Building Arabic Corpus Applied to Part-of-Speech Tagging

Rabab Ali Abumalloh, Hassan Maudi Al-Sarhan, Waheeb Abu-Ulbeh

Objective: This paper aimed to review corpus linguistics sources related to part-of-speech tagging and to build up a sufficient annotated corpus for the Arabic language that contains Arabic words and their grammatical tags. Methods/ Statistical Analysis: An in-depth survey conducted by the author's showed that there is a need for free tagged Arabic corpus that can be used in natural language processing researches. A corpus of 25,000 words collected manually from different web sources which ware written in Modern Standard Arabic. The collected words were tagged using Arabic language grammar books. Findings: The developed corpus can help the researchers in natural language processing applications. Applications/Improvements: This corpus needed to be expanded to include more words and their grammatical tags.