# Philippine Languages Online Corpora: Status, issues, and prospects

**Shirley Dita**
Department of English and Applied Linguistics, De La Salle University, Manila
shirley.dita@gmail.com

**Rachel Edita O. Roxas**
Center for Human Language Technologies, College of Computer Studies, De La Salle University, Manila
rachel.roxas@delasalle.ph

## Abstract

This paper presents the work being done so far on the building of online corpus for Philippine languages. As for the status, the Philippine Languages Online Corpora (PLOC) now boasts a 250,000-word written corpus of the eight major languages in the archipelago. Some of the issues confronting the corpus building and future directions for this project are likewise discussed in this paper.

## 1 Introduction

The 171 living Philippine languages have been the subject of linguistic investigations and descriptions all over the world (see Liao 2006; Quakenbush 2005; Reid 1981; *inter alia*). As there are controversial and interesting features of Philippine-type languages that are distinct from other Austronesian languages, Philippinists have focused on the different features of Philippine languages over the years. For instance, Brainard (1994) has looked at voice and ergativity; or the focus system (see Barlaan 1986); or case system (see Ramos 1997); and recently, Dita (2010) on pronominal system. But even with a considerable overlap in syntax and morphology, there is a wide range of typological variety found among the more than one hundred Philippine languages (Reid & Liao 2004). And since the plethora of research in Philippine linguistics has been done by non-Filipinos and/or non-Philippine residents, authors have utilized various means to get hold of data on Philippine languages. The methodological approaches of previous studies on language description can be summed up to three: 1) researchers come to the Philippines and stay in the place where the language is spoken for a time; 2) researchers work with a native speaker of the language who currently resides abroad (close to the researchers); and 3) researchers use printed or published materials about the language of interest. It is against this scenario that building a corpus of Philippine languages was conceptualized.

In Dita, Roxas, & Inventado (2009), the design and scope of the first phase of the corpus building were described. As was mentioned, the primary consideration of the data collection was the comparability of the texts in the languages included. Hence, the first phase of the project included a rather limited text type and category, that is, only written texts with two categories: literary and religious. Although there was a plan then to include journalistic and academic texts, it has been observed that not all languages have these text types.

In what follows, we will describe the second phase of the Philippine Languages Online Corpora (henceforth, PLOC), and its current status in terms of data collection and analysis, its distinguishing features, and the issues encountered in the corpus building. Recommendations and future prospects are then outlined towards the end of the paper.

## 2 Architecture and Parameters

The initial idea was to pattern PLOC after the International Corpus of English where every variety includes a one million-word collection of both written and spoken texts. And as Dita et al. (2009) have emphasized, the first phase of the project faced serious time constraints. This led to the decision to include the most popular kind of text in any Philippine language: religious and literary texts.

### 2.1 Size and Scope

The standing goal of the project is to provide a comparable corpus of as many Philippine lan-



*Proceedings of the 9th Workshop on Asian Language Resources*, pages 59–62,
Chiang Mai, Thailand, November 12 and 13, 2011.

guages as possible. To be able to do this, the first phase of the project consisted of the four top languages of the Philippines (Tagalog, Cebuano, Ilocano, and Hiligaynon) and the Filipino Sign Language (FSL). The second phase includes the next four top languages (Bikol, Kapampangan, Pangasinense, and Waray-waray). Hence, the project now consists of the eight major languages in the Philippines and the FSL.

| Language | Native Speakers (Millions) | Percentage of Population |
|---|---|---|
| Tagalog | 17 | 24.0 |
| Cebuano | 15 | 21.0 |
| Ilocano | 8 | 11.0 |
| Hiligaynon (3 dialects) | 7 | 10.0 |
| Bicolano (5 dialects) | 3.5 | 7.0 |
| Waray-waray | 2.4 | 4.6 |
| Kapampangan | 1.9 | 3.7 |
| Pangasinan | 1.1 | 2.3 |
| Maguindanao (2 dialects) | 1 | 1.7 |
| Total | 56.9 | 87 |

**Table 1. The major Philippine languages**

Initially, there was a plan to pattern the PLOC after the structure and design of the International Corpus of English (ICE) project. As reported by Bautista (2004), the Philippine component of the International Corpus of English (ICE-PHI) which is composed of over one million words (1,106,778 words, to be exact) of spoken and written English, is the first mega-word electronic corpus produced in the Philippines. The PLOC is envisioned to be the first multi-million-word electronic and online corpora of Philippine languages. But as there are more languages included in the project, the 1 million size was reduced to 250,000 words for each language, and the written and spoken was reduced to written only.

To date, PLOC only has the written genre and has two types of writing so far: religious and literary. Hence, for each language included, there are 100,000 words of religious texts and another 150,000 literary texts. Although the limitation of the data poses drawbacks, it is the comparability of the data that was taken into account.

### 2.2 Some Issues

The collection of texts had its own share of challenges. Even with the advancement of technology or the availability of OCRs, the kind of texts needed for the corpora could hardly be scanned. Religious and literary texts are usually the oldest existing literature on any given language. Some of these texts had to be manually encoded, proofread, verified by a native speaker, tagged for corpus use, then uploaded. In most cases, a particular language has more literary text than religious while in some cases, there are more religious than literary. It is in this case that texts need to be carefully chosen to avoid corpus disproportion.

One of the inevitable limitations of the corpora is the issue of representativeness. As earlier mentioned, the scope of texts for PLOC only includes religious and literary which is very specific. Although a corpus, as Leech (1991) puts it, is ideally a representative of the language variety it is supposed to represent, the PLOC do not claim any representativeness of the languages included. But even with this limitation, the corpora still promise to be reliable bases for any linguistic investigation or analysis of Philippine languages.

### 2.3 Corpus Tagging

After the data is encoded, verified, and proofread, the next stage is the corpus tagging. In cases where the data exhibit some formatting styles such as texts in boldface, italics, or underline, the codes , , <u></u> are placed before and after the texts which are boldfaced, italicized, and underlined, respectively. For special characters, the following are used: <pd></pd> for non-end of sentence and <lbl></lbl> for special labels such as bible verses. As far as the data is concerned, these tags are sufficient to cover the special formatting styles and characters found in the documents.

## 3 Corpus Processing and Tools

### 3.1 Online Repository

One distinguishing feature of PLOC is its being accessible online, unlike most corpora where they are distributed separately. The data repository is called *Palito,* a Tagalog term which literally means 'stick'. The online repository requires a username and password which can be requested from the website master(s) upon verification of identity. Users can either upload or download a document, both subject to the ap-

proval of the webmaster. For users who want to contribute to the existing corpus, they send in their documents, the webmaster verifies accuracy and genuineness of the data then uploads them to the online repository. The repository automatically indexes the documents so they can easily be tracked. *Palito* can be accessed through the following link: ccs.dlsu.edu.ph:8086/Palito


**Figure 1**. Screenshot of Palito's front page

### 3.2 Features of Palito

Another feature of the repository is its internal browser. For users who want to search for a particular document, such as short story or song, users are to click the 'document' icon then type in the specific name of the document. All files under this category or name are then displayed.

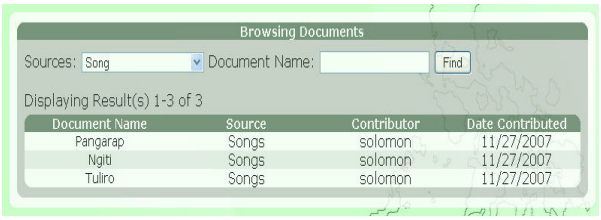


**Figure 2.** *Palito*'s internal browser

For the word frequency feature, the search key is simply typed in the 'filter words'. After which, the results are displayed which makes it more convenient for users who are interested in the distribution of a particular lexical item across the documents. As can be seen in figure 3, the results display the results in descending order, from the files with the most number of the keyword to the least.

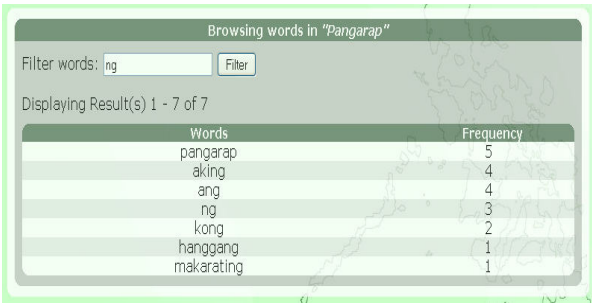


**Figure 3.** *Palito*'s word frequency feature

Finally, *Palito* is equipped with concordance which generates a list of the occurrences of a specific word under search. The concordance r is an indispensable tool for any corpus as it conveniently displays all occurrences of the keyword in the entire collection. Figure 4 shows how concordance works for PLOC.

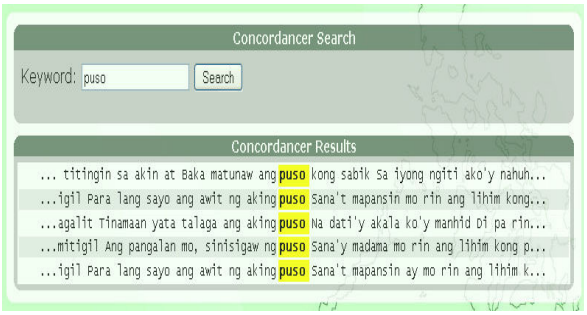


**Figure 5**. A screenshot of *Palito's* concordancer

Aside from the data in the Philippine corpus, the software tools will also aid language researchers everywhere in analyzing the Philippine languages, such as comparing different usages of the same word, analyzing collocates, and finding and analyzing phrases and idioms.

## 4 Conclusion

This paper has presented the work done so far in the building of the PLOC. As Dita et al. (2009) have reported, there were many things to consider in conceptualizing the first phase of the project. And since the primary consideration was the comparability of texts in different languages, the scope was rather limited to written genre in general, and to literary and religious texts, in particular. The second phase of the project has so far completed a 2-million word corpus of the eight major Philippine languages (Taga-

log, Cebuano, Ilocano, Hiligaynon, Bicol, Kapampangan, Pangasinense, and Waray). In summary, the present PLOC now contains a 250,000-word written texts of the eight major languages in the Philippines.

There are many plans for the expansion of the PLOC. First, we plan to collect a 250,000-word counterpart in spoken texts. The spoken texts will consist of dialogues and monologues: face-to-face conversations for dialogues and speeches or radio/tv commentaries for monologues. Second, we plan to extend the coverage of written texts by including journalistic writing and other 'creative' writings such as advertisements, internet texts and the like. The plan is to get a one million-word of written and spoken corpus, for every language. If this is achieved, the expansion for the PLOC is to include as many Philippine languages as possible. When all these plans are achieved, there will be more chances of comparing features cross-linguistically, following the classic works of Blake (1906), Tsuchida, Yamada, Constantino, & Moriguchi (1989), and Constantino (1965), to name a few.

The PLOC, as it is envisioned to be, will make a significant contribution to Philippine linguistics. Doing linguistics by then, to quote Bautista (2004), would mean "sitting in one's armchair and introspecting and thinking of sample sentences to exemplify particular structures" (p. 1), as opposed to going through the laborious task of fieldwork to collect data from different informants.

**Acknowledgments**

This project has been partially funded by the National Commission for Culture and the Arts, Philippine Government.

## References

D. Biber, S. Condrad, and R. Reppen. 1998. *Corpus linguistics: Investigating language structure and use.* Cambridge: CUP.

E. Constantino. 1965. The sentence patterns of twenty-six Philippine languages. *Lingua 15*, 71-124.

F. R. Blake. 1906. Contributions to comparative Philippine grammar (Part 1). *Journal of the American Oriental Society 27,* 317-396.

G. Leech. 1991. The state of the art in corpus linguistics. In K. Ajmer & B. Altenberg (Eds.), *English Corpus Linguistics: Linguistic Studies in Honor of Jan Svartvik,* (pp. 8-29). London: Longman.

H. Liao. 2006. Philippine linguistics: The state-of-the-art 1981-2005. Paper presented at the Annual Lecture of the Andrew Gonzalez, FSC Distinguished Professorial Chair in Linguistics and Language Education on March 4, 2006. De La Salle University, Manila, Philippines.

J. S. Quakenbush. 2005. Philippine linguistics from an SIL perspective: Trends and prospects. In H. Liao & C.R.G. Rubino (Eds.), *Current issues in Philippine linguistics and anthropology: Parangal kay Lawrence A. Reid* (pp. 3-27). Manila: Linguistic Society of the Philippines and SIL Philippines.

L. Reid. 1981. Philippine linguistics: The state of the art: 1970-1980. In D. V. Hart (Ed.), *Philippine studies: Political science, economics, and linguistics* (pp. 212-273). DeKalb: Center for Southeast Asian Studies, Northern Illinois University.

L. Reid and H. Liao. 2004. A brief syntactic typology of Philippine languages. *Language and Linguistics 5*(2), 433-490.

M. L. S. Bautista. 2004. An Overview of the Philippine Component of the International Corpus of English (ICE-PHI). *Asian Englishes*, *7*(2), 8-26.

P. M. Lewis. (Ed.) 2009. *Ethnologue: Languages of the world* (16th ed.) Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/

R. Barlaan. 1986. *Some major aspects of the focus system in Isnag*. Ph.D. dissertation, University of Texas at Arlington.

Shigeru. Tsuchida, Y. Yamada, E. Constantino, and T. Moriguchi. (1989). *Batanic languages: Lists of sentences for grammatical features.* Tokyo: The University of Tokyo.

Shirley. N. Dita. 2010. A morphosyntactic analysis of the pronominal system of Philippine languages. *Proceedings of the 24th Pacific Asia Conference in Language, Information & Computation* (pp. 45-59). Tokyo: Waseda University Press.

Shirley. N. Dita, R.E.O Roxas, and P. Inventado. 2009. Building Online Corpora of Philippine Languages. *Proceedings of the Twenty-third Pacific Asia Conference on Language, Information and Computation,* 646-653.

S. Brainard. 1994. *Voice and ergativity in Karao.* Ph.D. dissertation, University of Oregon.

T. V. Ramos. 1997. *Case system of Tagalog verbs*. Ph.D. dissertation, University of Hawaii.