Reflections on the Awareness and Progress of Natural Language Processing (NLP) Research in the Philippines

Rodolfo C. Raga, Jr.
Jose Rizal University
80 Shaw Blvd., Mandaluyong City Philippines
rodolfo.raga@jru.edu

ABSTRACT

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. The number of top tech companies pouring funds into NLP research as well as the growing number of reputable foreign universities offering NLP courses makes it evident that NLP plays a crucial role in the business arena today. At the same time, if harnessed and properly directed, NLP technologies can be used as an instrument in addressing the language problems that afflict most multilingual countries. However, despite these potentials, in the Philippines, it can be observed that the level of awareness of students in this field of research seemingly remains low. A low level of awareness can potentially have inhibiting effects on the progress of this field of research. This article attempts to provide a thumbnail sketch of this situation. The discussion is supported by results from an online survey of students from various HEIs and by looking at the breadth and attributes of research papers presented to the annual National Natural Language Processing Research Symposium (NNLPRS). The challenges of promoting awareness in NLP research, the roles of NLP advocates within educational institutions, as well as the effects of research facilities that support NLP research are some of the issues discussed in this paper.

Keywords

Natural Language Processing, Awareness, Education, Research.

1. INTRODUCTION

Very briefly, Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things (Chowdhury, 2003). NLP research primarily focuses on gathering knowledge on how people use and understand their natural language so that appropriate tools and techniques can be developed that enable computer systems to mimic this level of understanding and allow them to manipulate languages to perform useful tasks.

The rising interest in this technology along with its ever increasing application in various domains indicates that a review of its status and progress is necessary. Studies assessing the progress of NLP research had already been conducted in the past. Jones (2001) for example, reviewed the history of NLP seeking to identify successive trends that reflect concerns with different problems or the development of approaches for solving NLP problems. The review served to provide a marker of where the NLP research field stands after more than four decades of efforts. Eisele & Ziegler-Eisele (2002) summarized contributions and discussions from two language technology workshops to be able to present some visions of NLP-related applications and to sketch milestones that may help to measure the progress towards developing these applications. Mote (2012) focused on describing the methods and processes of recent NLP literature in order to analyze some of the challenges that researchers are faced with regards to implementing a functional language-understanding machine. Cambria & White (2014) looked into the different schools of thought of NLP research in order to provide insights on the evolution of current NLP technologies and suggests near future research directions. Gambäck, et al (2005) described the evaluation of an NLP course, as well as the performance of the students to provide a sketch of his experiences on introducing NLP in Ethiopia. Most recently, Kurian (2014) conducted a survey of the progress made in this field in order to explain the difficulties and challenges faced by the NLP research community in India.

Given this background, relative to students, and, in the context of a developing and multilingual country like the Philippines, this article will attempt to contribute further by presenting some perspective towards the status of NLP research in the country.

In this regard, this article has the following objectives: (1) introduce the role and importance of NLP and identify motivations as to why Filipino students should be concerned about it; (2) assess the current status of awareness among students with regards this technology; and (3) examine research that has been done previously on NLP in order to identify issues in the progress of research on this technology.

2. THE ROLE OF NLP TECHNOLOGY

The multitude of reasons why Filipino students need to be aware and immerse themselves in NLP research can probably be summarized into two distinct and critical aspects.

1. The vital role that NLP technology plays in the current business arena, and
2. The huge potential of NLP technology as an instrument for addressing the language problems in the country.

2.1 Importance of NLP in the business arena

Short of asking each and every IT-based company in the industry, the critical role that NLP plays in the business arena can be implicitly felt if we take a look at the sort of research projects and products that IT companies currently engage themselves with (Church & Rau, 1995). Table 1 for example, lists some of the biggest names in the IT industry (in alphabetical order) and the NLP-based research projects/products that they are currently

Philippine Computing Journal Dedicated Issue on Natural Language Processing, pages 1–9
Vol. XI No. 1 August 2016

funding/developing/promoting. These companies pour in huge amounts of investments into their NLP-based research and products. If these big companies are focusing this much effort on NLP, it is certain that other companies are also lined-up to get a slice of the pie.

Table 1: NLP-based products and research projects of big IT companies

	Company	NLP Category	Research Projects/Products
1	Apple	Speech Recognition	Siri
2	DELL	Data Analytics	Social Net Analytics Pulse (SNAP)
3	Facebook	Speech Recognition	WIT.AI and Graph Search
4	Google	Information Retrieval	Hummingbird and Majel
5	HP	Data Analytics	Autonomy and Vertica
6	IBM	Question & Answering	Watson
7	Intel	Natural Language Recognition	IndiSys
8	Microsoft	Lexical Resource Builder	MindNet and NLPWin
9	Sun MicroSystem	Language Translation	Translation Editor
10	Yahoo	Data Analytics	SkyPhrase

To this end, the first important question to ask is: Why are big companies, even those whose line of expertise seem to fall under hardware manufacturing, suddenly became interested in developing NLP-based products and applications? A big chunk of the answer lies in the advent of the Social Web (Cambria & White, 2014). The introduction of this Internet phenomenon in the last few decades has led to the progressive growth of online activities. This trend is not only applied to individual users; various organizations, enterprises, companies, and corporations also joined in to migrate and/or offer their services, products, and business over the World Wide Web (WWW). The effect of such migration is that the amount of naturally occurring texts being generated over the web also skyrocketed (examples include customer feedback, competitor information, client emails, tweets, press releases, legal filings, and product/engineering documents). These large amounts of data represent a very valuable information resource for business organizations. Because not only can they be used as an input for guiding strategic decisions, but, more importantly, they can be processed and used to provide perspectives on the unique profiles and preferences of the body of customers of each business establishment (Žižka & Dařena, 2013). However, much of this data is extremely unstructured and its huge size makes it unsuitable for direct human processing and consumption (Grimes, 2005). This is where Natural Language Processing (NLP) becomes useful (Mote, 2012). NLP is a form of technology that can be adapted both as a mechanism for processing this huge amount of data; allowing either efficient knowledge discovery of information (Larson & Watson, 2013) and/or enabling a more natural human-computer interface. It can also be adapted to tap through the enormous volume of unstructured texts in any language.

A follow-up question that can be asked is: Why do students need to be concerned about businesses' interest in NLP? A foremost response would be because maintenance of their NLP-based products, projects, and capabilities would also require these companies to maintain a workforce with solid knowledge and background in NLP. This scenario suggests that one of the most critical employability skills that companies will be looking for in their future employees is the ability to develop and work with NLP applications and/or conduct NLP oriented research. This trend is not only observable among big IT companies. Samsung and Sony, for example, both highly successful electronics and smart appliance companies, have already started looking for applicants with research and development experience oriented towards NLP (See figure 1).

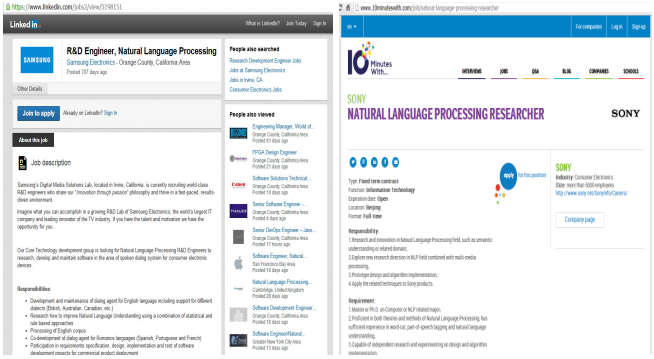


Figure 1: Job Ads from Sony and Samsung for applicants with NLP background

This implies that NLP could serve as a jump-off point for any student wanting to open career opportunities not only in the IT industry but in all related industries as well. In fact, the possibility of this implication has already permeated the air of education in many developed countries. The website www.quora.com[1], for example, listed more than two dozen reputable U.S. universities offering NLP course along with several other universities from the United Kingdom, Hongkong, Singapore, Germany, and Canada. The number of reputable universities offering NLP courses makes it evident that NLP is a crucial technology that students need to be acquainted with.

2.2. The Role of NLP in addressing the language problems in the country.

Another point to consider when considering the importance of NLP is its potential as an instrument in addressing the language problems in our country. Once again, there are two aspects that students need to be made aware of with regards this problem:

(1) The so-called language barrier problem, and

(2) The extinction of languages in the Philippines.

[1] https://www.quora.com/What-are-the-best-schools-for-studying-natural-language-processing

2.1.1 Language Barrier Problem

When talking about the language barrier, we refer to situations where people find it difficult to communicate with each other because they are using different languages. In the Philippines, the language barrier situation was clearly observed by Dr. Andrew Gonzales when he wrote in his paper:

> *"the Philippine state and Philippine society have not developed enough to be a 'crystallized' nation or a unified culture, manifested in part by the failure of the society really to consider the national language for purposes beyond the symbolic" (Gonzalez, 2003)*

The fact that each geographic region has its own mother tongue generates a strong ethnic loyalty among Filipinos to their own region. This, in turn, generates the language barrier problem. One effect of this problem is that the government now struggles with the problem of inefficient communication with its citizenry (Martin, 2012), thus building a wall between leaders and constituents. Professor Renato Constantino (1970) wrote about this situation when he said:

> "*People don't even think it is their duty to know or that they are capable of understanding the national problems. Because of the language barrier, they (the masses) are content to leave everything to their leaders. This is one of the root causes of their apathy, their regionalism or parochialism."*

As technology can play a critical role in the development of any country (Bogliacino, Perani, Pianta, & Supino, 2009), the involvement with emerging technologies like NLP can have a boosting effect on the status of our country as it can be used to promote better citizenship and government by minimizing hindrances caused by language barriers (Kurian, 2014). The lack of qualified people trained in this technology, therefore, can be a severe obstacle to the progress of this cause since a prerequisite to the creation of applications under each specific NLP field is the education and training of computer professionals skilled in the localization and development of language processing resources (Space, 2001; Daoud & El-Seoud, 2010).

2.1.2 Language Extinction Problem

Besides having many languages separated by wide regional boundaries, linguists note that the Philippines are also experiencing a sort of language convergence where high levels of borrowing between languages causes some languages to be abandoned and altogether become extinct (McFarland, 2004). An extinct language is a language that no longer has any speakers, or that is no longer in current use. According to experts, unless urgent measures are adopted, the Philippines are bound to witness a high rate of language extinction (Molina, 2012). This phenomenon can be observed not only in minority languages (Headland, 2003) but even in major Filipino languages as well (Anderson & Anderson, 2007).

NLP technology can provide tools that can contribute in slowing down the extinction of languages by enabling more efficient documentation and translation of endangered languages (Ptaszynski, Kazuki, & Momouchi, 2013). In the Philippines, efforts in applying NLP technology to local languages are already on-going (Roxas, 2010; Roxas, Alcantara, & Borlongan, 2010; Roxas, Cheng, & Lim, 2009), however, documenting and describing endangered languages are far too challenging that more works and contributions are still needed.

3. METHOD

With the above discussed issues in mind, the aim of this article is to draw a thumbnail sketch of the status of NLP research in the Philippines. This objective can be achieved by way of examining a two-fold picture: first, the level of awareness of NLP technology among students of various universities in the Philippines and second, the progress of NLP-based research activities conducted in the last twelve years as reflected in the research papers submitted to the annual National Natural Language Processing Research Symposium (NNLPRS).

As the starting point of this work, a pilot electronic survey was prepared for distribution to students of various Higher Education Institutions within Luzon. The survey consists of a small demographic section and an NLP awareness section which includes questions that seeks to find out the students' level of awareness and understanding of the field of Natural Language Processing. The survey was constructed and deployed using Google forms[2]. Survey participants were recruited by sending the survey URL to faculty contacts and requesting them to ask their students to answer the online survey. The internal consistency of Likert-scale questions was calculated using Cronbach's alpha and individual weighted average for each item was calculated by summing the response's weight scores and dividing by the total number of responses for each item.

A total of 618 students from 13 different universities responded to the online survey. However, only survey responses from Computer Science and Information Technology students were considered. Table 2 provides a summary of the demographic data of the respondents. Overall, the group of universities with participating students included 8 private universities, 4 state colleges, and 1 local university[3]. Of the 618 student respondents, only 401 (64.8%) provided complete and valid responses. While these valid responses may be relatively small, they potentially represent the students' level of awareness of NLP technology in the Philippines.

After taking the survey, we then looked at the range of research papers presented to the NNLPRS from 2004 to 2016. The NNLPRS is a regular gathering of researchers working on the analysis, processing, and generation of human languages by computing systems in the Philippines. It is organized by the Computing Society of the Philippines' Special Interest Group on Natural Language Processing (CSP SIG-NLP). A total of 156 papers were parsed to collect relevant data of interest. Observations from the point of view of the number of papers presented per year, authorship and institutional affiliation of the authors, focus NLP subfields, and supported languages were conducted. Although the number of papers submitted to the NNLPRS may not be exhaustive of all the NLP research conducted in the Philippines, it still can give a feel of the shape of the recent status of NLP research in the country.

---

[2] https://docs.google.com/a/jru.edu/forms/d/1jd1hGXXsWXV4UDoKgvdckdDcoco4NVdARAtfhhN4J6c/viewform?c=0&w=1

[3] Of which, only 1 state college and 2 private universities have authors that previously participated in the NNLPRS.

Table 2: Frequency Table of Respondents

Characteristic	Frequency	Percentage
CS and IT Students	401	100%
Type of Student		
Full time	313	18%
Part time	72	78%
Full time w/ work	16	4%
Year Level		
1st Year	25	6%
2nd Year	71	18%
3rd Year	214	53%
4th Year	76	19%
Irregular	15	4%
Age Group		
15-18	174	43.39
19-22	208	51.87
23-26	16	3.99
27 and over	3	0.75

4. RESULTS

Results of both the student survey and research paper survey have highlighted a number of key issues that will need to be reinforced in order to promote and sustain the continued growth of NLP research in the country. Some of these issues are described below.

4.1 Status of Student Awareness

Using Cronbach's alpha, the internal consistency of all five point scale questions discussed in this section was computed at 0.83.

4.1.1 Degree of Student Awareness of NLP

The level of awareness of students about the existence of any emerging technology is very critical because it could influence their level of interest, which, in turn, could dictate the outcome of the technical skills they will be able to develop. For this reason, a closed-ended item was included in the survey, which asked students, whether or not they are aware of the existence of NLP technology. Results indicate that a majority of students have no awareness about NLP technology at all. As shown in figure 2, 70% of the students (n=281) indicated they have never heard about NLP before.

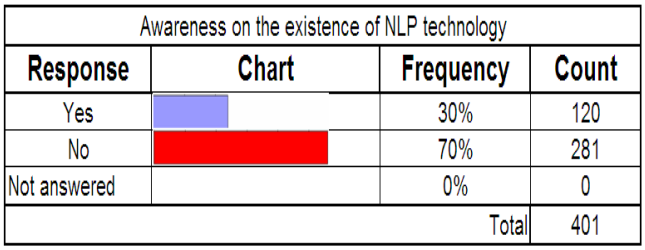


Figure 2. Student Awareness on NLP Technology

From the same figure, we can also see that only 30% of the students (n=120) are aware of the existence of NLP technology. This result mostly corroborates the intuitive observation of the weak awareness of NLP technology among Filipino students.

4.1.2 Awareness on the Nature of NLP Technology

One possible factor in fostering disinterest in any field of study is the misunderstandings that students may have about the discipline. Generally, the more misunderstandings the students have on a particular technology, the more they will fear it, and the less likely they will inquire and/or be interested about it. As a further means of assessing student awareness, students were also asked a question that examined their awareness of the nature of NLP technology. Results indicate that, among the students who indicated that they have already heard of NLP, as Figure 3 illustrates, the majority (n=95, 79.2%) admitted that they have little or no knowledge about its nature.

Only a small percentage (n=25, 20.8%) were moderately or extremely aware that natural language processing (NLP) is concerned with implementing within computers the ability to understand a normal human language. Weighted average was computed at 2.7. This deficiency in proper background knowledge can also serve as a factor in further inhibiting students' interest in this technology.

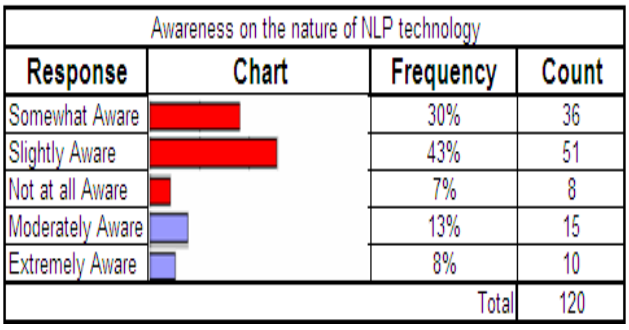


Figure 3. Student Awareness on the Nature of NLP

4.1.3 Awareness on Existence of Commercial Applications Utilizing NLP

It is important for students to realize how a particular technology can benefit the lives of many people and society in general. This is also another critical factor that can affect students' interest because the function and role that a technology can take within a society directly defines its commerciability; and students are always attracted towards developing commerciable technology. To reflect on students' awareness of this aspect, they were asked whether they are aware that many commercial applications powered by NLP already exist that have provided people with a more efficient means to do things, like Apple iPhone's Siri and the Speaktoit apps for Android users.

With a weighted average of 2.9, results with regards this aspect, as shown in figure 4, indicate that the majority of students (n=79, 66%) who have already heard about NLP has little or no awareness of the role of NLP in these applications. Only 41 (34%) indicated that they were aware of applications that make use NLP technology.

Again, this lack of knowledge on the usefulness of NLP on various facets of people's lives can project an unappealing image of the technology to students.

Awareness on commercial applications powered by NLP technology			
Response	**Chart**	**Frequency**	**Count**
Somewhat Aware		21%	25
Slightly Aware		33%	40
Not at all Aware		12%	14
Moderately Aware		16%	19
Extremely Aware		18%	22
		Total	120

Figure 4. Student Awareness on the existence of commercial applications powered by NLP

4.1.4 Students' Interest to Learn more about NLP

The survey instrument was also designed to be informational to ensure that each respondent will be able to learn something about NLP technology while answering the survey. The purpose of this is so we can ask the students at the end of the survey, how interested they are in learning more about NLP. The result for this item is very encouraging and is depicted in figure 5.

With a weighted average of 4.3, the results indicate that given the chance, an overwhelming majority of the students (n=367, 92%) is interested to learn more about NLP and its applications. Only a few (n=12, 3%) indicated that they would not be interested whereas the rest (n=22, 5%) are not sure. This positive result lends support to observations in foreign universities where students actually found NLP courses to be interesting, challenging and informative (Dale, Mollá-Aliod, & Schwitter, 2003).

Student respondent's interest to learn more about NLP			
Response	**Chart**	**Frequency**	**Count**
Agree		92%	367
Disagree		3%	12
Not Sure		5%	22
		Total	401

Figure 5. Students' implied interest to learn more about NLP

4.2 Progress of NLP Research

4.2.1 Research Paper Submissions to the NNLPRS

The graph in figure 6 depicts the number of papers presented to the NNLPRS on an annual basis. The numbers indicate that, in general, the level of interest in NLP research in the country is still very low. The average contribution is only 13 papers per symposium. In addition, as figure 6 depicts, the number of submitted papers is relatively unsteady and does not follow an increasing pattern.

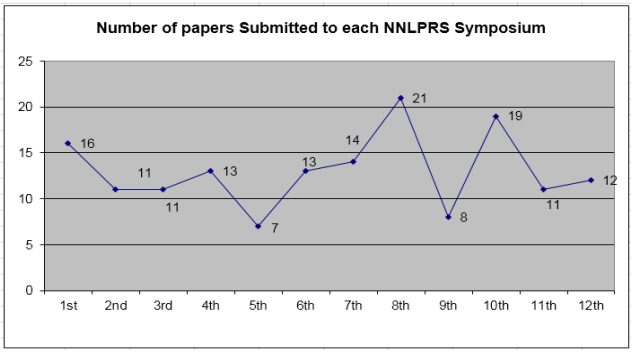


Figure 6: Number of papers submitted to the NNLPRS

Paper Distribution by Topic Focus

Table 3 gives more detailed information on the status of research progress by mapping each paper to the specific NLP subfield that it focuses on[4]. The table also lists the number of presented papers in each category per symposium. These data were gathered by using the set of relevant topics identified in the 12th NNLPRS as NLP subfield categories. Each paper was then hard classified in each category based on whether: (1) the paper mentions the topic on its Title, (2) the paper identifies the topic in its Keywords section, and (3) the paper associates with the topic through its description in its Abstract section.

A close look at the data presented in Table 3 allows us to draw two initial observations: First, the coverage of NLP subfields, per NNLPRS, is relatively sparse. The symposium with the highest topic coverage is the 8th NNLPRS with about 67% of the topics covered, the rest all have topic coverage of less than 50%. Also, of the eighteen subtopic categories only five topics, namely, Machine translation, Text summarization & generation, Language Resources & Evaluation, Audio Processing, and Syntax & grammar, have a relatively high rate of presence across all symposiums (83%, 75%, 67%, 58% and 58% respectively[5]) all the rest have 50% and below. Secondly, the intensity[6] of papers submitted per topic is also relatively weak and projects an ebb and flow pattern across consecutive symposiums. The highest intensity of papers on a given topic occurred during the 7th symposium when researchers from UP-Diliman concurrently presented their papers focused on Audio Processing. Furthermore, it can also be observed that trends in terms of recurrence of high degrees of topic intensity were limited to only a few topics, this can be indicative of sporadic and fragmented research activities.

In a way, both these observations reflect a very small growth rate for NLP. It is also indicative of a lack of focus and consolidated efforts. This scenario then raises a new and bigger challenge for the NLP community which has already shown itself to be active and working. What can be done in order to achieve wider topic coverage and higher research intensity?

---

[4] Papers presented by foreign authors were excluded as well as papers authored at foreign institutions by someone who has since moved to a foreign country.

[5] Computed by counting the total number of times the topic has been covered in a symposium over the total number of symposiums

[6] The number of papers submitted per topic

In all likelihood, the shortcomings in these two aspects likely stem from the low level of interest among faculty and students in this field of research. As will be discussed in more detail in the following subsection, this low level of interest could probably be mitigated by imposing certain support structure within educational institutions.

4.2.2 Authorship and Institutional Affiliation

Driving a technology forward initially requires for it to be included in the academic curriculum to make it more common and thereafter accessible to the public (Depositario et al., 2011; Kurian, 2014). Relative to this, it can be posited that the amount of interest of both faculty and students with regards any field of research is always proportional and linked to the level of support

Table 3: Paper distribution by topic focus, magnitude, and NNLPRS symposium

	NNLPRS												
NLP Subfield Categories	**1st**	**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**	**8th**	**9th**	**10th**	**11th**	**12th**	**Total**
Language resources and evaluation	5			2		2		5	3	5	2	2	**26**
Machine translation	2	3	4	4	1	1		1	1		2	1	**20**
Audio processing					1	3	6	2		5	1	1	**19**
Text summarization and generation		4	1	2	1	1	2	2			1	1	**15**
Syntax and grammar	1		1	2	1	4		1				2	**12**
Word sense disambiguation	2	3	4					2	1				**12**
WordNets and ontologies				2	1		1	2		2	1		**9**
Sentiment analysis and opinion mining							1	1		1	2	3	**8**
Information extraction	1	1				1	1	2	1				**7**
Sociolinguistics and NLP										4		2	**6**
Discourse analysis					1	1	1	1	1				**5**
Corpus building	1		1								2		**4**
Language Generation	2				1		1						**4**
Named entity recognition				1					1	1			**3**
Information retrieval	2							1					**3**
Culturomics							1						**1**
Machine learning for natural language										1			**1**
Textual Entailment								1					**1**
Total # of Papers :	**16**	**11**	**11**	**13**	**7**	**13**	**14**	**21**	**8**	**19**	**11**	**12**	
% topic coverage per symposium :	44.4	22.2	27.8	33.3	38.9	38.9	44.4	66.7	33.3	38.9	38.9	38.9	

Note: Many of the papers actually employ machine learning techniques (e.g., sentiment analysis through machine learning), however, as hard classification was employed, in such cases, the more specific area of application was used as the category of the paper (i.e., sentiment analysis)

provided by Educational Institutions. To gain some foothold with regards this aspect, Table 4 briefly summarizes the amount of paper contributions along with the total number of researchers who have contributed to the NNLPRS and their institutional affiliations.

At a glance, the data in this tableshows that out of the 2,299 HEIs in the Philippines, only 19 have authors that have already participated in the NNLPRS. However, particular attention should be given to the fact that the majority of authors came from DLSU and UP-Diliman with 202 and 67 authors respectively. A closer look into these HEIs revealed that both institutions have curriculum integrated with courses whose aim is to study languages through research and publication. At the same time, both institutions have established research laboratories/centers dedicated to research in Language Technologies and/or Linguistics[7]. The data can be further analyzed if we take a look at

[7] DLSU has a Center for Language Technologies while UP-D has a Department of Linguistics

the context of authorship. The relative frequency of co-authorship according to Subramanyam (1983) can be used as a measure of research collaboration using the following formula:

$$C = \frac{N_m}{N_m + N_s}$$

Where:

C = degree of collaboration

N_m = # of multi-authored papers

N_s = # of single-authored papers

Table 4: Number of papers submitted and number of contributing authors per Institution[8]

	Institution	# of Paper(s)	# of Author(s)
1	Dela Salle University	86	202
2	UP-Diliman	20	67
3	Polytechnic University of the Phils	9	36
4	UP-Cebu	8	10
5	Jose Rizal University	8	6
6	National University	6	24
7	Saint Louis University	5	14
8	UP-LB	4	7
9	Mindanao State University-IIT	4	4
10	University of San Carlos	2	2
11	Ateneo de Zamboanga University	2	4
12	Arellano University	2	3
13	University of the Cordilleras	1	3
14	University of Santo Tomas	1	4
15	Philippine Military Academy	1	1
16	New Era University	1	2
17	Colegio de San Juan de Letran-Calamba	1	1
18	Ateneo de Davao University	1	3
19	AMACC-Baguio	1	1

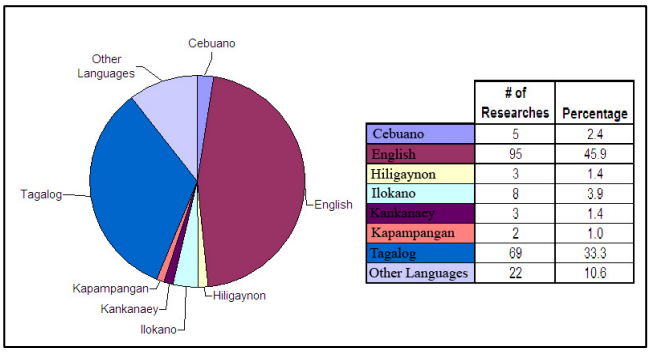
DLSU has a total of 74 multi-authored and 12 single-authored papers while UP-D has 17 multi-authored and 3 single-authored papers. Given this, their collaboration index would be around 47.4% and 10.9% respectively[9]. Below these values, the remaining institutions garnered collaborative index of only 5.5% and below. These values indicate that NLP research collaboration is higher in learning institutions with established research laboratories/centers. This finding is also supported by the work of Egghe et al. (2007) who found that scientific environment is a determining factor of collaboration, and that collaboration is higher in fields with large research laboratories.

Another interesting observation that we can make out of the figures in table 4 is that there are also institutions without established research laboratories and/or NLP-based course offerings, but are still able to produce some amount of foundational work. These institutions include PUP, UP-Cebu, JRU, NU, and SLU. A closer look at the authorship of papers from these institutions revealed that the set of papers often has a common author. Often, the common author is the key person who already received NLP training or who have co-authored with another researcher who did. It is this key person who also acts as the 'principal liaison person' between the NLP community and the Educational Institution, providing assistance to the institution in terms of identifying research issues and promoting NLP research both to faculty and students.

These observations highlight the effect of institutional support through curriculum integration and provision of appropriate research facilities as well as the presence of individual NLP champions within institutions in promoting interest and work in NLP research. It seems reasonable to expect that shortcomings in terms of both topic coverage and research intensity can be addressed if these support structures are only put in place in as many educational institutions as possible.

4.2.3 Language Research Breakdown

As previously mentioned in section 2, NLP applications can also be used to address the language problems in the country. Hence, another aspect of NLP research which deserves a closer look is the degree to which various researches have focused on Philippine languages. Figure 7 shows a breakdown of the languages covered by NLP research thus far.



	# of Researches	Percentage
Cebuano	5	2.4
English	95	45.9
Hiligaynon	3	1.4
Ilokano	8	3.9
Kankanaey	3	1.4
Kapampangan	2	1.0
Tagalog	69	33.3
Other Languages	22	10.6

Figure 7: Number of papers submitted to the NNLPRS per Philippine language focus

As the figure illustrates, NLP research in the country has not reached a level of a wide coverage of Philippine languages. Out of 186 languages and dialects, only a few have been subjected to NLP research. English is still the predominant language of focus (45.9%) followed by Tagalog (33.3%). This is not unexpected as many researchers started working on Machine Translation systems using this language pair. However, it points out that there is a lot more thrust needed to promote widespread focus on local language computing.

[8] Foreign institutions and authors were excluded from this table.

[9] With the total number of multi-authored papers divided by the total number of papers submitted to the symposium.

Further development of language processing techniques attuned to local languages will present ample research opportunities in all subareas of NLP. For instance, large collection of annotated corpora is often needed to develop and train NLP-based applications. At the current, there is an acute scarcity of these valuable corpora among Philippine languages. Therefore, more researchers are needed to systematically collect, document, and if possible distribute corpora from various languages, especially the near extinct ones. The same goes true for other lexical resources such as lexicons, dictionaries, part-of-speech taggers, etc. Compiling these resources will require the collective efforts of many NLP researchers with varied backgrounds. Although building these resources are but the initial step among many in NLP research, it is considered as a very critical component because the quality of any NLP system is directly dependent on the resources the system has access to (Granger, 2002).

5. DISCUSSION

NLP is a very potent research field and its application has tremendous potential to provide students with opportunities to not only sharpen their employability skills, but to also contribute to solving language problems that act as factors hindering the growth of a multi-lingual and developing country like the Philippines.

Results obtained from the pilot survey indicated a very low level of awareness among students with regards NLP research. In addition, among those who have already heard about it, only 21% admitted that they knew what it was all about. This means that there are still large numbers of students that have not been exposed to the concepts, frameworks, and benefits of NLP. This initial finding, although indicative, stresses the need for more intense promotion of this technology. NLP is a very complex field with a multitude of diverse subtopics and a lot of open questions. It is nearly impossible for only one or two institutions to cover all those subtopics and address all those questions. A more consolidated effort is needed to make the Philippines at par with technologically advanced countries. All is not lost, however, as results of the survey also indicate that students show willingness to learn more about NLP if they can only be adequately informed.

Undoubtedly, these huge populations of students will be difficult to reach out to by merely conducting school-to-school seminars. Ultimately, it behooves upon individual educational institutions to inform their students. Analysis of attributes of papers submitted to the NNLPRS identified best practices from well performing institutions that can enhance collaboration and promote the conduct of foundational research work in other institutions. Putting up research facilities to support NLP work, for example, seemingly promotes higher degrees of collaboration and research output. But simple initiatives like identifying and training potential NLP champions within the faculty roster will also have positive outcomes. The challenge then will become to develop a sense of personal responsibility for educational institutions to recognize the value of NLP and start acting towards integrating it into their curriculum and providing serious encouragement for NLP-based research and development activities.

Naturally, to assist interested educational institutions, the role of the CSP SIG-NLP on all this is to provide facilitation and setting-up of standards, since they are the ones who have the technical expertise to identify an effective standard curriculum as well as the list of concepts that needs to be covered for each course. They are also the ones who are in the best position to come-up with NLP-based learning materials that is more suited to Philippine perspectives. According to Mote (2012), “despite the proliferation of research materials (or because of it) there remains a barrier to the student who wishes to gain a preliminary understanding of this topic”. Developing learning materials suited to the environment, skills, and technical capabilities of Filipinos will allow uninitiated faculty and students to more effectively understand the nature and functions of NLP applications and activities and enable them to apply it to as many Philippine languages as possible; such an approach has already been proven in other countries (Bharati, Chaitanya, Sangal, & Ramakrishnamacharyulu, 1995).

There are already existing literatures that indicate the positive effects of exposing students to NLP courses (Dale et al, 2003; Zhang & Lo, 2010; Sangal, 2011). Furthermore, as NLP is bound to be a critical technology in the future, this type of educational reform will not only benefit the HEIs and faculty members, but the entire country as well, who will eventually reap the rewards of the potential of this technology (Manyika, Chui, Bughin, Dobbs, Bisson, & Marrs, 2013).

6. ACKNOWLEDGMENTS

This work was supported by the Research Department and the College of Computer Studies and Engineering (CSE) of Jose Rizal University. The author also gratefully acknowledges the assistance provided by Mr. John Sherwin Eroma in compiling the survey results from Google forms. Acknowledgment of gratitude is likewise extended to the following those who made valuable suggestions or who have otherwise contributed to the distribution of the URL for the online survey forms: Mr. Epifanio Cao, Ms. Jennifer Carpio, Mr. Joselito Carpio, Mr. Dennis Cruz, Mr. Napoleon Hermoso, Mr. Bryan Lamarca, Ms. Amy Maddalora, Mr. Nathaniel Oco, Ms. Rose Rodavia, Mr. Jonathan Santiago, Mr. Isagani Tano, Ms. Mia Villarica.

7. REFERENCES

[1] Anderson, V. B., & Anderson, J. N. (2007). Pangasinan—An endangered language? Retrospect and prospect. Philippine Studies, 116-144.

[2] Bharati, A., Chaitanya, V., Sangal, R., & Ramakrishnamacharyulu, K. V. (1995). Natural language processing: a Paninian perspective (pp. 65-106). New Delhi: Prentice-Hall of India.

[3] Bogliacino, F., Perani, G., Pianta, M., & Supino, S. (2009, September). Innovation in developing countries. The evidence from innovation surveys. InFIRB conference Research and Entrepreneurship in the Knowledge-based Economy, Milano: Universita L. Bocconi.

[4] Cambria, E., & White, B. (2014). Jumping NLP curves: a review of natural language processing research [review article]. IEEE Computational Intelligence Magazine, 9(2), 48-57.

[5] Chowdhury, G. G. (2003). Natural language processing. Annual review of information science and technology, 37(1), 51-89.

[6] Church, K. W., & Rau, L. F. (1995). Commercial applications of natural language processing. Communications of the ACM, 38(11), 71-79.

[7] Constantino, R. (1970). The mis-education of the Filipino. Journal of contemporary Asia, 1(1), 20-36.

[8] Dale, R., Mollá-Aliod, D., & Schwitter, R. (2003). Natural language processing in the undergraduate curriculum. In Proceedings of the fifth Australasian conference on Computing education-Volume 20 (pp. 9-13). Australian Computer Society, Inc.

[9] Daoud, D., & El-Seoud, S. A. (2010). Human Factors Required for Building NLP Systems. Technology for Facilitating Humanity and Combating Social Deviations: Interdisciplinary Perspectives: Interdisciplinary Perspectives, 249.

[10] Depositario, D. P. T., Aquino, N. A., & Feliciano, K. C. (2011). Entrepreneurial Skill Development Needs Of Potential Agri-Based Technopreneurs. Journal of ISSAAS [International Society for Southeast Asian Agricultural Sciences](Philippines).

[11] Egghe, L., Goovaerts, M., & Kretschmer, H. (2007). Collaboration and productivity: An investigation into ''Scientometrics'' journal and ''UHasselt'' repository. COLLNET Journal of Scientometrics and Information Management, 1(2), 33–40.

[12] Eisele, A., & Ziegler-Eisele, D. (2002, August). Towards a road map on human language technology: Natural Language Processing. In Proceedings of the 2002 COLING workshop: A roadmap for computational linguistics-Volume 13 (pp. 1-22). Association for Computational Linguistics.

[13] Gambäck, B., Eriksson, G., & Fourla, A. (2005, June). Natural language processing at the school of information studies for Africa. In Proceedings of the Second ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics (pp. 49-56). Association for Computational Linguistics.

[14] Gonzalez, A. (2003, November). Language planning in multilingual countries: The case of the Philippines. Paper presented at the Conference on Language Development, Language Revitalization, and Multilingual Education in Minority Communities in Asia, Bangkok.

[15] Granger, S. (2002). A bird's-eye view of learner corpus research. Computer learner corpora, second language acquisition and foreign language teaching, 3-33.

[16] Grimes, S. (2005). Structure, Models and Meaning, Is 'Unstructured'data merely unmodeled?. Intelligent Enterprise, Mar.

[17] Headland, T. N. (2003). Thirty endangered languages in the Philippines. Work Papers of the Summer Institute of Linguistics, University of North Dakota Session, 47, 1-12.

[18] Jones, K. S. (2001). Natural Language Processing: a historical review. University of Cambridge, 2-10.

[19] Kurian, C. (2014). A Review On The Progress Of Natural Language Processing In India. International Journal of Advances in Engineering & Technology, 7(5), 1420.

[20] Larson, K., & Watson, R. T. (2013). The Impact of Natural Language Processing-Based Textual Analysis Of Social Media Interactions On Decision Making. In ECIS (p. 70).

[21] Manyika, J., Chui, M., Bughin, J., Dobbs, R., Bisson, P., & Marrs, A. (2013). Disruptive technologies: Advances that will transform life, business, and the global economy (Vol. 180). San Francisco, CA: McKinsey Global Institute.

[22] Martin, I. P. (2012). Expanding the role of Philippine languages in the legal system: The dim prospects. Perspectives in the Arts and Humanities Asia, 2(1).

[23] McFarland, C. D. (2004). The Philippine language situation. World Englishes, 23: 59–75. doi: 10.1111/j.1467-971X.2004.00335.x

[24] Molina, G. (2012). Disappearing Languages in the Philippines. Retrieved September 29, 2015, from http://www.ethnicgroupsphilippines.com/2012/05/12/disappearing-languages-in-the-philippines-2/

[25] Mote, K. (2012). Natural Language Processing-A Survey. arXiv preprint arXiv:1209.6238.

[26] Ptaszynski, M., Kazuki, M., & Momouchi, Y. (2013) NLP for Endangered Languages: Morphology Analysis, Translation Support and Shallow Parsing of Ainu Language.

[27] Roxas, R.E. (2010) Practical Applications of Human Language Technology: the Philippine Experience. Philippine Computing Journal, Volume 5, Number 2, December 2010.

[28] Roxas, R.E., Alcantara, D.L. and Borlongan, A.M. (2010) Language Documentation and Applications in the Philippines: Implications for Mother Tongue-Based Multilingual Education. Philippine Education Research Journal.

[29] Roxas, R. E., Cheng, C., & Lim, N. R. (2009). Philippine language resources: trends and directions. In Proceedings of the 7th Workshop on Asian Language Resources (pp. 131-138). Association for Computational Linguistics.

[30] Sangal, R. (2011). A Research Oriented Undergraduate Curriculum: Design Principles and Concrete Realization.

[31] Space, C. (2001). Curriculum Development Guidelines: New ICT Curricula for the 21st Century. Brochure free of charge on request from Cedefop,(download: www. career-space. com.).

[32] Subramanyam, K. (1983). Bibliometric studies of research collaboration: A review. Journal of information Science, 6(1), 33-38.

[33] Zhang, M., & Lo, V. M. (2010). Undergraduate computer science education in China. In Proceedings of the 41st ACM technical symposium on Computer science education (pp. 396-400). ACM.

[34] Žižka, J., & Dařena, F. (2013). Discovering Opinions from Customers' Unstructured Textual Reviews Written in Different Natural Languages. E-Marketing in Developed and Developing Countries: Emerging Practices: Emerging Practices, 137.