Towards the Development of a Rule-Based Filipino Morphological Analyzer and Generator

Kristine Mae Adlaon
ITE Program
Univ.of the Immaculate Conception
Father Selga St., Davao City
kadlaon@uic.edu.ph

Charibeth Cheng
College of Computer Studies
De La Salle University
2401 Taft Ave, Manila Philippines
charibeth.cheng@dlsu.edu.ph

Maristella Aquino
College of Computer Studies
De La Salle University
2401 Taft Ave, Manila Philippines
maristella_aquino@dlsu.edu.ph

Ervin Fernandez
College of Computer Studies
De La Salle University
2401 Taft Ave, Manila Philippines
ervin_fernandez@dlsu.edu.ph

Kevin Villanueva
College of Computer Studies
De La Salle University
2401 Taft Ave, Manila Philippines
kevin_villanueva@dlsu.edu.ph

ABSTRACT

This paper presents a rule-based morphological analyzer and generator (MAG) for Filipino language. The rules are expressed using the two-level PC-Kimmo [2] format, which allows one rule to be used for both morphological analysis and generation. We modified the basic two-level rule format to include information on grammatical changes, which is not handled by the basic format. The system is composed of three modules; namely the (1) Rule Compiler, which processes rules to be used by the analyzer and generator, the (2) Analyzer Module which determines the root word of a given word, and the (3) Generation Module which produces the list of transformed words from a given root word. Both the Analyzer and Generator modules detect morphological changes or phenomena that occurred during transformation, including affixation and reduplication. The MAG can process nouns, verbs, and adjectives. The Generation module was tested using 13,937 words, with an accuracy of 68.42%. On the other hand, the Analyzer module was tested on 16,540 Tagalog words, with an accuracy of 83.84%. The accuracy may be improved by considering all orthography rules for Filipino, as discussed in Komisyon sa Wikang Filipino's "Binagong Gabay sa Ortograpiya ng Wikang Filipino" [1] by National Artist Virgilio S. Almario. Phonetic information such as stress and glottalization of words may also be included as another attribute for the generation and analysis.

General Terms

Algorithms, Languages

Keywords

Morphological Analyzer, Morphological Generator, Rule-based approach, Tagalog Morphology, and Morphotactic Rules

1. INTRODUCTION

The Philippines is an archipelago composing of more than seven thousand and one hundred (7,100) islands, with over one hundred (100) distinct languages. Tagalog is the top Philippine language, and is the primary basis of the national language Filipino. It is spoken in Metro Manila and its nearby provinces. Tagalog has a complex verbal system exhibiting morphological phenomena such as affixation, stress shifting, consonant alternation, and reduplication for determining parts of speech, aspect, and voice which includes the use of 2 various particles, prefixes, infixes, suffixes, and circumfixes. This makes Tagalog much more morphosyntactically complex than a language like English which makes less use of markers and morphemes for determining parts of speech and focus as it does syntactic arrangements.

Nelson [8] states that Tagalog has been phonologically and lexically influenced by languages such as Spanish, English, and Indonesian, however, its morphosyntactic properties have remained wholly Tagalog. Verbal inflection to indicate aspect differs according to the affix class of the verb. *Ka-* marks a recently completed action of the verb, and is often followed by the adverbial particle *lang*. This is seen with the prefix *ka-* added to the root *kain*, as in *kakakain ko lang ng karne*, meaning 'I have just eaten some meat'. The reduplication signals action started. Reduplication is the repetition of parts or all of the affix or of the root. For instance the root, *gawa*, could be infixed with *-um-* to produce the form *gumawa*, after which reduplication could form the word *gagawa* or *gumagawa*, thus reduplicating the initial consonant and vowel combination of the root. However, if the root begins with a vowel, like a word *abot*, loosely meaning 'to grab', then the infix with *-um-* is attached before the reduplicated initial vowel to form *umaabot*.

Works that handled Tagalog Morphology include Bonus [3], Fat [5], Fortes [6], and Nelson [8]; which will be discussed in the next section. Rule-based approaches for morphological analysis and generation make use of hand-crafted rules. Previous related works focus mainly on developing either a Morphological Analyzer or a Generator for Tagalog only. This study investigates the use of a rule-based approach in developing a single system for Morphological Analysis and Generation that uses the same set of rules.