# Philippine Language Resources: Trends and Directions

**Rachel Edita Roxas** **Charibeth Cheng** **Nathalie Rose Lim**
Center for Language Technologies
College of Computer Studies
De La Salle University
2401 Taft Ave, Manila, Philippines
rachel.roxas, chari.cheng, nats.lim@delasalle.ph

## Abstract

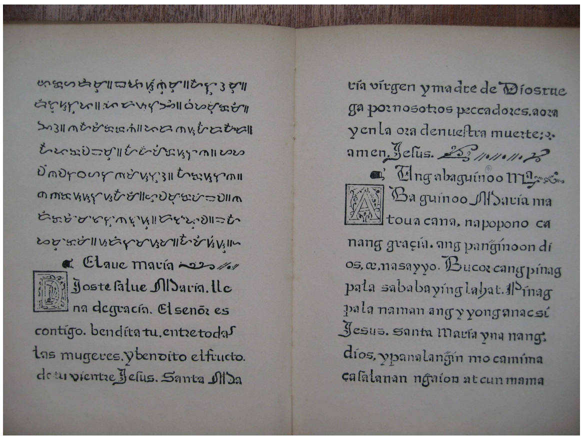
We present the diverse research activities on Philippine languages from all over the country, with focus on the Center for Language Technologies of the College of Computer Studies, De La Salle University, Manila, where majority of the work are conducted. These projects include the formal representation of Philippine languages and the processes involving these languages. Language representation entails the manual and automatic development of language resources such as lexicons and corpora for various human languages including Philippine languages, across various forms such as text, speech and video files. Tools and applications on languages that we have worked on include morphological processes, part of speech tagging, language grammars, machine translation, sign language processing and speech systems. Future directions are also presented.

## 1 Introduction

The Philippines is an archipelagic nation in Southeast Asia with more than 7,100 islands with 168 natively spoken languages (Gordon, 2005). These islands are grouped into three main island groups: Luzon (northern Philippines), Visayas (central) and Mindanao (southern), and various Philippine languages distributed among its islands.

Little is known historically about these languages. The most concrete evidence that we have is the Doctrina Christiana, the first ever published work in the country in 1593 which contains the translation of religious material in the local Philippine script (the Alibata), Spanish and old Tagalog (a sample page is shown in Figure 1, courtesy of the University of Sto. Tomas Library, 2007). Alibata is an ancient Philippine script that is no longer widely used except for a few locations in the country. The old Tagalog has evolved to the new Filipino alphabet which now consists of 26 letters of the Latin script and "ng" and "ñ".

The development of the national language can be traced back to the 1935 Constitution Article XIV, Section 3 which states that "...Congress shall make necessary steps towards the development of a national language which will be based on one of the existing native languages..." due to the advocacy of then Philippine President Manuel L. Quezon for the development of a national language that will unite the whole country. Two years later, Tagalog was recommended as the basis of the national language, which was later officially called *Pilipino*. In the 1987 Constitution, Article XIV, Section 6, states that "the National language of the Philippines is Filipino. As it evolves, it shall be further developed and enriched on the basis of existing Philippine and other languages." To date, Filipino that is being taught in our schools is basically Tagalog, which is the predominant language being used in the archipelago.[1]



El aue maría
Dios te ſalue Maria. lle
na de gracia. El señor es
contigo. bendita tu, entre todas
las mugeres. y bendito el fructo.
de tu vientre Jeſus. Santa Ma

ria virgen y madre de Dios rue
ga por noſotros peccadores. aora
y en la ora de nueſtra muerte:
amen. Jeſus.
Ang abaguinoo Ma
Aba guinoo Maria ma
toua cana, napopono ca
nang gracia. ang panginoon di
os, ꝯ. nasa yyo. Bucor cang pinag
pala sa babaying lahat. Pinag
pala naman ang y yong anac si
Jesus. santa Maria yna nang
dios, ypanalangin mo cami ma
caſalanan ngaion at cun mama

Figure 1: Alibata, Spanish and Old Tagalog sample page: Doctrina Christiana (courtesy of the University of Sto. Tomas Library, 2007)

Table 1 presents data gathered through the 2000 Census conducted by the National Statistics

[1] Thus, when we say Filipino, we generally refer to Tagalog.

*Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009*, pages 131–138,
Suntec, Singapore, 6-7 August 2009.

Office, Philippine government, on the Philippine languages that are spoken by at least one percent of the population.

| Languages | Number of native speakers |
|---|---|
| Tagalog | 22,000,000 |
| Cebuano | 20,000,000 |
| Ilokano | 7,700,000 |
| Hiligaynon | 7,000,000 |
| Waray-Waray | 3,100,000 |
| Capampangan | 2,900,000 |
| "Northern Bicol" | 2,500,000 |
| Chavacano | 2,500,000 |
| Pangasinan | 1,540,000 |
| "Southern Bicol" | 1,200,000 |
| Maranao | 1,150,000 |
| Maguindanao | 1,100,000 |
| Kinaray-a | 1,051,000 |
| Tausug | 1,022,000 |
| Surigaonon | 600,000 |
| Masbateño | 530,000 |
| Aklanon | 520,000 |
| Ibanag | 320,000 |

Table 1. Philippine languages spoken by least 1% of the population.

Linguistics information on Philippine languages are available, but as of yet, the focus has been on theoretical linguistics and little is done about the computational aspects of these languages. To add, much of the work in Philippine linguistics focused on the Tagalog language (Liao, 2006, De Guzman, 1978). In the same token, NLP researches have been focused on Tagalog, although pioneering work on other languages such as Cebuano, Hiligaynon, Ilocano, and Tausug have been made. As can be noticed from this information alone, NLP research on Philippine languages is still at its infancy stage.

One of the first published works on NLP research on Filipino was done by Roxas (1997) on IsaWika!, a machine translation system involving the Filipino language. From then on most of the NLP researches have been conducted at the Center for Language Technologies of the College of Computer Studies, De La Salle University, Manila, Philippines, in collaboration with our partners in academe from all over the country. The scope of experiments have expanded from north of the country to south, from text to speech to video forms, and from present to past data.

NLP researches address the manual construction of language resources literally built from almost non-existent digital forms, such as the grammar, lexicon, and corpora, augmented by some automatic extraction algorithms. Various language tools and applications such as machine translation, information extraction, information retrieval, and natural language database interface, were also pursued. We will discuss these here, the corresponding problems associated with the development of these projects, and the solutions provided. Future research plans will also be presented.

## 2 Language Resources

We report here our attempts in the manual construction of Philippine language resources such as the lexicon, morphological information, grammar, and the corpora which were literally built from almost non-existent digital forms. Due to the inherent difficulties of manual construction, we also discuss our experiments on various technologies for automatic extraction of these resources to handle the intricacies of the Filipino language, designed with the intention of using them for various language technology applications.

We are currently using the English-Filipino lexicon that contains 23,520 English and 20,540 Filipino word senses with information on the part of speech and co-occurring words through sample sentences. This lexicon is based on the dictionary of the Commission on the Filipino Language (Komisyon sa Wikang Filipino), and digitized by the IsaWika project (Roxas, 1997). Additional information such as synsetID from Princeton WordNet were integrated into the lexicon through the AeFLEX project (Lim, *et al.*, 2007b). As manually populating the database with the synsetIDs from WordNet is tedious, automating the process through the SUMO (Suggested Upper Merged Ontology) as an InterLingua is being explored to come up with the Filipino WordNet (Tan and Lim, 2007).

Initial work on the manual collection of documents on Philippine languages has been done through the funding from the National Commission for Culture and the Arts considering four major Philippine Languages namely, Tagalog, Cebuano, Ilocano and Hiligaynon with 250,000 words each and the Filipino sign language with 7,000 signs (Roxas, *et al.*, 2009). Computational features include word frequency counts and a concordancer that allows viewing co-occurring words in the corpus. Mark-up conventions followed some of those used for the ICE project.

Aside from possibilities of connecting the Philippine islands and regions through language, crossing the boundaries of time are one of the goals (Roxas, 2007a; Roxas, 2007b). An unexplored but equally challenging area is the collection of historical documents that will allow research on the development of the Philippine languages through the centuries, one of which is the already mentioned Doctrina Christiana which was published in 1593.

Attempts are being made to expand on these language resources and to complement manual efforts to build these resources. Automatic methods and social networking are the two main options currently being considered.

### 2.1 Language Resource Builder

Automatic methods for bilingual lexicon extraction, named-entity extraction, and language corpora are also being explored to exploit on the resources available on the internet. These automatic methods are discussed in detail in this section.

An automated approach of extracting bilingual lexicon from comparable, non-parallel corpora was developed for English as the source language and Tagalog as the target language (Tiu and Roxas, 2008). The study combined approaches from previous researches which only concentrated on context extraction, clustering techniques, or usage of part of speech tags for defining the different senses of a word, and ranking has shown improvement to overall F-measure from 7.32% to 10.65% within the range of values from previous studies. This is despite the use of limited amount of corpora of 400k and seed lexicon of 9,026 entries in contrast to previous studies of 39M and 16,380, respectively.

The NER-Fil is a Named Entity Recognizer for Filipino Text (Lim, *et al.*, 2007a). This system automatically identifies and stores named-entities from documents, which can also be used to annotate corpora with named-entity information. Using machine learning techniques, named entities are also automatically classified into appropriate categories such as person, place, and organization.

AutoCor is an automatic retrieval system for documents written in closely-related languages (Dimalen and Roxas, 2007). Experiments have been conducted on four closely-related Philippine languages, namely: Tagalog, Cebuano and Bicolano. Input documents are matched against the n-gram language models of relevant and irrelevant documents. Using common word pruning to differentiate between the closely-related Philippine languages, and the odds ratio query generation methods, results show improvements in the precision of the system.

Although automatic methods can facilitate the building of the language resources needed for processing natural languages, these automatic methods usually employ learning approaches that would require existing language resources as seed or learning data sets.

### 2.2 Online Community for Corpora Building

PALITO is an online repository of the Philippine corpus (Roxas, *et al.*, 2009). It is intended to allow linguists or language researchers to upload text documents written in any Philippine language, and would eventually function as corpora for Philippine language documentation and research. Automatic tools for data categorization and corpus annotation are provided by the system. The LASCOPHIL (La Salle Corpus of Philippine Languages) Working Group is assisting the project developers of PALITO in refining the mechanics for the levels of users and their corresponding privileges for a manageable monitoring of the corpora. Videos on the Filipino sign language can also be uploaded into the system. Uploading of speech recordings will be considered in the near future, to address the need to employ the best technology to document and systematically collect speech recordings especially of nearly extinct languages in the country. This online system capitalizes on the opportunity for the corpora to expand faster and wider with the involvement of more people from various parts of the world. This is also to exploit on the reality that many of the Filipinos here and abroad are native speakers of their own local languages or dialects and can largely contribute to the growth of the corpora on Philippine languages.

## 3 Language Tools

Language tools are applications that support linguistic research and processing of various language computational layers. These include lexical units, to syntax and semantics. Specifically, we have worked on the morphological processes, part of speech tagging and parsing. These processes usually employ either the rule-based approach or the example-based approach. In general, rule-based approaches capture language processes by formally capturing these processes which would require consultations and inputs from linguists. On the other hand, example-based approaches employ machine learning methodologies where automatic learning of rules is performed based on manually annotated data that are done also by linguists.

### 3.1 Morphological Processes

We have tested both rule-based and example-based approaches in developing our morphological analyzers and generators. Rule-based morphological analysis in the current methods, such as finite-state and unification-based, are predominantly effective for handling concatenative morphology (e.g. prefixation and suffixation), although some of these techniques can also handle limited non-concatenative phenomena (e.g. infixation and partial and full-stem reduplication) which are largely used in Philippine languages. TagMA (Fortes-Galvan and Roxas, 2007) uses a constraint-based method to perform morphological analysis that handles both concatenative and non-concatenative morphological phenomena, based on the optimality theory framework and the two-level morphology rule representation. Test results showed 96% accuracy. The 4% error is attributed to d-r alteration, an example of which is in the word *lakaran*, which is from the root word *lakad* and suffix *-an*, but *d* is changed to *r*. Unfortunately, since all candidates are generated, and erroneous ones are later eliminated through constraints and rules, time efficiency is affected by the exhaustive search performed.

To augment the rule-based approach, an example-based approach was explored by extending Wicentowski's Word Frame model through learning of morphology rules from examples (Cheng and See, 2006). In the WordFrame model, the seven-way split re-write rules composed of the canonical prefix/beginning, point-of-prefixation, common prefix substrings, internal vowel change, common suffix substring, point-of-suffixation, and canonical suffix/ending. Infixation, partial and full reduplication as in Tagalog and other Philippine languages are improperly modeled in the WordFrame model as point-of-prefixation as in the word *(hin)-intay* which should have been modeled as the word *hintay* with infix *–in-*. Words with an infix within a prefix are also modeled as point-of-prefixation as in the word *(hini-)hintay* which should be represented as infix *–in* in partial reduplicated syllable *hi-*. In the revised WordFrame model (Cheng and See, 2006), the non-concatenative Tagalog morphological behaviors such as infixation and reduplication are modeled separately and correctly. Unfortunately, it is still not capable of fully modeling Filipino morphology since some occurrences of reduplication are still represented as point-of-suffixation for various locations of the longest common substring. There are also some problems in handling the occurrence of several partial or whole-word reduplications within a word. Despite these problems, the training of the algorithm that learns these re-write rules from 40,276 Filipino word pairs derived 90% accuracy when applied to an MA. The complexity of creating a better model would be computationally costly but it would ensure an increase in performance and reduced number of rules.

Work is still to be done on exploring techniques and methodologies for morphological generation (MG). Although it could be inferred that the approaches for MA can be extended to handle MG, an additional disambiguation process is necessary to choose the appropriate output from the many various surface form of words that can be generated from one underlying form.

### 3.2 Part of Speech Tagging

One of the most useful information in the language corpora are the part of speech tags that are associated with each word in the corpora. These tags allow applications to perform other syntactic and semantic processes. Firstly, with the aid of linguists, a revised tagset for Tagalog has been formulated (Miguel and Roxas, 2007), since a close examination of the existing tagset for languages such as English showed the insufficiency of this tagset to handle certain phenomena in Philippine languages such as the lexical markers, ligatures and enclitics. The lexical marker *ay* is used in inverted sentences such as *She is good* (*Siya ay mabuti*). Ligatures can take the form of the word *na* or suffixes *–ng* (*-g*), the former is used if the previous noun, pronoun or adjective ends with a consonant (except for *n*), and the latter if the previous word ends with a vowel (or *n*).

Manual tagging of corpora has allowed the performance of automatic experiments on some approaches for tagging for Philippine languages namely MBPOST, PTPOST4.1, TPOST and TagAlog, each one exploring on a particular approach in tagging such as memory-based POS, template-based and rule-based approaches. A study on the performance of these taggers showed accuracies of 85, 73, 65 and 61%, respectively (Miguel and Roxas, 2007).

### 3.3 Language Grammars

Grammar checkers are some of the applications where syntactic specification of languages is necessary. SpellCheF is a spell checker for Filipino that uses a hybrid approach in detecting and correcting misspelled words in a document (Cheng, *et al.*, 2007). Its approach is composed of dictionary-lookup, n-gram analysis, Soundex and character distance measurements. It is implemented as a plug-in to OpenOffice Writer. Two spelling rules and guidelines, namely, the Komisyon sa Wikang Filipino 2001 Revision of the Alphabet and Guidelines in Spelling the Filipino Language, and the Gabay sa Editing sa Wikang Filipino rulebooks, were incorporated into the system. SpellCheF consists of the lexicon builder, the detector, and the corrector; all of which utilized both manually formulated and automatically learned rules to carry out their respective tasks.

FiSSAn, on the other hand, is a semantics-based grammar checker. Lastly, PanPam (Jasa, *et al.*, 2007) is an extension of FiSSAn that also incorporates a dictionary-based spell checker (Borra, *et al.*, 2007).

These systems make use of the rule-based approach. To complement these systems, an example-based approach is considered through a grammar rule induction method (Alcantara and Borra, 2008). Constituent structures are automatically induced using unsupervised probabilistic approaches. Two models are presented and results on the Filipino language show an F1 measure of greater than 69%. Experiments revealed that the Filipino language does not follow a strict binary structure as English, but is more right-biased.

A similar experiment has been conducted on grammar rule induction for the automatic parsing of the Philippine component of the International Corpus of English (ICE-PHI) (Flores and Roxas, 2008). The ICE-PHI corpora consist of English texts with indigenous words and phrases during speaker context switching. The markup language followed the standards specified by the ICE group, which is headed by ICE-GB. Constituent rule induction is performed from manually encoded syntactically bracketed files from the ICE-PHI, and will be used to parse the rest of the corpus. Manual post editing of the parse will be performed. The development of such tools will directly benefit the descriptive and applied linguistics of Philippine English, as well as other Englishes, in particular, those language components in the ICE.

Various applications on Philippine languages have been created at the Center for Language Technologies, College of Computer Studies, De La Salle University to cater to different needs.

## 4 Language Applications

### 4.1 Machine Translation

The Hybrid English-Filipino Machine Translation (MT) System is a three-year project (with funding from the PCASTRD, DOST), which involves a multi-engine approach for automatic language translation of English and Filipino Roxas, *et al.*, 2008). The MT engines explore on approaches in translation using a rule-based method and two example-based methods. The rule-based approach requires the formal specification of the human languages covered in the study and utilizes these rules to translate the input. The two other MT engines make use of examples to determine the translation. The example-based MT engines have different approaches in their use of the examples (which are existing English and Filipino documents), as well as the data that they are learning.

The system accepts as input a sentence or a document in the source language and translates this into the target language. If source language is English, the target language is Filipino, and vise versa. The input text will undergo preprocessing that will include POS tagging and morphological analysis. After translation, the output translation will undergo natural language generation including morphological generation. Since each of the MT engines would not necessarily have the same output translation, an additional component called the Output Modeler was created to determine the most appropriate among the translation outputs (Go and See, 2008). There are ongoing experiments on the hybridization of the rule-based and the template-based approaches where transfer rules and unification constraints are derived (Fontanilla and Roxas, 2008).

The rule-based MT builds a database of rules for language representation and translation rules from linguists and other experts on translation from English to Filipino and from Filipino to English. We have considered lexical functional grammar (LFG) as the formalism to capture these rules. Given a sentence in the source language, the sentence is processed and a computerized representation in LFG of this sentence is constructed. An evaluation of how comprehensive and exhaustive the identified grammar is will be considered. Is the system able to capture all possible Filipino sentences? How are all possible sentences to be represented since Filipino exhibits some form of free word order in sentences? The next step is the translation step, that is, the conversion of the computerized representation of the input sentence into the intended target language. After the translation process, the computerized representation of the sentence in the target language will now be outputted into a sentence form, or called the generation process. Although it has been shown in various studies elsewhere and on various languages that LFG can be used for analysis of sentences, there is still a question of whether it can be used for the generation process. The generation involves the outputting of a sentence from a computer-based representation of the sentence. This is part of the work that the group intends to address.

The major advantage of the rule-based MT over other approaches is that it can produce high quality translation for sentence patterns that were accurately captured by the rules of the MT engine; but unfortunately, it cannot provide good translations to any sentence that go beyond what the rules have considered.

In contrast to the rule-based MT which requires building the rules by hand, the corpus-based MT system automatically learns how translation is done through examples found in a corpus of translated documents. The system can incrementally learn when new translated documents are added into the knowledge-base, thus, any changes to the language can also be accommodated through the updates on the example translations. This means it can handle translation of documents from various domains (Alcantara, *et al.*, 2006).

The principle of garbage-in-garbage-out applies here; if the example translations are faulty, the learned rules will also be faulty. That is why, although human linguists do not have to specify and come up with the translation rules, the linguist will have to first verify the translated documents and consequently, the learned rules, for accuracy.

It is not only the quality of the collection of translations that affects the overall performance of the system, but also the quantity. The collection of translations has to be comprehensive so that the translation system produced will be able to translate as much types of sentences as possible. The challenge here is coming up with a quantity of examples that is sufficient for accurate translation of documents.

With more data, a new problem arises when the knowledge-base grows so large that access to it and search for applicable rules during translation requires tremendous amount of access time and to an extreme, becomes difficult. Exponential growth of the knowledge-base may also happen due to the free word order nature of Filipino sentence construction, such that one English sentence can be translated to several Filipino sentences. When all these combinations are part of the translation examples, a translation rule will be learned and extracted by the system for each combination, thus, causing growth of the knowledge-base. Thus, algorithms that perform generalization of rules are considered to remove specificity of translation rules extracted and thus, reduce the size of the rule knowledge-base.

One of the main problems in language processing most especially compounded in machine translation is finding the most appropriate translation of a word when there are several meanings of source words, and various target word equivalents depending on the context of the source word. One particular study that focuses on the use of syntactic relationships to perform word sense disambiguation has been explored (Domingo and Roxas, 2006). It uses an automated approach for resolving target-word selection, based on “word-to-sense” and “sense-to-word” relationship between source words and their translations, using syntactic relationships (subject-verb, verb-object, adjective-noun). Using information from a bilingual dictionary and word similarity measures from WordNet, a target word is selected using statistics from a target language corpus. Test results using English to Tagalog translations showed an overall 64% accuracy for selecting word translation.

Other attempts on MT are on Tagalog to Cebuano (Yara, 2007), and Ilocano to English (Miguel and Dy, 2008). Both researches focus on building the language resources for the languages Cebuano and Ilocano, respectively, since focus on the Philippine languages have so far been on Tagalog. It is also important to note that these contributions are local researches being done where the languages are actually spoken and actively in usage.

## 4.2 Sign Language Processing

Most of the work that we have done focused on textual information. Recently, we have explored on video and speech forms.

With the inclusion of the Filipino sign language in a corpora building project (Roxas, *et al.*, 2009), video formats are used to record, edit, gloss and transcribe signs and discourse. Video editing merely cuts the video for final rendering, glossing allows association of sign to particular words, and transcription allows viewing of textual equivalents of the signed videos.

Work on the automatic recognition of Filipino sign language involves digital signal processing concepts. Initial work has been done on sign language number recognition (Sandjaja, 2008) using color-coded gloves for feature extraction. The feature vectors were calculated based on the position of the dominant-hand's thumb. The system learned through a database of numbers from 1 to 1000, and tested by the automatic recognition of Filipino sign language numbers and conversion into text. Over-all accuracy of number recognition is 85%.

Another proposed work is the recognition of non-manual signals focusing on the various parts of the face; in particular, initially, the mouth is to be considered. The automatic interpretation of the signs can be disambiguated using the interpretation of the non-manual signals.

### 4.3 Speech Systems

PinoyTalk is an initial study on a Filipino-based text to speech system that automatically generates the speech from input text (Casas, *et al.,* 2004). The input text is processed and parsed from words to syllables, from syllables to letters, and assigned prosodic properties for each one. Six rules for Filipino syllabication were identified and used in the system. A rule-based model for Filipino was developed and used as basis for the implementation of the system. The following were determined in the study considering the Filipino speaker: duration of each phoneme and silences, intonation, pitches of consonants and vowel, and pitches of words with the corresponding stress. The system generates an audio output and able to save the generated file using the mp3 or wav file format.

A system has been developed at the Digital Signal Processing Laboratory at the University of the Philippines at Diliman to automatically recognize emotions such as anger, boredom, happiness and satisfaction (Ebarvia, *et al.*, 2008).

## 5 Future Directions

Through the Center for Language Technologies of the College of Computer Studies, De la Salle University, and our partners, varied NLP resources have been built, and applications and researches explored. Our faculty members and our students have provided the expertise in these challenging endeavors, with multi-disciplinary efforts and collaborations. Through our graduate programs, we have trained many of the faculty members of universities from various parts of the country; thus, providing a network of NLP researchers throughout the archipelago. We have organized the National NLP Research Symposium for the past five years, through the efforts of the CeLT of CCS-DLSU, and through the support of government agencies such as PCASTRD, DOST and CHED, and our industry partners. Last year, we hosted an international conference (the 22nd Pacific Asia Conference on Language, Information and Computation) which was held in Cebu City in partnership with UPVCC and Cebu Institute of Technology. We have made a commitment to nurture and strengthen NLP researches and collaboration in the country, and expand on our international linkages with key movers in both the Asian region and elsewhere. For the past five years, we have brought in and invited internationally-acclaimed NLP researchers into the country to support these endeavors. Recently, we have also received invitations as visiting scholars, and participants to events and meetings within the Asean region which provided scholarships, which in turn, we also share with our colleagues and researchers in other Philippine universities.

It is an understatement to say that much has to be explored in this area of research that interleaves diverse disciplines among technology-based areas (such as NLP, digital signal processing, multi-media applications, and machine learning) and other fields of study (such as language, history, psychology, and education), and cuts across different regions and countries, and even time frames (Cheng, *et al*., 2008). It is multi-modal and considers various forms of data from textual, audio, video and other forms of information. Thus, much is yet to be accomplished, and experts with diverse backgrounds in these various related fields will bring this area of research to a new and better dimension.

## References

D. Alcantara and A. Borra. 2008. Constituent Structure for Filipino: Induction through Probabilistic Approaches. *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation (PACLIC).* 113-122.

D. Alcantara, B. Hong, A. Perez and L. Tan. 2006. *Rule Extraction Applied in Language Translation – R.E.A.L. Translation*. Undergraduate Thesis, De la Salle University, Manila.

A. Borra, M. Ang, P. J. Chan, S. Cagalingan and R. Tan. 2007. FiSSan: Filipino Sentence Syntax and Semantic Analyzer. *Proceedings of the 7th Philippine Computing Science Congress*. 74-78.

D. Casas, S. Rivera, G. Tan, and G. Villamil. 2004. *PinoyTalk: A Filipino Based Text-to-Speech Synthesizer*. Undergraduate Thesis. De La Salle University.

C. Cheng, R. Roxas, A. B. Borra, N. R. L. Lim, E. C. Ong and S. L. See. 2008. e-Wika: Digitalization of Philippine Language. *DLSU-Osaka Workshop*.

C. Cheng, C. P. Alberto, I. A. Chan and V. J. Querol. 2007. SpellChef: Spelling Checker and Corrector for Filipino. *Journal of Research in Science, Computing and Engineering.* 4(3), 75-82.

C. Cheng, and S. See. 2006. The Revised Wordframe Model for Filipino Language. *Journal of Research in Science, Computing and Engineering*. 3(2), 17-23.

D. Dimalen and R. Roxas. 2007. AutoCor: A Query-Based Automatic Acquisition of Corpora of Closely-Related Languages. *Proceedings of the 21st PACLIC*. 146-154.

E. Domingo and R. Roxas. 2006. Utilizing Clues in Syntactic Relationships for Automatic Target Word Sense Disambiguation. *Journal of Research for Science, Computing and Engineering.* 3(3), 18-24.

E. Ebarvia, M. Bayona, F. de Leon, M. Lopez, R. Guevara, B. Calingacion, and P. Naval, Jr. 2008. Determination of Prosodic Feature Set for Emotion Recognition in Call Center Speech. *Proceedings of the 5th National Natural Language Processing Research Symposium (NNLPRS)*. 65-71.

D. Flores and R. Roxas. 2008. Automatic Tools for the Analysis of the Philippine component of the International Corpus of English. *Linguistic Society of the Philippines Annual Meeting and Convention.*

G. Fontanilla and R. Roxas. 2008. A Hybrid Filipino-English Machine Translation System. *DLSU Science and Technology Congress*.

F. Fortes-Galvan and R. Roxas. 2007. Morphological Analysis for Concatenative and Non-concatenative Phenomena. *Proceedings of the Asian Applied NLP Conference*.

K. Go and S. See. 2008. Incorporation of WordNet Features to N-Gram Features in a Language Modeller. *Proceedings of the 22nd PACLIC*, 179-188.

Gordon, R. G., Jr. (Ed.). 2005. *Ethnologue: Languages of the World*, 5th Ed. Dallas,Texas: SIL International. Online version: www.ethnologue.com.

M. Jasa, M. J. Palisoc and J. M. Villa. 2007. *Panuring Panitikan (PanPam): A Sentence Syntax and Semantics-based Grammar Checker for Filipino.* Undergraduate Thesis. De La Salle University, Manila.

H. Liao. 2006. *Philippine linguistics: The state of the art (1981-2005)*. De La Salle University, Manila.

N. R. Lim, J. C. New, M. A. Ngo, M. Sy, and N. R. Lim. 2007a. A Named-Entity Recognizer for Filipino Texts. *Proceedings of the 4th NNLPRS.*

N. R. Lim, J. O. Lat, S. T. Ng, K. Sze, and G. D. Yu. 2007b. Lexicon for an English-Filipino Machine Translation System. *Proceedings of the 4th National Natural Language Processing Research Symposium*.

D. Miguel and M. Dy. 2008. Anglo-Cano: an Ilocano to English Machine Translation. *Proceedings of the 5th National Natural Language Processing Research Symposium*. 85-88.

D. Miguel and R. Roxas. 2007. Comparative Evaluation of Tagalog Part of Speech Taggers. Proceedings of the 4th *NNLPRS*.

R. Roxas, P. Inventado, G. Asenjo, M. Corpus, S. Dita, R. Sison-Buban and D. Taylan. 2009. Online Corpora of Philippine Languages. *2nd DLSU Arts Congress: Arts and Environment*.

R. Roxas, A. Borra, C. Ko, N. R. Lim, E. Ong, and M. W. Tan. 2008. Building Language Resources for a Multi-Engine Machine Translation System. Language Resources and Evaluation. Springer, Netherlands. 42:183-195.

R. Roxas. 2007a. e-Wika: Philippine Connectivity through Languages. *Proceedings of the 4th NNLPRS.*

R. Roxas. 2007b. Towards Building the Philippine Corpus. *Consultative Workshop on Building the Philippine Corpus*.

R. Roxas. 1997. Machine Translation from English to Filipino: A Prototype. *International Symposium of Multi-lingual Information Technology (MLIT '97)*, Singapore.

I. Sandjaja. 2008. *Sign Language Number Recognition*. Graduate Thesis. De La Salle University, Manila.

P. Tan and N. R. Lim. 2007. FILWORDNET: Towards a Filipino WordNet. *Proceedings of the 4th NNLPRS.*

E. P. Tiu and R. Roxas. 2008. Automatic Bilingual Lexicon Extraction for a Minority Target Language, *Proceedings of the 22nd PACLIC.* 368-376.

J. Yara. 2007. A Tagalog-to-Cebuano Affix-Transfer-Based Machine Translator. *Proceedings of the 4th NNLPRS.*