# **Estorya** : a Part of Speech Tagger for Cebuano/Visayan Language

**Vince Mariel Sayon**
**Christian Mae Mendez**
**Khryss Bea Ayuste**
**Jim Salarza**

**MAY 2017**

The Cebuano language (Binisaya, or simply Bisaya, to its native speakers) is the second most widely-spoken native language in the Philippines next to Tagalog, and is used mostly in the Visayas and Mindanao regions. The language is divided into six main dialects, namely: Traditional, Urban, Negrense, Boholano, Leyteño, and Mindanao dialects, depending on the location where the speakers are based. [1]

However, despite its wide coverage, Cebuano language has very limited resources when it comes to grammar, text, and speech processing tools and methods in the field of developing automated interlingual/multilingual technologies, which basically rely heavily on having a large collection of data at hand. Although this may not be a problem to huge languages such as English, which has multitudes of resources, it is an especially huge problem for “resource-poor” languages such as Cebuano.[2]

The scarcity was partly due to the fact that before the government approved the K-12 Curriculum in the Philippines, Cebuano - was not really taught in educational institutions before the implementation. The MTB-MLE (Mother Tongue-Based Multilingual Education) now teaches the Mother Language (Lingua Franca), which is defined in Sec. 4 of the Republic Act No. 10533 as “the language or languages first learned by a child, and is identified as a native language which he/she knows best.”[3]

With the new Curriculum in the Philippines, comes plenty of barriers and challenges in the implementation of the Mother Tongue (specifically Cebuano). The lack of reliable resources for Cebuano grammar and spelling, for instance, in making modules. Due to the lack of resources, there are major difficulties when it comes to translations especially when trying to incorporate it to English-based studies such as Science and Mathematics. [4]

One of the most useful tools in Natural Language Processing is the Part of Speech (POS) Taggers, which assigns tags (e.g. noun, verb, etc.) to every tokenized word within a pre-read sentence. POS taggers are one of the important tools in developing reliable NLP technologies and is used in speech recognition, natural language parsing, information retrieval and extraction as well as in developing language corpus. It is used mainly in translators in order to try to construct an accurate version of the translated sentence using the new language’s grammar. In order to do this, the POS tagger has to parse the sample sentence into a series of simple words. It then tokenizes the words, eliminating white spaces. After which, the tagger analyzes and identifies what each word’s category of POS is. This enables the tagger to identify the phrases in the text. [5]

POS taggers employ one or a combination of approaches in order to generate results. Ways in determining POS tags involve but are not limited to statistical approach(i.e. N-gram based and HMM or Hidden Markov Model), machine language approach, and rule-based approach (i.e. Brill). [5]

The Cebuano-Visayan part-of-speech categories is made up of the 8 major parts of .speech categories in the English language, namely: nouns, pronouns, verbs, adverbs, adjectives, prepositions, conjunctions, and interjections. These categories also have multiple subcategories depending on how the word is used in the sentence. For example, a noun could be proper, common, concrete, etc.[6] Using the subcategories in a POS tagger depends on how finely-grained the POS tagger is.

With these gathered information, the proponents saw an opportunity to perform a research on making a POS tagger/ API that delves on the Cebuano language, that future developers may use to create more reliable resources for Cebuano-based applications and programs, as well as educational purposes. The proponents wish to explore the capability of Machine Learning approach to improve tagging. To test the API, the proponents plan on implementing the POS tagger in a gamified educational application that will help children in their Mother Tongue studies.

## BIBLIOGRAPHY

[1] Anon. Information Retrieval. Retrieved May 6, 2017 from https://books.google.com/books/about/Information_Retrieval.html?id=nH9G7LCR1-UC

[2] Information Retrieval. Retrieved May 5, 2017 from https://books.google.com.ph/books?id=nH9G7LCR1-UC&pg=PA235&lpg=PA235&dq=available+tools+for+Cebuano+language+processing&source=bl&ots=Ef4aAxYw07&sig=5yhWhuZGyUIRqmWIeqbptH8lEhk&hl=en&sa=X&redir_esc=y#v=onepage&q=available tools for Cebuano language processing&f=false

[3] Department of Education, 2013. *Implementing Rules and Regulations (IRR) of Republic Act No. 10533 Otherwise Known as the Enhanced Basic Education Act of 2013*, DepEd Order No. 43, s. 2013

[4] "Mother Tongue-Based Learning Makes Lessons More Interactive And Easier For Students".October,2016.Retrieved 11 May 2017 from www.deped.gov.ph/press-releases/mother-tongue-based-learning-makes-lessons-more-interactive-and-easier-students.

[5] Hassan, F., 2006. *Comparison of Different POS Tagging Techniques for Some South Asian Languages*. Bachelor of Science in Computer Science and Engineering. BRAC University.

[6] Kilaton, Cesar. "Http://Www.Binisaya.Com/Content/Parts-Speech-Mga-Bahin-Sa-Pamulong". *Binisaya.com*. N.p., 2012. Web. 6 May 2017.