Abstract 58

Developing a Dependency Tag Set for Sinhala: Procedure and Issues

Chamila Liyanage[1], Senior Prof. W. M. Wijeratne[2]

[1]Language Technology Research Laboratory, University of Colombo School of Computing
cml@ucsc.cmb.ac.lk

[2]Department of Linguistics, University of Kelaniya
hiru_89@yahoo.co.uk

Dependency Grammar (DG) is considered as one of the prominent theories of syntax. In order to analyze a particular language on DG and to make an annotated Dependency Treebank, a Tag set is needed. The objective of this research is to compile a Dependency Treebank for Sinhala. As part of compiling, the Treebank a Tagset was developed. This study is designed to explore the procedure and issues of developing a dependency tagset, with special focus to Sinhala Language. Methodology of the study includes 1. Identify same grammatical categories from benchmark tagsets 2. Find out syntactico-semantic categories from traditional Sinhala grammar books 3. Analyze sentences extracted from UCSC Sinhala corpus and further identify grammatical categories 4. Verify the tagset. In literature no reported work has been done based on DG for Sinhala. However, syntactic analysis on other grammatical traditions, Sinhala grammar books and several tagsets were referred in this work. Among the referred tag sets, Stanford typed dependencies manual (Marneffe and Manning, 2016) and AnnCorra: TreeBanks for Indian Languages-Guidelines for Annotating Hindi TreeBank (Bharati et al, 2012) were selected as benchmark tagsets. To ensure uniformity of the tagsets many tags for the same grammatical categories were taken from the above benchmark tag schemas. Findings of the research introduce syntactico-semantic categories and levels of dependency relations of words in Sinhala. The tagset comprises 42 tags and can be used in related works on DG for Sinhala.

Key words: Computational Grammar, Dependency Annotation, Dependency Tag Set, Sinhala Grammar, Sinhala Linguistics