Jayaweera, A.J.P.M.P., Virtusa (Private) Limited.
N.G.J. Dias, Department of Statistics & Computer Science,University of Kelaniya

*Paper: Sustainability*

# Part of Speech (POS) tagger for Sinhala language

Sinhala is a morphologically complex and agglutinative language. Most of the features of the words are postpositionally affixed to the root word. This paper presents a POS (Part Of Speech) tagger for Sinhala language using Hidden Markov Model (HMM).

Part Of Speech tagging is one of the fundamental and important steps of any natural language processing task, which is the process of assigning a part-of-speech or other lexical class marker to each word in a sentence. This is important in every area of natural language processing (NLP) from speech recognition to machine translation, spelling and grammar checking to language-based information retrieval on the web. The tagger takes a sentence, a tagset and a corpus as input and gives the tagged sentence as output. The tagging process is done by counting the tag sequence probability $P(t_i|t_{i-1})$ and a word-likelihood probability $P(w_i|t)$ form the given corpus, where the linguistic knowledge is automatically extracted from the annotated corpus.

In this research, we use the tagset and the corpus developed by UCSC/LRTL (2005) under PAN Localization project. The current tagset consists of 29 morphological syntactic tags. An algorithm is presented in this paper for the implementation of POS tagging system for Sinhala language, which would enable users to reach more than 80% of the success rate.

**Keywords: Part-of-speech (POS), Morphology, lexical, lemma, word stream, affixes, algorithm, stochastic model, Hidden Markov Model (HMM), Natural language processing (NLP)**