# Developing a Tagalog Linguistic Inquiry and Word Count (LIWC) 'Disaster' Dictionary for Understanding Mixed Language Social Media: A Work-in-Progress Paper

**Amanda Andrei, Alison Dingwall, Theresa Dillon, Jennifer Mathieu**
MITRE Corporation
7515 Colshire Drive
Mclean, Virginia 22042 USA
aandrei@mitre.org

## Abstract

In the wake of super typhoon Yolanda (known internationally as Haiyan) in the Philippines in 2013, many individuals in the Philippines turned to social media to express their thoughts and emotions in a variety of languages. In order to understand and analyze the sentiment of populations on the ground, we used a novel approach of developing a conceptual Linguistic Inquiry and Word Count (LIWC) dictionary comprised of Tagalog words relating to disaster. This work-in-progress paper documents our process of filtering and choosing terms and offers suggestions for validating the dictionary. When results on how the dictionary was used are available, we can better assess the process for creating conceptual LIWC dictionaries.

## 1 Background

By engaging in a variety of social networking and blogging activities, individuals often reveal their "perceptions, attitudes, beliefs, and behaviors" (Maybury, 2010) through multiple social media platforms such as Facebook and Twitter. In addition, social media provides an important source for breaking news, especially during natural disasters and emergencies (Nagar et al., 2012; Crowe, 2012). During events such as the 2010 earthquake in Haiti and the 2011 tsunami in Japan, individuals turned to social media to report injuries, ask for assistance, and publish personal accounts (Gao et al., 2011; Abbasi et al., 2012). Likewise, the 2013 disaster of super typhoon Yolanda (known internationally as Haiyan) in the Philippines triggered a wide use of social media during the period of the storm.

### 1.1 Philippines

With its two official languages (English and Filipino) and dozens of other local languages and dialects, the Philippines has a complex and politicized history of multilingualism (Gonzalez, 1998; Nical et al., 2004; Ang, 1978). Both the grammar and vocabulary of Filipino (also known as Pilipino) is based primarily from Tagalog, a language originating from the regions surrounding the capital city of Manila, although some scholars argue that Filipino is essentially Tagalog (Ang, 1978; Baumgartner, 1989).

In 2011, the Philippines had the highest percentage of active online users in the world (Global WebIndex, 2011). In 2012, the nation had more than 10 million active Twitter users, which ranked it tenth in countries with the most Twitter users (Abuy, 2012). Tweets from the Philippines are in mixed languages, with 80% in English and the other 20% in Filipino languages (Pilkington, 2011). Furthermore, the Philippines is the most disaster-prone nation in the world (CDRC Admin, 2013; Bankoff, 2002), making it a prime candidate for analyzing sentiment in social media during and following a natural disaster.

### 1.2 Linguistic Inquiry and Word Count (LIWC)

As social media analysis continues to mature as a field, if social media is to be leveraged more effectively for disaster response and relief there is a need for more quantitative methods to supplement current qualitative techniques and subject matter expertise (Servi and Elson, 2012). Servi & Elson (2012) used the novel approach of combining mathematical algorithms with a social psychology tool, Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007), to detect and forecast emotional trends in Twitter in an unbiased way.

LIWC uses internal "dictionaries" of words which correspond to various domains of linguistic



*Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014*, pages 91–94,
Gothenburg, Sweden, April 26 2014. 

processes, psychological processes, personal concerns, and spoken categories. For instance, one dictionary under psychological processes is filled with *positive emotion* words (e.g., "love," "nice," and "sweet"). Another dictionary under personal concerns includes *death* words (e.g., "bury," "coffin," "kill"). When a researcher runs a text file through LIWC, the program compares the words in the file to the dictionaries and outputs a ratio of dictionary words to total words (e.g., 1% of all words are positive emotion words). Researchers have used LIWC to analyze a variety of texts, most notably newspaper coverage of a bonfire tragedy at Texas A&M (Gortner and Pennebaker, 2003), one of the earliest examples of using LIWC to understand emotions around disaster management.

First created in English, LIWC has also been translated into other languages. This project is developing a set of LIWC dictionaries in Tagalog in order to understand social media usage in the Philippines, particularly usage during the typhoon. Tagalog was chosen and distinguished from Filipino for three main reasons: namely, that more information about and translations in Tagalog are readily available, to highlight the fact that there are more Philippine languages beyond "Filipino" that could be translated as well, and because most tweets originated from Manila, where Tagalog is the main Philippine language spoken.

## 2 Process

On November 8, 2013, super typhoon Yolanda made landfall in the central Philippines. Over 11 million people were affected, with 2.5 million people in need of food aid and over 2,000 people dead (BBC News Asia, 2013).

As the Philippines is linguistically diverse, there remained a need to also explore the tweets that were posted in mixed languages, primarily Tagalog since it is one of the major languages in the nation. A LIWC dictionary of disaster-related words in Tagalog was developed in order to gauge how many tweets during the period of the typhoon related to the disaster. To explore the impact of the typhoon from the publics perspective in social media, mixed language Twitter posts geographically restricted to the Philippines were analyzed.

Using a commercially available social search and analytics too which filters Twitter content based on variables such as location, time and date, tweet type (original, retweet, reply), language (although Tagalog is not included) and others, a volume of approximately 1.5 million tweets from the Philippines were identified within a two-week date range around the typhoon. This set was isolated based on restricting the tweets to those originating from the Philippines between the dates of November 3-18, 2013 and included any of the following terms: *typhoon*, *yolanda*, *haiyan*, *supertyphoon*, as well as corresponding hashtags.

A wide range of words and concepts relating to typhoons and disasters, such as *baha* (storm), *donasyon* (donation), *nagugutom* (starving), *patay* (dead), and *sagip* (save) were identified. Related terms were also identified and included in the search, such as *#bangon* (rise up), a nationalistic call of inspiration; *#walangpasok* (no entry), colloquially a school closing alert; *Libreng Tawag* (free calls), used to alert users which telecommunication companies were allowing no charge phone calls; and *PAGASA* (hope), which is also an acronym for the Philippine weather alert service.

In order to develop a clean data set, these terms were then narrowed down based on what would make appropriate inclusions for a LIWC dictionary, resulting in the discarding of hash tags, multi-word phrases, and proper nouns. Certain words were also found to be too broad (i.e., false positives), such as donasyon, which was used in non-disaster contexts just as frequently as in typhoon-related tweets within the date range analyzed. Words like nagugutom and patay were actually used more frequently in non-disaster contexts (e.g., "I am starving, I want a sandwich").

The dictionary was designed to include different grammatical forms of words. For example, for nouns, both *baha* (flood) and *bahang* (flood), where ng is a linking suffix, were included. For verbs, different tenses were included, e.g., *tulong* (help), *tumulong* (helped), and *tumutulong* (helps). In the case of the verbs, other forms of the words were searched, but not included in the dictionary if they were not frequently used in tweets. The complete dictionary is included in the Appendix.

## 3 Remarks and Future Work

The Tagalog LIWC disaster dictionary was developed to quickly explore and understand perceptions expressed on social media about the typhoon. While the terms were included for the 2013 typhoon, additional research and validation is required for generalization for understanding future

natural disasters. While social media can contain a wealth of information, the processes of filtering and searching for terms would benefit from a more rigorous standard of including words in the dictionary. For instance, researchers may want to consider what counts as high frequency for a word, e.g., if it appears over a certain absolute number of times or if it appears in high proportion compared to other words. Overall, a move from qualitative analysis to more quantitative analysis would clarify the connection between the dictionary and the source corpus.

The process of creating a conceptual LIWC dictionary should also be vetted against other use cases and concepts. For instance, the word *lindol* (earthquake) was included in the dictionary since earthquakes are common in the Philippines, although earthquake activity was not recorded during the typhoon. The dictionary could be evaluated or validated against other social media responses to other recent disasters, such as the October 2013 earthquake in Bohol, an island near the typhoon-struck areas, in order to see how users tweet about disasters.

Geography also plays an important role in how the disaster dictionary can be used. For the purposes of creating this dictionary, tweets were restricted to the Philippines. It would be worthwhile to examine if the same words in the dictionary occur if tweets were collected from different origins, such as Leyte (the island which sustained most of the damage) versus Manila (the capital city of the Philippines) versus a location with a large concentration of Filipino immigrants (such as California, USA).

Additionally, other concepts related to disaster management should be explored and considered for inclusion in the dictionary, such as words relating to property, family, and emotions. As the original (English-language) LIWC application already has categories for such concepts, future work would include translating the complete set of LIWC dictionaries into Tagalog while also including culturally specific words without exact translations. This work is currently in progress.

Furthermore, the areas hit by the typhoon speak and use social media in other Philippines languages in addition to Tagalog (primarily Cebuano and Waray). It may also be helpful to have dictionaries in other languages predominant in the area where a disaster occurs. This may be a difficult task to undertake, as translations for other Philippine languages are not as readily available as translations for Tagalog.

This paper details the process for creating the dictionary; how the dictionary was used in actual social media datasets concerning the typhoon is still in progress. Upon reviewing how the disaster dictionary was used, this process of creating concept LIWC dictionaries and its utility will be better assessed and validated. Since this tool and the additional LIWC dictionaries are still in their preliminary formats, there are no current plans to make the tools commercially available until they are reviewed and vetted by native Tagalog speakers. As the work progresses, the disaster dictionary will be maintained and kept up-to-date in order to include additional terms which may apply to future disasters.

## References


Mohammad-Ali Abbasi, Shamanth Kumar, Jose Augusto Andrade Filho, and Huan Liu. 2012. Lessons learned in using social media for disaster relief - ASU crisis response game. *Social Computing, Behavioral-Cultural Modeling and Prediction*, 7227, 282-289.

Abiel Abuy. 2012. "Twitter crosses 500 million mark, Philippines in the top 10 in terms of Twitter accounts." 2 Aug 2012. *KabayanTech*. http://kabayantech.com/2012/08/twitter-crosses-500-million-mark-philippines-in-the-top-10-in-terms-of-twitter-accounts/

Gertrudes R. Ang. 1978. The Filipino as a bilingual or multilingual: Some implications. *Philippine Quarterly of Culture and Society*, 187-189.

Gregory Bankoff. 2002. *Cultures of disaster: Society and natural hazard in the philippines.* Routledge-Curzon, New York, NY.

Joseph Baumgartner. 1989. The controversy about the national language: Some observations. *Philippine Quarterly of Culture and Society*, 168-172.

BBC News Asia. 2013. "Typhoon Haiyan: Aid in numbers." 14 Nov 2013. *BBC News*. http://www.bbc.co.uk/news/world-asia-pacific-24899006

CDRC Admin. 2013. "Philippines is most disaster-affected country in 2012." 8 Apr 2013. *Citizens' Disaster Response Center*. http://www.cdrc-phil.com/philippines-is-most-disaster-affected-country-in-2012/

Adam Crowe. 2012. *Disasters 2.0: The application of social media systems for modern emergency management.* CRC Press: Boca Raton, FL.

Juiji Gao, Geoffrey Barbier, and Rebecca Goolsby. 2011. Harnessing the crowdsourcing power of social media for disaster relief. *Intelligent Systems, IEEE*, 26(3), pp.10,14, May-June 2011.

Global WebIndex. 2011. "Global Map of Social Networking 2011." *GlobalWebIndex.Net*. https://globalwebindex.net/wp-content/uploads/downloads/2011/06/Global-Map-of-Social-Networking-GlobalWebIndex-June-20112.pdf

Andrew Gonzalez. 1998. The language planning situation in the Philippines. *Journal of Multilingual and Multicultural Development*. 19(5), 487-525.

Eva-Maria Gortner and James W. Pennebaker. 2003. The archival anatomy of a disaster: Media coverage and community-wide health effects of the Texas A&M bonfire tragedy. *Journal of Social and Clinical Psychology*. 22, 580-603.

Mark Maybury. 2010. "Social Radar for Smart Power." *The MITRE Corporation*. http://www.mitre.org/sites/default/files/pdf/10_0745.pdf

Seema Nagar, Aaditeshwar Seth, and Anupam Joshi. 2012. Characterization of social media response to natural disasters. *Proceedings of the 21st international conference companion on World Wide Web (WWW '12 Companion)*, ACM, 671-674.

Iluminado Nical, Jerzy J. Smolicz, and Margaret J. Secombe. 2004. Rural students and the Philippine bilingual education program on the island of Leyte. *Medium of instruction policies - Which agenda? Whose agenda?*, 153-176. Lawrence Erlbaum Associates, Mahwah, NJ.

James W. Pennebaker, Roger J. Booth, and Martha E. Francis. 2007. Linguistic Inquiry and Word Count: LIWC2007 Operators manual. *LIWC.net*.

Andy Pilkington. 2011 "Axe shows consistency is key to a successful multi-lingual page in the Philippines." 26 Oct 2011. *WaveMetrix*. http://wave.wavemetrix.com/content/axe-shows-consistency-key-successful-multi-lingual-page-philippines-00844

Les Servi and Sara Beth Elson. 2012. A mathematical approach to identifying and forecasting shifts in the mood of social media users. *The MITRE Corporation*. Bedford, MA.

Paul Schachter, and Fe T. Otanes. 1972. *Tagalog Reference Grammar*. University of California Press, Berkeley, CA.

# A Appendix

The completed dictionary is included in the following table. consists of the Tagalog and English columns. In some cases, multiple dictionary entries correspond to the same Tagalog lexeme. For example:

bagyong
bagyo-ng
storm-LIGATURE

For more on Tagalog grammar, see Schachter and Otanes 1972.

| **Tagalog** | **English** |
|---|---|
| bagyo | storm |
| bagyong | storm |
| baha | flood |
| bahang | flood |
| biktima | victims |
| hangin | wind |
| lindol | earthquake |
| lumikas | evacuate |
| nagsilikas | refugees |
| nasawi | casualty |
| sagip | rescue |
| sagipin | rescue |
| sinalanta | devastated |
| sugatan | wounded |
| tulong | help |
| tumulong | help |
| tumutulong | help |
| ulan | rain |