Some computational aspects of natural language understanding: A case study on Spoken Sinhala

Chandima H. de Silva\(^1\) and Neelakshi Chandrasena Premawardhena\(^2\)

\(^1\)Department of Statistics and Computer Science, University of Kelaniya, Kelaniya, Sri Lanka. E-mail: chandima@kln.ac.lk

\(^2\)Department of Modern Languages, University of Kelaniya, Kelaniya, Sri Lanka.

Computational Linguistics (CL) is a relatively new area of research that incorporates knowledge from both computer science and linguistics. CL is mainly concerned with the computational aspects of the human language faculty. Applied CL focuses on the practical outcome of modeling human language use. Natural language understanding plays a major role in designing effective CL systems. This paper attempts to address some issues related to transferring Spoken Sinhala into electronic form in the context of CL. Though studies have been conducted in relation to European languages, a study of this nature has not been done involving Sinhala.

Diglossia in modern Sinhala with a vast difference between the spoken and the written variety has paved for two separate areas of research today. Linguists refer to the two varieties as Literary and Spoken Sinhala. Although much research has been conducted on Literary Sinhala, the area of Spoken Sinhala has been somewhat neglected. The present study is based on data available from Spoken Sinhala, thus attempting to contribute to filling this void.

This paper concentrates only on reference devices (mechanisms employed by native speakers to keep track of referents in discourse, i.e. pronouns, nouns, honorific and kinship terms) in Spoken Sinhala. Unlike major European languages used in applied CL, Sinhala belongs to the category of languages that employs many sociolinguistic variables that are not signaled directly by the grammatical devices. Identifying speaker (first person) and addressee (second person) generally causes no great difficulties in reference tracking. However, identifying third persons in discourse poses many difficulties in Spoken Sinhala, for instance, due to heavy use of zero anaphora, gender-neutral pronouns ('eyaa'), deictic terms ('araya', 'meyaa', 'oyaa', 'eyaa'). Thus, this paper also discusses the possible difficulties that would arise in using Spoken Sinhala in CL.

Key words: Computational linguistics; Human language; Sociolinguistic variables; Literary Sinhala; Spoken Sinhala