# Urdu OCR Project

# Part of Speech Annotator v2.0

**June 24, 2012**

**CENTER FOR LANGUAGE ENGINEERING**
**UNIVERSITY OF ENGINEERING AND TECHNOLOGY, LAHORE**
**PAKISTAN**

# 1. Introduction

Annotator is an annotation tool to facilitate the process of manually tagging the text for part of speech (POS). The tool is supported for windows and Microsoft .NET Framework Version 4.0 http://www.microsoft.com/en-us/download/details.aspx?id=17718 is required to run it.

# 2. Annotator Tool

This section describes basic features of Annotator. The Annotator program has two components: Urdu POS Annotator.exe and tagset.txt. Urdu POS Annotator.exe is the main program. The tagset.txt file contains the tag set that will be used to annotate the text. The 1st line of tagset.txt contains the total number of tags in the file and remaining each line contains a single tag. Tagged files will be automatically saved in the Urdu POS Annotator TaggedFiles folder and the log word files will be saved in the Urdu POS Annotator LogFiles folder. Both these folder will have the same location as the directory of the current file opened in the Annotator.

## 2.1 Using Annotator

Double click the Urdu POS Annotator.exe to run the program. Click File→Open to open a text file for Annotation.

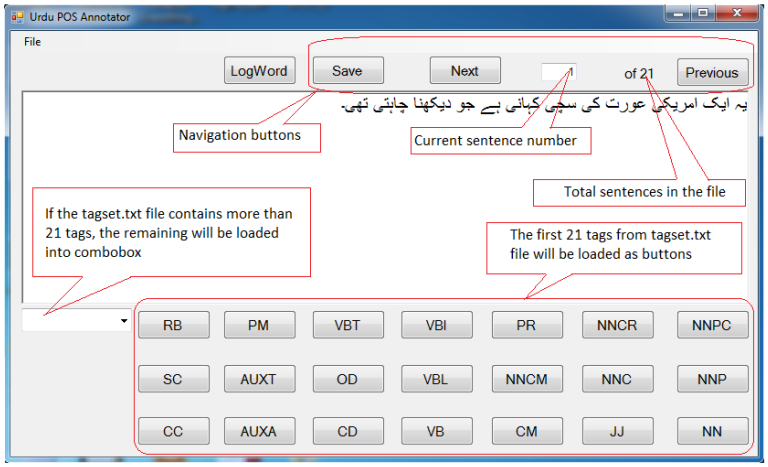


**Figure 1: Using Annotator**

The Annotator will show the 1st sentence of the opened file as shown in Figure 1. The file can be traversed using navigation buttons: Previous and Next; or by entering the sentence number in the current sentence number text field.

The first 21 tags from tagset.txt file will be loaded as buttons and if the tagset.txt file contains more than 21 tags the remaining will be loaded into combobox.

## 2.3 Tagged files Word log files

Tagged files will be automatically saved in the Urdu POS Annotator TaggedFiles folder. The current version of Urdu POS Annotator also facilitates the user with the word log option as well. The user can generate a log file in the Urdu POS Annotator LogFiles folder that will contain word, sentence and tag options on a separate line for each word for which the user clicks the LogWord button as shown in Figure 2.

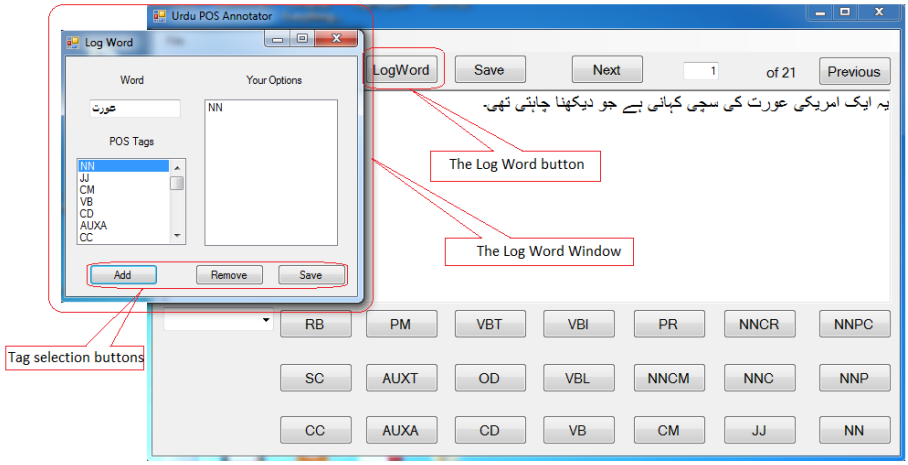


**Figure 2: The Log word window**

Each file opened in the Annotator will have its separate tagged and log file. The location of the folders will be the same as the file opened.