The Filipino Wordnet Construction

Ron Jeremy Bondoc[1], Alvin Garcia[2], John Bryan Lacaden[3], Yu Hun Ping[4], Allan Borra[5]
College of Computer Studies
De La Salle University Manila
2401, Taft Avenue, Manila
{[1]jeremy_bondy, [3]jb_lacaden, [4]chinaman_pingchaw}@yahoo.com , {[2]alvin.garcia, [5]borgz.borra}@delasalle.ph

ABSTRACT
We present a serious preliminary undertaking and manual population of the Wordnet for the Filipino language. Numerous Wordnets that cater to specific languages exist today except for the Filipino language. This became one of the main motivators to pursue this study, as well as, providing for a quality linguistic resource for a wide array of Natural Language Processing (NLP) tools dealing with the Filipino language. The blueprint for the Filipino Wordnet was the original Wordnet created in Princeton University and also served as the basis for the database design. The manual creation of the synonym sets (synsets) was also facilitated from the English Wordnet. A tool was developed to assist the manual creation and editing of synset entries in order to decrease the tediousness and complexity of manual population. This tool is equipped with several NLP tools that are relevant to facilitating the manual creation of synsets. The tools include a morphological generator and analyzer, a concordancer, and a bilingual translator. Evaluations were done on the tool's usability as well as the quality of the synsets created. Overall, the usability was very satisfactory while the quality of 200 synsets from more than 10,000 synsets with over 14,000 unique words created were high and only 9% had inaccuracies in definitions.

Keywords
Wordnet, SUMO, NLP, Synsets, Synset Creation Assistant

1. INTRODUCTION
Wordnet was created for the purpose of having a lexical-conceptual database. This database would house lexical units and their relationship to one another. These lexical units would be organized into semantic networks [6]. The initial idea of Wordnet was to be used as an aid to online dictionaries. As the work progressed, Wordnet developers became more ambitious. They sought Wordnet to not just instantiate hypotheses based on psychological research results, rather a dictionary based on psycholinguistic principles [5]. With this idea, Wordnet became more than just another online dictionary.

The complexity of developing Wordnet gave problems to the developers though. One of these problems is that it may become too redundant; this is due to the fact that one word may belong to more than one category (nouns, verbs, adjectives, adverbs, and function words). An example would be the word "*back*", there is a meaning of "*back*" found on the verb category that is the same as the meaning found on the adverb category. However, the advantage in this is that users can clearly identify and systematically exploit the fundamental differences in the semantic organization of those syntactic categories [5]. There are other types of Wordnet aside from the English Wordnet that was developed in Princeton University, some of these are the EuroWordnet, Estonian Wordnet, and the Arabic Wordnet which will be discussed in this document.

The English Wordnet, which was developed in Princeton University, consist an approximate of 57,000 noun word forms; these noun forms are then grouped into 48,800 synonym sets. Wordnet also contains an approximate of 19,500 adjectives that are grouped into 48,800 synonym sets. These numbers continue to grow with the help of different individual contributions, which is one of the advantages of an online database.

The Estonian Wordnet is composed of 7,678 synsets. As of now, these synsets are divided into 5,028 noun synsets and 2,650 verb synsets. The EuroWordnet (EWN) is also a multilingual database of different languages from Europe, which are Italian, German, French and Dutch. It is patterned after the Princeton Wordnet or the English Wordnet, which is also linked to an interlingual index (ILI). Via this index, the different languages of the European Wordnet are interconnected making it possible for different words from other languages go to other languages with similar words. The index for this Wordnet gives access to top-ontology of 63 semantic distinctions, which is the common semantic framework for all the languages. Although there are common semantic frameworks for all the languages, the specific properties of the languages are different for each of the Wordnets [13].

The Arabic Wordnet (AWN) is constructed under the methods used for developing the EWN because of its high compatibility with other Wordnets. The AWN is aligned to every other Wordnet developed either directly or indirectly by making use of the Inter-Lingual Index (or ILI) and the Suggested Upper Merged Ontology (SUMO). The database design of the AWN supports multiple languages and its interface also explicitly states being multilingual. A huge population of the Middle East and other African countries, and the language of Muslims is Arabic. There are only very few projects done in the field of computerized language and lexical resources for this widely used language [2]. This is the reason for the development of the Arabic Wordnet.

In relation to the Filipino Wordnet, the researchers were only able to locate one uncompleted work in creating a Filipino Wordnet - the work of Tan [12].

2. THE SOLUTION: MANUALLY CREATE SYNSETS
Prior to the inclusion of the Synset Creation Assistant Tool, the narration that follows describes the tedious and complex procedures of manually creating the synsets. First, a Microsoft Excel file/worksheet is created and divided into 7 columns. The 7 column headers are: Synset ID, English Word, English Synset, Part of Speech (POS), English Definition, Filipino Synset, and Filipino Description. Tools are prepared such as a browser with DEBVisDic plug-in, a credible set of online or hardcopy English-Filipino dictionaries and access to SUMO concepts [8] through Sigma Ontology Website which is an online tool for viewing SUMO (http://sigma.ontologyportal.org:4010/sigma/KBs.jsp).

DEBVisDic is an online Wordnet editor and browser that allows access primarily to English synset entries. After the preparations, Listing 1 outlines the process of adding or creating Filipino synsets.

1. Think of an English word;
2. Use DEBVisDic to search the synsets of the English word;
3. Choose one synset and translate to Filipino. This now becomes an entry for the Filipino synset;
4. List down Part of speech and English definition;
5. Use Synset ID to search for SUMO term;
6. Translate English Word into Filipino using a translator or dictionary (or one could do the translation);
7. List down Filipino equivalent of English Synset;
8. List down Filipino definition; and
9. Use Dictionary or perform direct/rough translation from English definition.

Listing 1. Steps in Manual Creation of Synsets

All of these steps involved a lot of copy-pasting mechanism as well as shifting from one window application to another window and back to the Excel worksheet and even referring and page-flipping to hardcopies (dictionaries). Mechanically being adept to this still entailed more than 10 minutes per synset entry based on experience from the study.

2.1 Tool: Synset Creation Assistant

The Filipino Wordnet Synset Creation Assistant Tool system is designed to help users manually create synsets. The Synset Assistant Creation Form is composed of numerous NLP tools to aid users in analyzing Filipino words and translating English synsets to Filipino. Figure 1 shows a screenshot of the Synset Creation Assistant for the Filipino Wordnet. Tools are integrated in one window that allows quick access to these different tools. In the same manner, the database allowed for mechanisms such as word/synset-locking to a user so modifications and deletions are controlled. But report mechanisms are in place to facilitate modifications. Viewing and referencing are readily available to all users of the database.

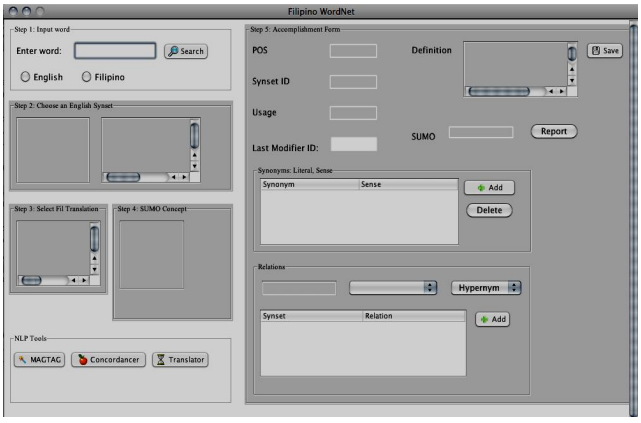

Figure 1. Screenshot of the Synset Creation Assistant Tool

As shown in Listing 1 on steps devised by the authors in the manual creation of synsets, the Synset Creation Assistant Tool's interface was designed with the sequence of steps in mind. The tool, though, remove windows switching between different tools and applications where disaggregated data are found. Moreover, the task of copying and pasting the data from the different windows to the Excel worksheet columns has also been removed.

The next subsections outline the components and tools that were integrated in the Synset Creation Assistant for the Filipino Wordnet.

2.1.1 The Filipino Wordnet Database

The database currently houses more than 10,000 Filipino synsets and over 14,000 unique words. The synsets were all manually translated and added. The database was built using the design of the English Wordnet's SQL version [5]. The entity-relation diagram (ERD) is shown in Figure 2.

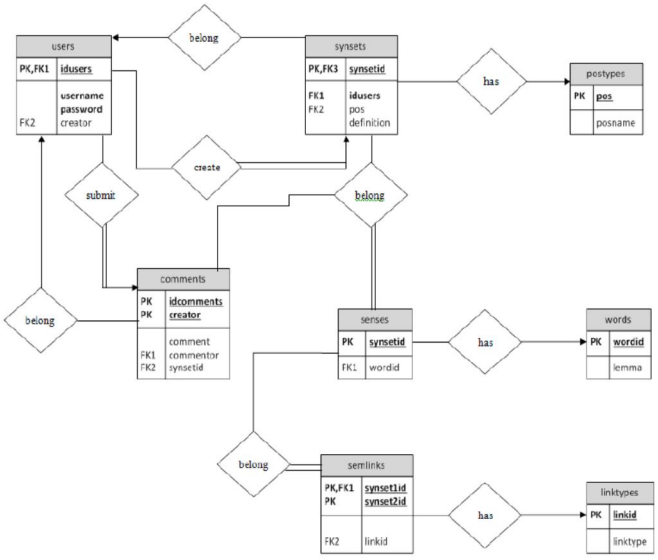

Figure 2. ERD for the Filipino Wordnet Database

2.1.2 SUMO

Suggested Upper Merged Ontology (SUMO) acts as the Interlingua of the Filipino Wordnet system, especially when connected to other Wordnets. SUMO offers the largest public axiomatized ontology [8]. It connects the Filipino Wordnet database and the Princeton Wordnet database using their SUMO counterparts. In order to access SUMO, the SIGMA inference engine has to be integrated to the system. SIGMA is a web-based tool which acts as the browser for SUMO. Figure 3 provides an insight on how SUMO is used to map the Filipino Wordnet synset entry "*hilamos*" to the Princeton English Wordnet by using the formal axioms of SUMO.

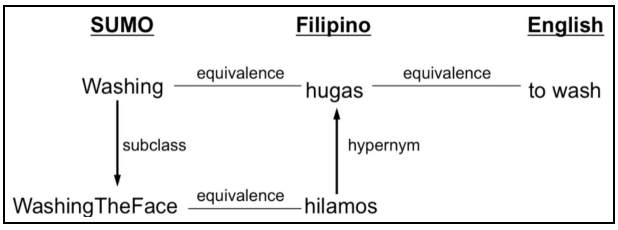

Figure 3. Relation of Filipino Wordnet Entry "Hilamos" to English Wordnet (from [3])

2.1.3 Concordancer

The current Concordancer being used is the found in the PALITO website [9]. PALITO is an online corpus management system. It provides mechanisms in storing and collecting various Philippine corpora in electronic form.

The Filipino Wordnet's Concordancer currently has seven Filipino texts which act as its corpora. The Filipino texts are all composed of Filipino legends and fables.

The Concordancer accepts a Filipino word and returns sentences from the corpora containing that queried word and arranged accordingly as shown in Figure 4. Such presentation of results by the Concordancer allows for analysis of the use of a given word in different contexts as shown by the words before and after it.

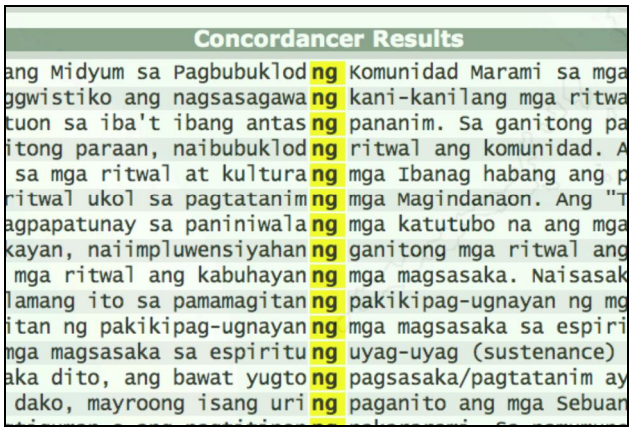

Figure 4. Concordancer Result for Word *'ng'* in Palito

The Concordancer, though, has a limited corpus: it only contains seven Filipino literary texts. These texts are made up of Filipino legends and fables. The legends and fables are made up of Filipino words that are now seldom being used, since they are considered old. These make the use of the Concordancer a bit limited and constraining.

2.1.4 MAGTag

The Morphological Analyzer and Generator for Tagalog (or MAGTag) implemented in the system is designed by Aquino and his colleagues [1]. For the Synset Creation Assistant, the original MAGTag was modified to suit the synset creation form. Some of the modifications done were to integrate the generator and analyzer in a single interface. The generator accepts a Filipino root word and generates a list of affixed forms of that word. The analyzer on the other hand accepts an affixed Filipino word and breaks it down and returns the root word and the affixes used along with their descriptions.

The MAGTag has two issues on generation and analysis. Mainly, the problem was over-generation of affixes/inflections to root words. It will continue on plugging affixes even though the resulting Filipino word does not exist in the Filipino language or the construction, anomalous. This is due to the generator having no means of checking for valid words.

The analyzer on the other hand has problems on POS tagging and on analyzing complex Filipino words. It can only analyze Filipino words that have one affix attached to it. Once the user inputs a complex Filipino word, only one affix will be identified. The POS tagging of the analyzer always displays "*verb*" as the POS.

2.1.5 Bilingual Translator

The translator used by the system is made up of English and Filipino words. The translator works both ways - it can translate from English to Filipino and Filipino to English. The translator currently contains about 1,100 words which were gathered from online dictionaries. This will make translating from English to Filipino or Filipino to English limited as well. This is not enough to cover the entire, or the majority, of the Filipino words.

2.1.6 User Interface

The user interface of the system is designed to look or function like the user interface of DEBVisDic [10]. The user interface is integrated with tools discussed earlier. The user interface makes use of two modes - the walkthrough mode and the advance mode. This is to accommodate both beginners and veteran users of the system. The user interface was also designed to be as guided as possible. It has step-by-step labels and tooltip texts which will help users in what to do.

3. TESTING AND RESULTS DISCUSSION

The Synset Creation Assistant Tool was tested by a total of 61 evaluators from De La Salle University, Manila. There were two phases of testing performed. The first part of the process is the evaluation of the user interface and the reliability of the tool to which 31 Computer Science student evaluators were surveyed. The next part of the evaluation is to gauge the usability of the different NLP tools and the overall usability of the system to which 30 non-Computer Science students were surveyed.

3.1 Database Design

The Filipino Wordnet database must be able to handle all the synsets translated to Filipino by the users. Moreover, duplication of synsets in the database is prevented from happening in order to keep the database free from multiple entries of a synset with possibly similar information. In addition to this, the database to handle the Filipino synsets is able to lock a *synset id* to only one user. A good way of making sure that the database is able to handle the Filipino synsets well is by using the tool developed to view the Filipino synsets made and modified by the users.

The first question in the usability testing tackled about the users being able to view information that are needed in creating synsets. The second question talked about being able to view or edit the synsets that were committed into the database. Questions three to six were about the various NLP tools (the translator, SUMO, MAGTag, and the Concordancer) and their functionalities.

Based on the results (Figure 5), it can be said that even non-Computer Science students who know very little about database systems, appreciate and can see that that database design for the FilWordnet system is effective in handling data. Also, the users seem satisfied that they immediately see any changes they made on their synsets by searching these again in the database. This is an important point since this assures the user that the changes he/she has made are really stored in the database. It is also seen that the NLP tools were able to properly do their tasks in guiding the users.

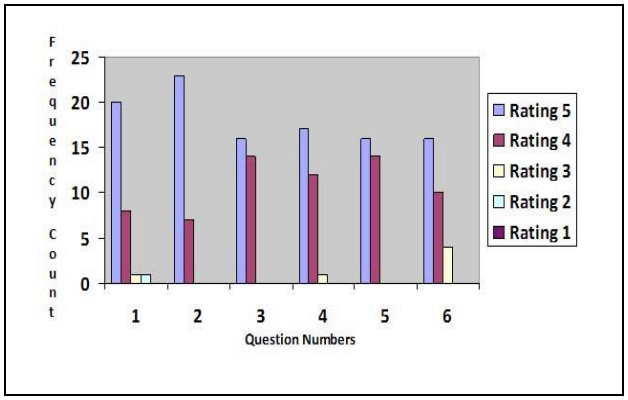

Figure 5. Results for the Usability Testing

3.2 Online and Offline Dictionaries

In order to avoid having the user transfer to other windows or switch to other resources, a database was created to store English words with corresponding Filipino translations, which serves as a bilingual dictionary for the system. The English and Filipino words in the database come from the online resources being used by the researchers previously. Most of the evaluators were satisfied with the bilingual dictionary used by the system.

It is important that the non-Computer Science student evaluators had a high mark for this since at least 8 of the 30 evaluators are under the Literature program of the College of Liberal Arts (CLA). With the bilingual dictionary still having a good mark with these non-Computer Science student evaluators, it can be said that the words in the bilingual dictionary can suffice for translating English synsets to Filipino. Currently, there are approximately 11,000 unique rows in the database. The current number of entries in the database however, is still insufficient since as noted by some of the evaluators, some English words have no translation/s in Filipino.

3.3 Concordancer

A good number of the evaluators were convinced by the use of the Concordancer in creating synset definitions. Fifty-six percent (56%) of the Computer Science student evaluators gave good marks. It means that the tool really helped the users in understanding a word by seeing sentences that use the word they are defining. Using context clues is definitely very helpful in creating definitions even without a dictionary. As for the non-Computer Science student evaluators, 66% were convinced with the use of the Concordancer in the system. Among the suggestions collected is to make the space for the resulting sentences of the Concordancer bigger so that the user does not need to scroll horizontally to view a sentence. These suggestions from the evaluators have already been addressed in the system.

3.4 SIGMA

The marks received for the obtaining of the SUMO concept of a synset is very effective as can be seen in the results from the evaluators. Initially, the interface design of the SUMO portion of the software makes it hard for the user to understand what it is for and tends to skip that step. To prevent this, the developers made a better design for this, highlighting the SUMO concept step and by putting a tooltip text, which describes the SUMO concept. The non-Computer Science student evaluators appreciated the speed and accuracy of getting the SUMO concept of a synset. They also found the SUMO concepts to be very useful since according to some evaluators, it is a way of grouping words via their concepts.

3.5 MAGTag

Among the items for usability, the use of MAGTag has the lowest mark with the two groups of evaluators. This can be attributed to the limit of the analyzer part returning only verb as the part-of-speech of any input word. This problem has been noticed by a lot of the evaluators even during the first part of the testing. On the other hand, the limitation for the generator part is that it over-generates affixed word forms since what it does is to plug-in its list of morphological rules to an input word. Using the generator component to identify the part-of-speech of the affixed word form may compensate for the analyzer's inability to recognize other parts-of-speech.

3.6 User Interface

The study made use of the ten heuristics of [7]. The questions used in the surveys talked about how the users' experience in using the interface of the system. The questions asked whether users were at a loss at some point when exploring the functionalities of the system. It also tackled whether users were guided by the system concerning on what to do at each step and what step was needed to be done next. Results are shown in Figure 6.

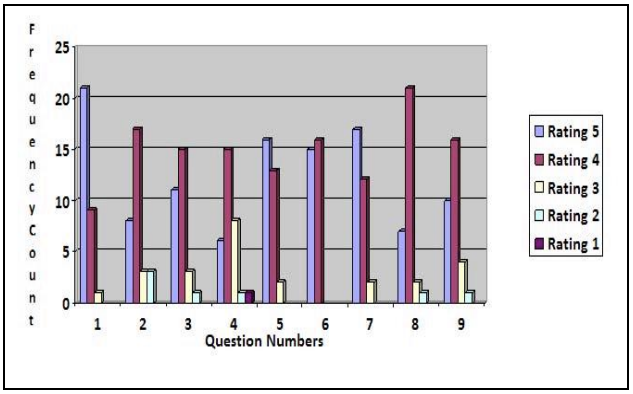

Figure 6. Results for the User Interface Testing

The design of the interface shows the most frequently used and needed information to the user at all times, hiding the less important information until it is needed. This is very important because the user may feel at a loss when there are so many things presented to him at the same time. This is the case with the previous interface developed since it immediately shows the user a window full of text boxes and labels wherein most of the items displayed to the user could be accessed and needed at a much later time. The design of the interface follows the process of creating the synset established during the study, commencing with an English word to translate, then adding Filipino word translations then the definition. The flow of adding information in the Synset Creation Assistant Tool is clearly based on the manual process of synset creation as performed by the authors. It can be seen also that the marks received from the non-Computer Science student evaluators were as good as the marks received from Computer Science student evaluators. It can be said that the system is usable by people who may not be as computer literate as the Computer Science students.

3.7 Evaluated Filipino Synsets

There are a total of 200 different synsets currently evaluated for this project. These 200 synsets were chosen randomly in the collection of synsets made by the authors. According to Crudo [4], even linguists themselves who translate English to Filipino texts, do not know all Filipino words which made it hard for them to evaluate all of the synsets that were given. Among the synsets which the evaluator had problems are those synsets which are words pertaining to fruits. For example, the word "*blackberry*" when translated to Filipino using the online dictionaries is "*lumboy*". The evaluator had a problem knowing whether the translation is correct because he does not know whether there is indeed a word "*lumboy*" in Filipino. In addition, the evaluator is also unknowledgeable as to what the Filipino translation of *blackberry* is. Other examples of synsets which the evaluator thought to be quite trivial are: *blueberry* and *fig* with the following Filipino translations, *alimanim* and *igos*, respectively.

Another problematic part of the evaluation is that some translations are considered to be *bastardized* English. Some are English words assimilated with Filipino morphology. The original English word is translated to Filipino by simply changing the spelling of the original. An example is the synset "*cherry*", which when translated to Filipino using the resources is called "*tseri*".

Because of the problems stated above, it can be said that the bilingual dictionary is not 100% accurate, however it cannot also be concluded that the dictionary is totally inaccurate since most of the words in the Filipino synsets were found to be correct. This result shows that other supplementary resources must be found in order to provide a purely accurate bilingual dictionary. Among the 200 synsets given to the resource persons for evaluation, all of the English terms were accurately translated in Filipino, although at least 9% of the total synsets evaluated whose Filipino translations seem to be inaccurate since the evaluator is not sure if such words in Filipino exist. These results are summarized in Figure 7.

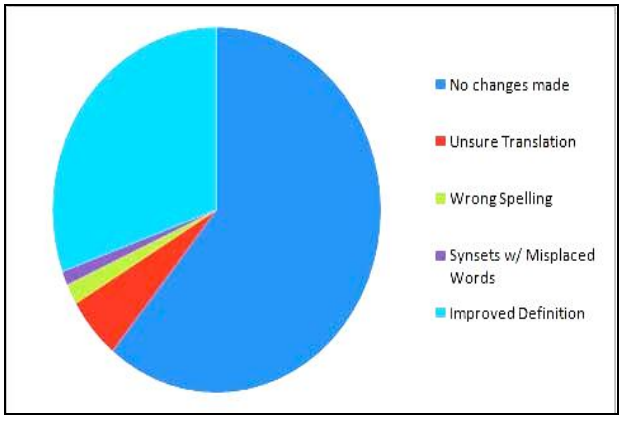

Figure 7. Results for the Synset Evaluation

The general assessment of the evaluators for the project is that according to them, it is a good way to preserve the Filipino language since it stores in a database the Filipino words, which may be forgotten due to the overtaking of the English language in the country [11]. The evaluators recommended providing a glossary for all the synsets, grouping them according to topics such as science, technology, and so on. However, the use of the SUMO concepts already provides this glossary-like functionality. In addition, one of the evaluators suggested having a dictionary called *Father Leo James English' English-Tagalog Dictionary*. The evaluators commend the research since this endeavor of using the Filipino language as part of academic research is usually set aside.

4. CONCLUSION AND RECOMMENDATION

Based on the test results obtained from the study, it can be concluded that the integration of the various NLP tools in the Synset Creation Assistant Tool developed is adequate in helping the user to produce quality synsets. In addition, the design of the interface is able to guide novice users through the system and also through the creation of Filipino synsets. With the use of the system, the Filipino synsets produced as base concepts for this project are found to be mostly correct with only some specific synsets which can be improved.

The 14,000 unique words and 10,000 synsets made for the Filipino Wordnet were created by Computer Science students who are speakers of both Filipino and English languages, while the creators of synsets for Wordnets for other languages are composed of professional linguists. Comparing the synsets made for the Filipino Wordnet with the synsets of the other Wordnets, it can be concluded that by using the resources and tools in this research helps in the creation of accurate synsets, in addition to the knowledge of the users on the Filipino and English languages. Having used the resources such as the Morphological Analyzer, and Generator, Concordancer, bilingual dictionaries, and an interface to facilitate in the synset creation will yield accurate and reliable synsets. It can be concluded that ample knowledge, a good interface design and reliable tools will help any person to produce accurate synsets, even in the absence of professional linguists.

It is recommended to further increase the number of Filipino texts in the Concordancer corpora. Errors found in both the analyzer and the generator of MAGTag, like over-generation, should be addressed in future builds. Packaging and distribution of the Filipino Wordnet is being undertaken to make it available to the Global Wordnet Consortium, Asian Wordnet, DEBVisDic Team and the Language Grid.

Further linguistic evaluation of the current Filipino Wordnet synset entries is imperative for more accurate performance of NLP tools capitalizing this linguistic resource. Moreover, the public release and interface provision for updating and further populating synset entries is hoped to increase the usability and scope of this linguistic resource.

5. REFERENCES

[1] Aquino, M., Fernandez, E. & Villanueva, K. (2007). *MAGTag: A Rule-Based Tagalog Morphological Analyzer and Generator*. Undergraduate Thesis, College of Computer Studies, De La Salle University, Manila.

[2] Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., et al. (2006). Introducing the Arabic Wordnet Project. In *Proceedings of the Third International Wordnet Conference* (GWC-2006), 295-299, Jeju Island, Korea, January 22-26, 2006. DOI=10.1.1.128.1517

[3] Borra, A., Pease, A., Roxas, R. & Dita, S. (2010). Introducing the Filipino Wordnet. In *Principles, Construction and Application of Multilingual Wordnets:*

Proceedings of the 5th Global WordNet Conference (GWC-2010), 306-310, Mumbai, India, January 31–February 4, 2010. P. Bhattacharyya, C. Fellbaum & P. Vossen (eds.), ISBN 978-81-8487-083-1, New Delhi.

[4] Crudo, C. (2010). Personal communication, July 21, 2010.

[5] Miller, G., Beckwith, R., Fellbaum, C., Cross, D. & Miller, K. (1990). *Introduction to Wordnet: An Online Lexical Database*. International Journal of Lexicography.

[6] Morato, J., Marzal, M., Llorens, J.,& Moreiro, J. (2004). Wordnet Applications. In *Proceedings of the Second Global Wordnet Conference* (GWC-2004), 270-278, Brno, Czech Republic, January 20-23, 2004. P. Sojka, K. Pala, P. Smrz, C. Fellbaum & P. Vossen (eds.), ISBN 80-210-3302-9, Masaryk University, Brno.

[7] Nielsen, J. (2005). *Ten Usability Heuristics.* Retrieved July 2010, http://www.useit.com/papers/heuristic/heuristic*list:html*

[8] Niles, I. & Pease, A. (2001). Origins of the IEEE Standard Upper Ontology. In *working Notes of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology*, 37-42, Seattle, DOI=10.1.1.76.8966.

[9] PALITO (2011). *Palito: Building the Philippine Corpus*. Retrieved from http://ccs.dlsu.edu.ph:8086/Palito

[10] Rambousek, A. & Kudlej, M. (2002). *DEBVisDic Wordnet Editor and Browser Based on Debii Platform*. Retrieved May 2, 2006, http://www.nytud.hu/cescl/proceedings/Rambousek-Kudlej*CESCL.pdf*

[11] San Juan, Michael (2010). Personal communication, July, 2010.

[12] Tan, P. & Lim, N. R. (2007). Towards a Filipino Wordnet. In *Proceedings of the 4th National Natural Language Processing Research Symposium*, 26–31, De La Salle University, Manila, June 14-16, 2007.

[13] Vossen, P., Bloksma, L., Rodriguez, H., Climent, S., Calzolari, N., Roventini, A., et al. (1998). *The EuroWordnet Base Concepts and Top Ontology (Rapport technique)*. EuroWordnet.