

Title: Consistency of Linguistic Annotation

Author: Akshay Aggarwal

Institute: Institute of Formal and Applied Linguistics

Supervisors: RNDr. Daniel Zeman, PhD, Institute of Formal and Applied Linguistics; Koldo Gojenola, PhD, Computer Languages and Systems, University of Basque Country (UPV-EHU), Spain

Abstract: This thesis attempts at correction of some errors and inconsistencies in different treebanks. The inconsistencies can be related to linguistic constructions, failure of the guidelines of annotation, failure to understand the guidelines on annotator's part, or random errors caused by annotators, among others. We propose a metric to attest the POS annotation consistency of different treebanks in the same language, when the annotation guidelines remain the same. We offer solutions to some previously identified inconsistencies in the scope of the Universal Dependencies Project, and check the viability of a proposed inconsistency detection tool in a low-resource setting. The solutions discussed in the thesis are language-neutral, intended to work with multiple languages with efficiency.

Keywords: Annotation Consistency; Annotation Inconsistency; Error Mining; Language Independent; Universal Dependencies; UD Project; Syntax; Morphology