Building Online Corpora of Philippine Languages

Shirley N. Dita[a], Rachel Edita O. Roxas[b], and Paul Inventado[b]

[a]Department of English and Applied Linguistics, De La Salle University
2401 Taft Avenue, Manila, Philippines
shirley.dita@dlsu.edu.ph
[b]College of Computer Studies, De La Salle University
2401 Taft Avenue, Manila, Philippines
{rachel.roxas, paul.inventado}@dlsu.edu.ph

Abstract. This paper aims at describing the building of the online corpora on Philippine languages as part of the online repository system called *Palito.* There are five components of the corpora: the top four major Philippine languages which are Tagalog, Cebuano, Ilocano and Hiligaynon and the Filipino Sign Language (FSL). The four languages are composed of 250,000-word written texts each, whereas the FSL is composed of seven thousand signs in video format. Categories of the written texts include creative writing (such as novels and stories) and religious texts (such as the Bible). Automated tools are provided for language analysis such as word count, collocates, and others. This is part of a bigger corpora building project for Philippine languages that would consider text, speech and video forms, and the corresponding development of automated tools for language analysis of these various forms.

Keywords: corpora, Philippine languages, corpus-building, concordancer, sign language

1 Introduction

With 168 languages spoken natively in the archipelago (Gordon 2005), Philippine linguistics has persistently become a fertile ground for investigation. Foreign linguists exhibit a remarkable interest on the richness of Philippine languages. The studies of Constantino (1971), McKaughan (1971), Reid (1981), Quakenbush (2005), and Liao (2006) have outlined the development of Philippine linguistics in the last 25 years. With all the researches that have been done, Liao (2006) underscores the pressing need to document major and minor languages in the Philippines. And since the vast majority of studies in Philippine languages are done by non-Filipinos, Liao emphasizes the demand for Filipino linguists to be involved in the documentation of Philippine languages. Additionally, she highlights that in the last 25 years, there had been 14 M.A. theses and 16 Ph.D. dissertations written about Philippine-type languages, but only one of them was written by a Filipino (i.e., Daguman 2004).

Researchers from all parts of the world have continually visited the Philippines to gather data for their studies. Others, however, have to resort to other means to obtain the needed data. For instance, Davis, Baker, Spitz and Baek (1998) came up with a grammar of Yogad (spoken in northern Luzon) in which the basis was just a native speaker of the language who now resides in Texas. Some would use published texts as corpus for their study (cf. Liao 2004). Others (cf. Ruffolo 2005; Rubino 1997; among others) would have a trip or two only to the country for data gathering due to travel nuisances. Thus, the accessibility of databank on

Copyright 2009 by Shirley N. Dita, Rachel Edita O. Roxas, and Paul Inventado

Philippine languages is a major concern for those interested to further explore issues concerning Philippine linguistics.

The online corpora of Philippine languages project poses several benefits. It is especially of great importance to those interested in Philippine languages, both locally and internationally. For the local researchers, the availability of Philippine data will enthuse linguists to delve into their own languages. For foreign researchers, the databank will expedite their studies on the one hand, and will pave way for other areas of studies, on the other. With this project, those interested with Philippine languages need not come personally to the Philippines to gather necessary data. Hence, easy access to Philippine data, as Dita (2007) conjectures, will pave the way to more researches on various fields such as syntax, semantics, pragmatics, sociolinguistics, and others. The online data on Philippine languages is then a valuable source to any linguist, or researcher, for that matter. As Meyer (2002:11) puts it, "linguists of all persuasions have discovered that corpora can be very useful resources for pursuing various linguistic agendas."

Thus, a language corpus is an indispensable component in researches on various aspects of the study of the nature and functions of natural language, and its multi-faceted applications such as language education, lexicography, and natural language processing. There have been various applications being developed across the country which are an English-Filipino machine translation system (which includes part-of-speech taggers, morphological analyzers and generators, English-Filipino lexicon for translation), and spell checkers for Tagalog, to name a few.

A corpus is a collection of documents in a particular language or languages that are to be stored, managed and analyzed in digital form. Francis (1982:7) defines corpus as "a collection of texts assumed to be representative of a given language, dialect, or other subset of language, to be used for linguistic analysis." In addition, Engwall (1992:167) gives another definition of corpus: "a closed set of texts in machine-readable form established for general or specific purposes by previously defined criteria." It is indispensable in researches on various aspects of the study of the nature and function of natural language, and its multi-faceted applications such as language education, lexicography, and natural language processing. In the last three decades or so, investigations of language use have focused on corpus-base approaches as these "provide a means of handling large amounts of language and keeping track of many contextual factors at the same time" (Biber, Conrad, & Reppen 1998).

The International Corpus of English (ICE) which started in mid-1990s has been one of the seminal works of corpus linguistics. With almost 20 countries participating in this project, the ICE has provided a comparative study of the different Englishes of the world. Greenbaum (1996) provides a comprehensive discussion of the common design of ICE corpora in which all countries should prescribe to. Although the majority of the member countries have completed the project, some are still working for the completion of these corpora. This fact clearly illustrates that corpus building is no mean feat.

The availability and accessibility of the Philippine component of the International Corpus of English (ICE-PHI) paved the way to various researches by linguists worldwide. The countless possibilities of corpus linguistics and the detailed description of the creation and completion of the ICE-PHI, as provided by Bautista (2004), have been the impetus for the online corpora of Philippine languages.

Corpora have been developed since then on other languages, both major and minor languages such as Spanish, German, Chinese, Japanese, Malay and Thai. The Malay corpus which is focused on the study of classical Malay literature features more than 4 million words and 95 texts, including 80,000 verses. The LOTUS (Large vOcabulary Thai continuous Speech Corpus) Thai corpus has 5,000 vocabularies for speech recognition. Another Thai corpus has 44,000 images of handwritten characters.

With the existing infrastructure as provided by the Internet that virtually connects people from various physical locations, contributing to development of such collection of documents is now a reality.

2 The Project

2.1 An overview

This project, called the online corpora of Philippine languages, is divided into five components: four of these represent the top four languages spoken in the Philippines, i.e., Tagalog, Cebuano, Ilocano, and Hiligaynon and the fifth component represents the Filipino Sign Language (FSL). The technical aspect of the project is actually considered another component. This interdisciplinary project is a joint effort of the following departments and institutions: the National Commission for the Culture and Arts (NCCA), the Komisyon sa Wikang Filipino (KWF); the Philippine Federation of the Deaf and the Philippine Deaf Resource Center, the College of Computer Studies, the Department of English and Applied Linguistics, the Department of Literature, and the Department of Filipino, respectively.

The project was given three months for the data gathering and completion. Although the original plan was to come up with one million words of the top four Philippine languages, as patterned after the ICE design, the time constraint compelled the project director and coordinators to reduce the number of words to 250,000 and to scrap altogether the spoken aspect of the corpus. To achieve comparability of the four languages, the 250,000-corpus has only two components: literary texts and religious texts. The Philippine Sign Language, on the other hand, boasts 7,000 signs in video format. The technical aspect of the project is discussed in separate section, see section 3.0.

2.2 Scope and limitation

There had been several issues that needed to be settled before the project was finalized. Among these issues are the types of texts that should be included in it, the number of samples for each text type and the length of each text sample.

A number of limitations were seen inevitable in the project. The issue of comparability was a major concern. Since the present project would eventually become a part of a bigger project, the text type was a crucial issue in the corpus building. Other Philippine languages may not have some text types that some languages have. Hence, the corpus only included literary and religious types. Literary texts comprise 150,000 and religious comprise 100,000 of the corpus. Included in the literary texts should only be prose forms such as novels or short stories, legends, or epics. Songs, proverbs, riddles, and poems are therefore not included in the scope of literary texts as these types of literature exhibit deviances in lexicon, semantics and syntax. On the other hand, religious texts come mainly from bible verses. Prayers, religious songs or chants are therefore not allowed due to the nature of their construction.

With these two text types included in the corpus-building, the issue of representativeness is another problem. The samples in the corpus may not be a representative of the type and time of the collected texts. As Biber (1993) opines, the extent to which a sample includes the full range of variability in a population. Sampling is then an imminent limitation of a corpus building. As is well established that a corpus can never be fully representative, Leech (1991) clarifies that whatever findings the corpus yield can be generalized to a larger hypothetical corpus.

In addition, the corpus only contains written texts, given the intricacies of collecting and, eventually, transcribing spoken texts. However, there is a strong recommendation to include spoken texts in the expansion of the project. As this project is the initial work to a larger Philippine Languages Corpora, it is hoped that spoken texts be included in the succeeding projects.

It should be noted that the corpus does not provide an English translation of the languages. In keeping with the objective of the project which is to provide a digital corpus for those interested in Philippine languages, the project thus focuses on the native languages only.

2.3 Philippine Languages Corpora Text Type and Categories

The types and categories that have been agreed upon by the project coordinators went through several revisions. As earlier mentioned, the original plan was to pattern after the ICE-PHI's text types: written and spoken. But with the time constraint dictated by the project, the number of words was halved and the text type was reduced to written texts only. Of the written texts, it was not possible to subdivide it further to printed and non-printed since other Philippine languages may not have these types. Hence, the categories under written texts were simplified into literary and religious texts, respectively. The literary category included only prose forms. As is the nature of poetry, the extent of an utterance is not definite. It is rather difficult to identify the number of sentences or the end of sentences.

The literary texts constituted a bigger part of the corpus, that is, 150,000 words. The religious texts, on the other hand, comprised the other 100,000 words. Just like the literary texts, religious texts included only the prose forms. Bible verses are the best representative of this type. Prayers and religious songs do not also have clear-cut sentence end and may therefore yield problems in the future.

Another rationale that prompted the choice for religious and literary texts is the time component of the utterances. The language of the literary texts is more contemporary, whereas, the religious is more formal and antiquated. Such characteristic may be helpful for any comparison or contrast of the languages of these text types.

The Philippine Sign Language was composed of illustrations of letters, basic terms and basic expressions. As is the nature of sign language, the letters or words are illustrated in the form of an action which is captured in video. A caption of the foregoing accompanies the actions.

2.4 Method of text collection

Of all the components, the collection of Philippine sign languages was indeed the most difficult task. Not only did it necessitate elaborate room set-up, it also demanded more human resources as the technical assistance is more complex. The intricacies of digitizing the signs called for the expertise of quite a number of people. The Philippine Deaf Resource Center kindly managed all these details of text collection.

As for the Philippine languages, the initial stage of the project was devoted to looking for potential materials for inclusion in the corpus. The issue on copyright of published materials was mainly the difficulty that the coordinators had to deal with. After identifying the materials that qualified under each category, securing permission from authors and publishers came next, then the collected materials were encoded by the research assistants followed by a careful proofreading of the texts.

The markup of the corpus came after the data collection stage. Textual markup, as Nelson (1996) suggests, is necessary since the features of the original text are lost when it is converted into a computerized text file. A list of characters that need special coding was provided. Among these include the styles of formatting such as 'bold', 'underlined', and 'italicized.' Likewise, characters such as non-end sentence periods and labels called for special codes. The mark-up of the texts was carefully edited and checked by the language coordinators of each component.

2.5 Some caveats

With the aforementioned limitations of this project, it is just appropriate to mention some caveats. Just like in the building of ICE-PHI, Bautista (2004) has explained that only convenience sampling presented a feasible way of building corpus. With this, the samples

included in this corpus were the ones that are readily available to the language coordinators and those whose authors and publishers willingly resolve copyright issues. For instance, some of the Hiligaynon texts came from the language coordinator's own publications. Hence, there is no claim at all of representativeness of samples included in this corpus.

The orthography utilized by the texts is also preserved. If some bible verses exhibit antiquated language or Spanish orthography, such characteristics were regarded as feature of the language in a particular period of time. The treatment and analysis of the data then lie in the hands of linguists who would be utilizing the corpus for their own researches.

In addition, the texts included in the corpus are authentic and the instances of code switches are therefore inevitable. There was no attempt to edit these utterances.

The FSL, on the other hand, provides basic terms, expressions and discourse in video form.

The mark-up of the texts does not guarantee 100% accuracy, too. Although the language coordinators have diligently proofread the texts, there are still possibilities of errors. Even with these shortcomings, the corpus still promises to be significant and can be a solid basis for any linguistic investigation or analysis on Philippine languages.

3 The Technical Aspect

3.1 Corpora Data

The project director will oversee the five projects, one for each language (Tagalog, Cebuano, Ilocano and Hiligaynon), one for the Filipino sign language, and one for the technical project. For each of the languages (including the Filipino sign language), the project coordinator is responsible for the collection, validation, and uploading of the documents into the Philippine corpus portal. The development of the website and the setting of the standards and formats for the documents will be set by the technical project coordinator.

3.2 Online Repository

A very important aspect of data collection is storing data and keeping track of it. Thanks to new technologies, digitizing data allows a lot of functionalities that make storing and tracking data easier. Moreover, through the power of the internet, data is made even more accessible to anyone around the world. The purpose of the online repository is to use these technologies as leverage in the data collection process. Figure 1 shows the screen shot of the online repository called *PALITO*.

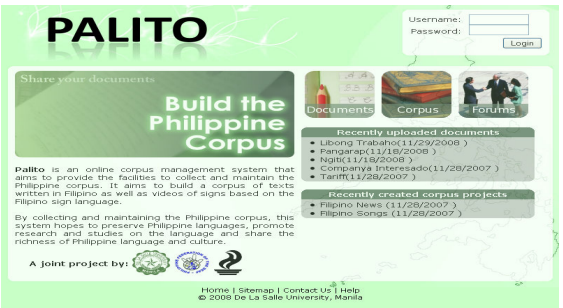

Figure 1: Screenshot of Palito's front Page

The online repository allows its users to submit their documents into the repository for storage. The repository automatically indexes these documents so they can easily be tracked.

Users of the repository can now look for specific documents using its internal search engine and most importantly use different linguistic tools that are very useful for language research.

One available tool is the document search feature which allows users to look for specific documents in the repository. A screenshot is shown in Figure 2 where all songs in the repository are listed.

Figure 2: Searching for documents in Palito

Another feature is the word frequency feature which displays the number of word occurrences in a single document or a group of documents. This is shown in Figure 3 where all words containing "ng" is listed.

Figure 3: Looking at word frequencies in a document

Lastly, one of the most important tools in the repository is the concordancer. It generates a concordance, which is list of the occurrences of a specified word in a document or a group of documents, with its immediate context. Figure 4 shows an example of a concordancer search for the word "puso" (heart) on a document. As seen, it lists all occurrences of the word its corresponding immediate context.

Figure 4: Using concordancer on a document

These will not be the only tools present in the repository since more tools will be developed in the future to make research and analysis much easier. They also serve as incentive for users in building the online repository by submitting their documents, as it becomes easier for them to analyze their own documents apart from access to documents submitted by other users.

3.3 Automated Tools

A concordancer is a software the automatically constructs a concordance, which is an alphabetical list of the principal words used in a book or body of work, with their immediate contexts. Existing concordancers include AntConc, ApSIC Xbench, WordSmith, MonoConc, GlossaNet, and CorpusEye. To illustrate the use of a concordance, let's consider AntConc. AntConc is a concordancer for Windows, Mac OS X, and Linux systems developed by Laurence Anthony of Waseda University, Japan. AntConc can generate "keyword-in-context" concordance lines and concordance distribution plots. It also has tools to analyze word clusters (lexical bundles), n-grams, collocates, word frequencies, and keywords. Although the AntConc was originally designed for use in classrooms, it has a powerful set of tools that are useful to researchers, including wildcard and regular expression searches. One useful feature of the program is the ability to process texts in almost any language in the world, including Asian languages, such as Chinese, Japanese, and Korean.

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

4 Conclusions and Recommendations

This paper has presented a detailed description of the project, "Online corpora of Philippine languages." The overview, scope and limitations, text types and categories, method of data collection, including the caveats were provided in order to point out the intricacies of corpus-building projects. With these, directions for future studies can be made. As clearly established, the present project is an introductory project towards the building of PALITO, an online repository of Philippine languages. It is therefore recommended that similar corpus be built for the other Philippine languages; first, the other major languages, then eventually, the minor Philippine languages.

In addition, a corpus cannot be called such without the spoken component of it. It is therefore highly recommended that spoken data be included in the corpus building.

The inclusion of as many Philippine languages is the ultimate objective of this corpus-building project. This goal can only be attained with the cooperation and involvement of more linguists and researchers who are willing to document as many Philippine languages as possible.

References

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

Biber, D. 1993. Representativeness in corpus design. *Literary and Linguistic Computing, 8(4),* 243-257.

Biber, D., S. Conrad, and R. Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Constantino, E. 1971. Tagalog and other major languages of the Philippines. In T. A. Sebeok, ed., *Current Trends in Linguistics, Vol. 8: Linguistics in Oceania*, pp.112-154. The Hague and Paris: Mouton.

Daguman, J. 2004. *A grammar of Northern Subanen.* Ph.D. thesis, Le Trobe University, Australia.

Davis, P, J. Baker, W. Spitz and M. Baek. 1998. *The grammar of Yogad: A functional explanation*. LINCOM Studies in Austronesian Linguistics 01. München and Newcastle: Lincom Europa.

Dita, S. N. 2007. *A reference grammar of Ibanag*. Ph.D. thesis, De La Salle University, Philippines.

Engwall, G. 1992. Comments. In J. Svartvik, ed., *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 1992,* pp. 164-169. New York and Berlin: Mouton de Gruyter.

Francis, G. 1982. Problems of assembling and computerizing large corpora. In S. Johansson, ed., *Computer corpora in English language research,* (7-24). Bergen: Norwegian Computer Center for the Humanities.

Gordon, R. G., Jr., ed. 2005. *Ethnologue: Languages of the World*, Fifteenth edition. Dallas,Texas: SIL International. Online version: http://www.ethnologue.com/

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

Liao, H. 2004. Transitivity and ergativity in Formosan and Philippine languages. Ph. D. thesis, University of Hawaii at Manoa.

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

McKaughan, H. P. 1971. Minor languages of the Philippines. In T. A. Sebeok, ed., *Current Trends in Linguistics*, *Vol. 8: Linguistics in Oceania*, pp.155-167. The Hague and Paris: Mouton.

Meyer, C. F. 2002. *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press.

Nelson, G. 1996. Markup systems. In S. Greenbaum, ed., *Comparing English worldwide: The International Corpus of English,* pp. 36-53. Oxford: Oxford University Press.

Quakenbush, J. 2005. Philippine linguistics from an SIL perspective: Trends and prospects. In Hsiu-chuan Liao and Carl R. Galvez Rubino (eds.), *Current issues in Philippine linguistics and anthropology: Parangal kay Lawrence A. Reid*, 3-27. Manila: Linguistic Society of the Philippines and SIL Philippines.

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

Rubino, C. R. G. 1997. *A Reference Grammar of Ilocano*. PhD thesis, University of California, Santa Barbara.

Ruffolo, Roberta. 2004. *Topics in the morpho-syntax of Ibaloy, Northern Philippines*. Ph.D. thesis, Australian National University.