmas’ and ‘Pasko’. The plots show annual repeating patterns, with the initial peak found at beginning of September. The next significant peak is on the first day of December, slowly increasing to its peak on December 25, then sharply decreasing by the first week of January. Hence, it can be said that a typical Filipino’s excitement over the thought of Christmas is triggered by certain events: (1) the entry of September which is the first of the ‘-Ber’ months leading to Christmas, (2) the entry of December, the month of Christmas, and (3) the actual day of Christmas.

2.2.2 Investigating the effect of Typhoon Ondoy

On September 24, 2009, the Philippines’ capital of Manila was inundated with floods when it was visited by Super Typhoon *Ondoy* (international code name: Ketsana). Images and videos of flood and devastation filled the storylines of news agencies all over the world and in various online social networking sites for many weeks. Many professed *Ondoy* to be a natural disaster that will never be erased in the collective Filipino memory. Looking at graph of normalized frequency plots for the search word ‘Ondoy’ of Figure 3, we see quite a different story. Approximately one month after *Ondoy* visited, news reportage has almost completely subsided. It can be seen that exactly one year later, news reporting on the topic Ondoy was only slightly revived.

[8] *Pasko* is the Filipino word for ‘Christmas’

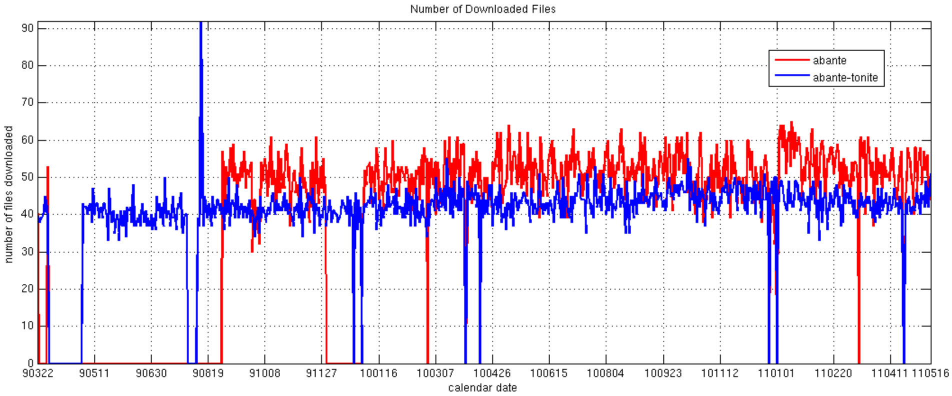

Figure 1: Distribution of downloaded files from Abante and Abante-tonite websites. *The format for the calendar date is YYMMDD where YY is the year, MM is the month, and DD is day the files were downloaded.*

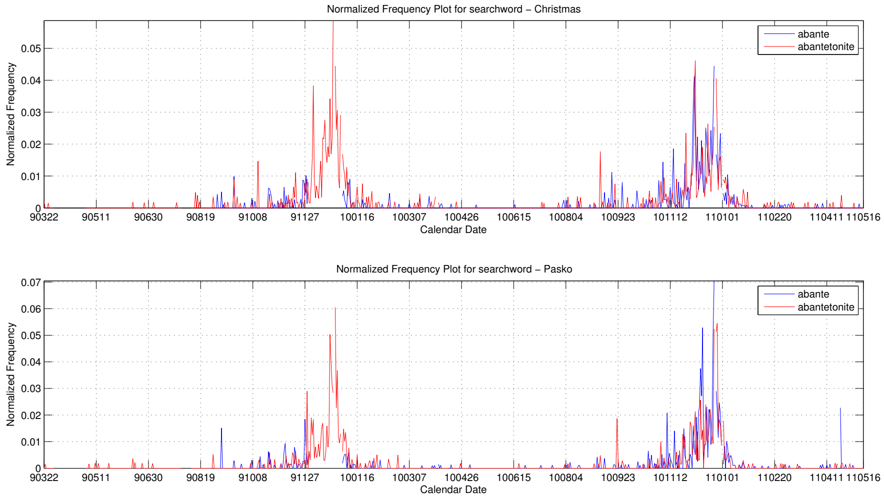

Figure 2: Frequency Plots for ‘Christmas’ (upper plot) and ‘Pasko’ (lower plot). *The format for the calendar date is YYMMDD where YY is the year, MM is the month, and DD is day the files were downloaded.*

2.2.3 Tracking the rise of popularity of the Ampatuans

Figure 4 shows the normalized frequency plots when the search word is ‘Ampatuan’. Data shows that the Ampatuans were relatively under the media radar before the Nov. 23, 2009 massacre of 57 civilians and media people in Maguindanao. It was widely believed that the massacre was masterminded by the heads of the Ampatuan clan, in order to deter their political foe, Ismael Mangudadatu from running against them in the May, 2010 national elections. The next significant peak near Apr. 19, 2010 is when DOJ Acting Secretary Agra absolved the Ampatuans, resulting in an immediate public outrage. The plots clearly show that since the day of the massacre, the Ampatuans have consistently remained to be a favorite news fodder for the Philippine media.

2.2.4 Looking at an example of a linguistic fad

The normalized frequency plot of a linguistic fad such as ‘Jejemon’ (see Figure 5) shows a quick rise, followed by a period of relatively stable frequency plot, but ending with an abrupt fall of mentions in news and media articles. Of all the words that qualified as finalists for the Sawikaan 2010 Word of the year sponsored by the UP-SWF, only *namumutbol* did not have at least 1 entry in the Data Set. The rest of the word finalists for

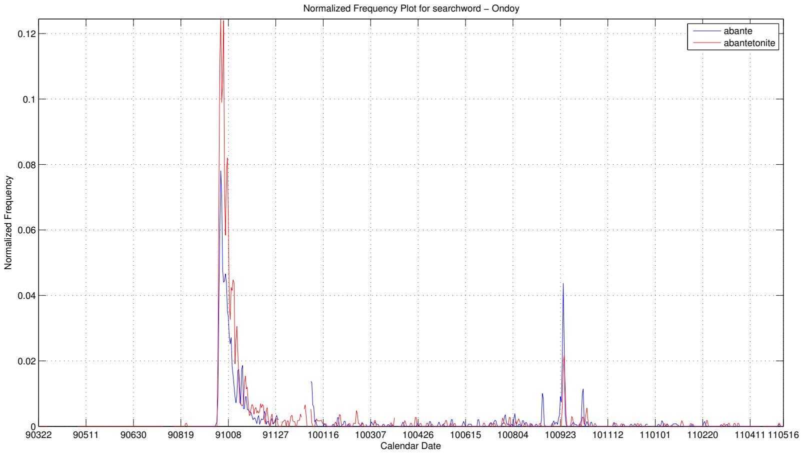

Figure 3: Frequency Plots for 'Ondoy'. *The format for the calendar date is YYMMDD where YY is the year, MM is the month, and DD is day the files were downloaded.*

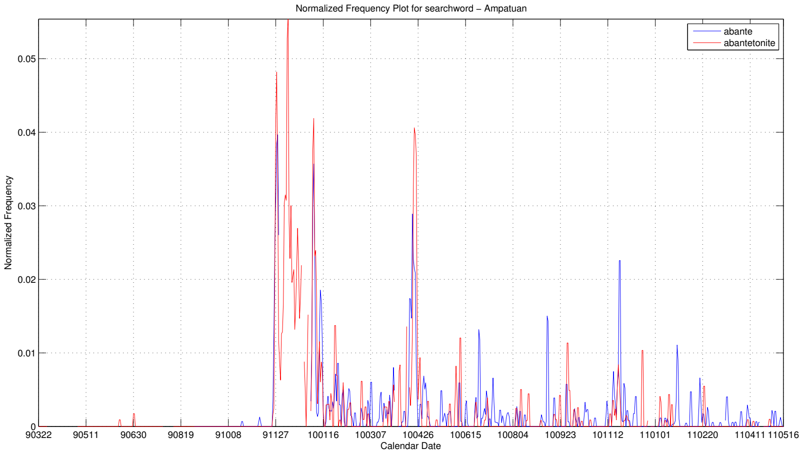

Figure 4: Frequency Plots for 'Ampatuan'. *The format for the calendar date is YYMMDD where YY is the year, MM is the month, and DD is day the files were downloaded.*

Sawikaan 2010 are *Ondoy, Ampatuan, load, Jejemon, solb, miskol, spam,* and *emo*.

2.2.5 Looking at an example of an enduring Filipino news item

Refer to Figure 6 to see the plot for the search phrase 'Cory Aquino'. Former Philippine President Cory Aquino, who was the transitory president after 26 years of Marcos dictatorship, has again been relatively popular in media news starting with rumors of her showing signs of succumbing to colon cancer early July of 2009, with reports peaking on the week of her death on August 1, 2009. Her first death anniversary shows strong public commemoration. Cory Aquino's popularity has remained consistent throughout the whole year.

3 Conclusions and Future Work

In this paper, we have described the *Bantay-Wika project*: it's goals, history, accomplishments and plans for future work. Aimed at helping language planners in attaining the goal of standardizing the use of the Philippine national language, *Bantay-Wika* is first and foremost seen as an objective and effective monitoring mechanism for tracking actual language use. Allowing the conventionalization phenomenon that characterizes the development of a language to be tracked over time enables language agencies such as the UP-SWF to test the effectiveness of their rolled out language

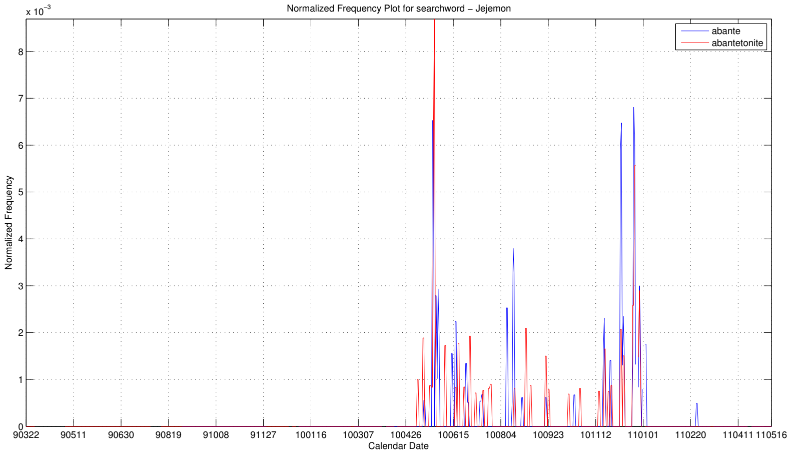

Figure 5: Frequency Plots for 'Jejemon'. *The format for the calendar date is YYMMDD where YY is the year, MM is the month, and DD is day the files were downloaded.*

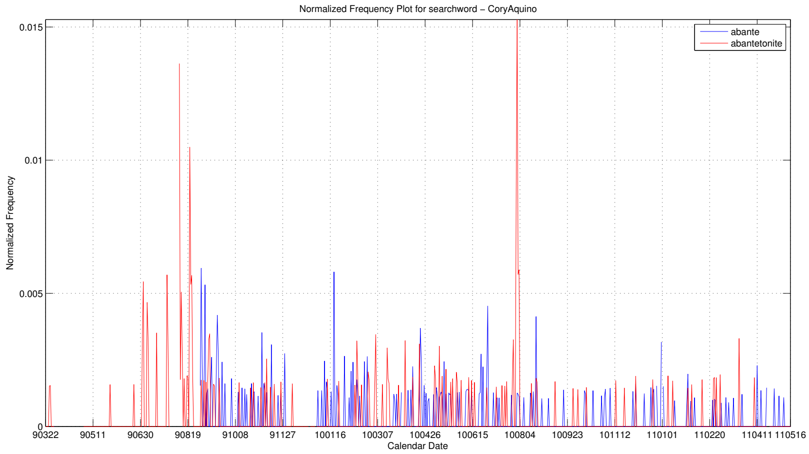

Figure 6: Frequency Plots for 'Cory Aquino'. *The format for the calendar date is YYMMDD where YY is the year, MM is the month, and DD is day the files were downloaded.*

policies and strategies, and design appropriate interventions that would aid strategies that turn out to be ineffective. In this regard, one strong motivation behind the Bantay-Wika project is to ultimately develop computational models for linguistic change, thus allowing language policy makers to reliably forecast the trajectory of language development for each language policy being implemented. Secondly, culturomic analysis of downloaded news articles covering a two-year period has provided us with more insights into the Filipino Christmas tradition, the Filipino outlook in the face of natural and man-made calamities that has attracted global attention, and linguistic fads. Observing the fluctuations in the intensity of news reportage of topics that have pervaded the Philippine national consciousness has given us an objective view of the favorite topics of Filipinos, perhaps a reflection of the values that Filipinos collectively hold dear. Efforts are currently being made to increase the coverage of the text corpora to include written work representative of each stage of the Philippine nation's colorful history starting from the 1900's to the present. With larger and more comprehensive text corpora, the authors are hopeful that more insights into how the Filipino sentiment changes with each milestone in the Philippines' history will be made apparent, thus offering a fresh look at Philippine history in a way not possible before. Finally, the data, the tools for analysis developed for this project, and the actual research findings open up different opportunities for conducting further quantitative investigative work in the areas of Philippine lexicography and diachronic linguistics, and even aid in the further understanding of Filipino philosophy. Truly, it can be said that in the case of the *Bantay-Wika project*, the use of technology has greatly expanded the reaches of scholarly inquiry.

Acknowledgments

The authors are grateful to the UP Diliman - Office of the Chancellor for funding the *Bantay-Wika project*. The first author would like to thank the Department of Science and Technology (DOST) for the Engineering Research and Development for Technology (ERDT) scholarship grant given to him.

References

Daniel M. Abrams and Steven H. Strogatz. 2003. Modeling the dynamics of language death. *Nature*, 424.

C. Ferguson. 1968. Language development. *Language Problems of Developing Nations*, pages 27–35.

Andrew B. Gonzalez. 2003. Language planning in multilingual countries: The case of the Philippines. In *Conference on Language Development, Language Revitilization and MultilingualEducation in Minority Communitites in Asia*, Bangkok, Thailand.

M. Paul Lewis. 2009. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 16th edition.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, pages 176–182.

UP Sentro ng Wikang Filipino. 2004. *Gabay sa Pagbaybay*. UP - Sentro ng Wikang Filipino.

UP Sentro ng Wikang Filipino. 2008. *Gabay sa Ispeling*. UP - Sentro ng Wikang Filipino.

Surian ng Wikang Pambansa. 1976. *Mga Tuntunin sa Ortograpiyang Fiipino*. Surian ng Wikang Pambansa.

Surian ng Wikang Pambansa. 1987. *Alpabeto at Patnubay sa Ispeling ng Wikang Filipino*. Surian ng Wikang Pambansa.

Rachel Edita Roxas, Charibeth Cheng, and Nathalie Rose Lim. 2009. Philippine language resources: trends and directions. In *7th Workshop on Asian Language Resources*, Singapore.

Komisyon sa Wikang Filipino. 2001. *Revisyon ng Alfabeto at Patnubay sa Ispeling ng Wikang Filipino*. Komisyon sa Wikang Filipino.

Komisyon sa Wikang Filipino. 2009. *Gabay sa Ortograpiyang Pilipino*. Komisyon sa Wikang Filipino.

Lope K. Santos. 1940. *Balarila ng Wikang Pambansa*. Surian ng Wikang Pambansa.