Corpus Editor Tool for Part of Speech (POS) Annotation

User Guide

June 2018

University of Moratuwa

Center for National Language Processing

Table of Contents

1 Introduction

This document explains how to use the Corpus Editor tool. It is a standalone web application that can be used to edit a POS-tagged Sinhala corpus. This tool can be used for:

- Correcting the tags of the words using word search
- Correcting the tags of the words using tag search
- Suggesting the tags for undefined words
- Producing a unique word list

2 Importance of the Tool

It is important in following aspects.

- Improving the quality and accuracy of a POS annotated corpus by correcting mis-tagged words
- Identifying new tags for the words in a language
- Identifying misspelled words

3 Running the Application

After downloading the html, follow the steps below to start the application.

I. Open the downloaded application
Right click on the downloaded html ⟶ Select open in folder option ⟶ Go to the location of downloaded ⟶ Double click and open / Right click and open

II. As the application starts, it shows main window as shown in the Figure 1.1

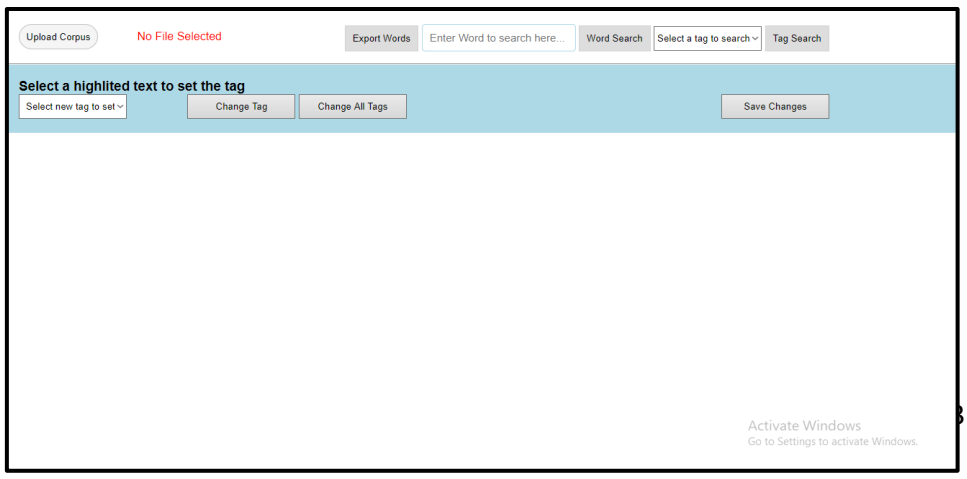

Figure 1.1 Corpus Editor at start up

4 Using the Tool

4.1 Upload a corpus/file

In order to upload a file, follow the steps given below.

a. Click the “upload corpus”. Then the open window will appear as shown in Figure 1.2

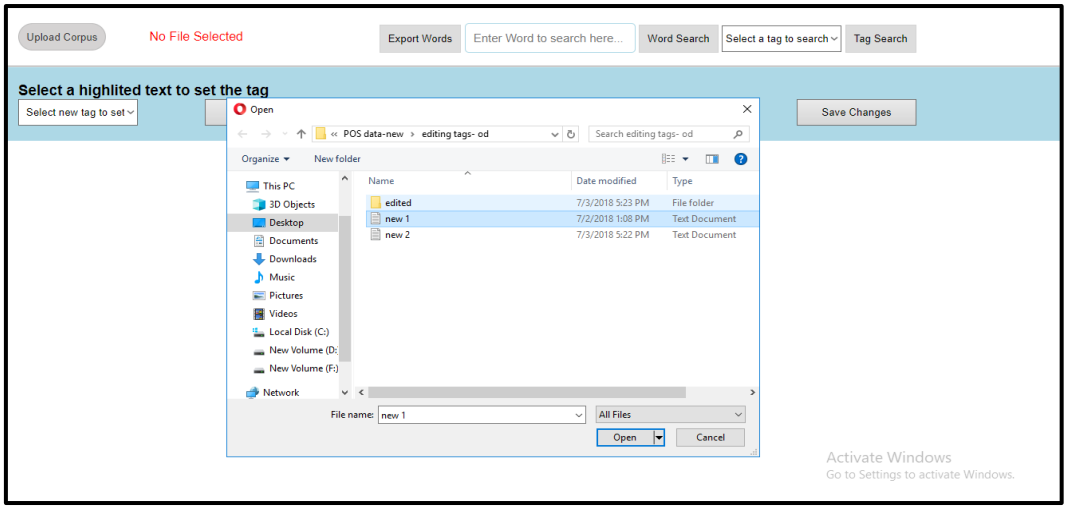

Figure 1.2 Uploading a file

b. Browse the files to be uploaded (the file should include the tagged data)
c. After uploading completed, a message appears as ‘file has uploaded’ as Shown in the Figure 1.3

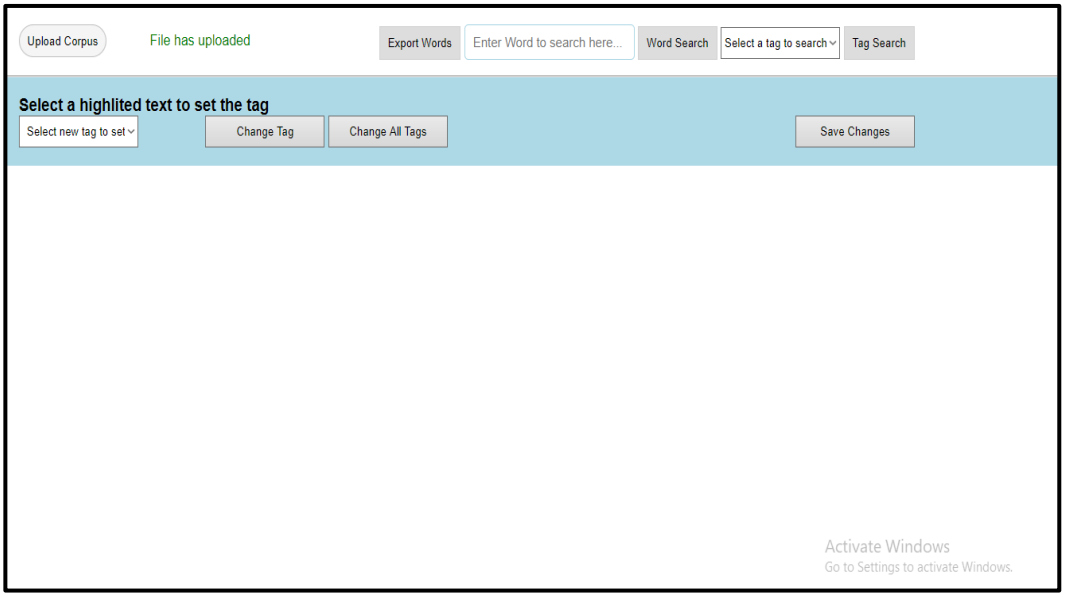

Figure 1.3 After uploading a file

4.2 Work Flow

The following work flow should be used for the edition.

STEP 1

- Generate and download a unique words list
    - Export Words ⟶ click DOWNLOAD link to download the file (below to the SAVE CHANGES button)
    - Double click on the “DOWNLOAD COMPLETE” box and open the unique words list (saved in the same folder as the corpus file). (Figure 4.2.1)

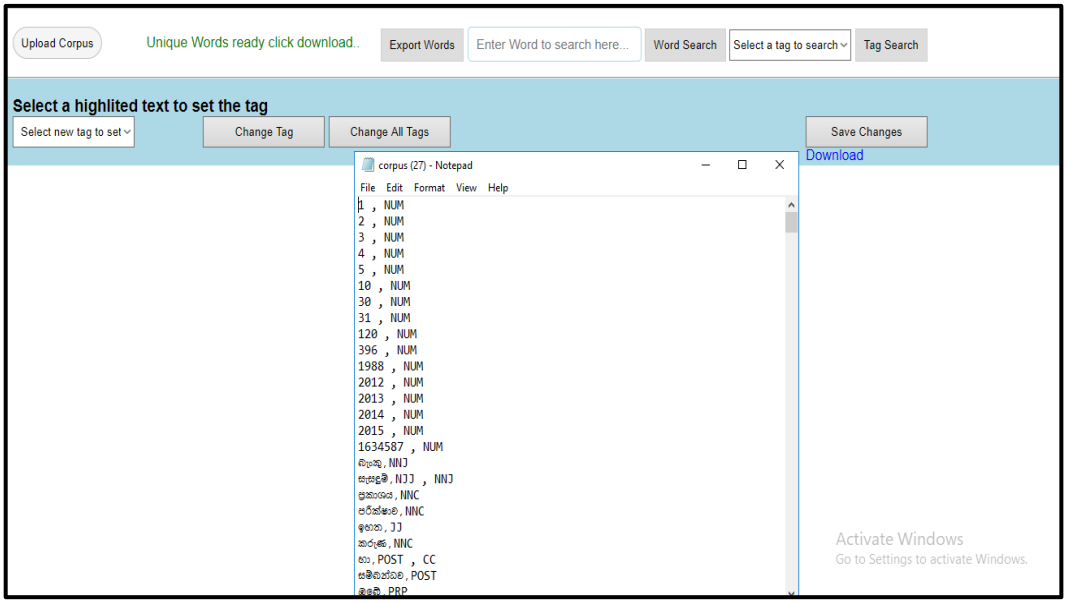

Figure 4.2.1 – Downloaded Unique words list

STEP 2

- Copy a word from the unique words list/ type a word and paste it in the box before the WORD SEARCH button
- Click on the WORD SEARCH button
- It shows all the tags in the corpus relevant to the word entered (Figure 4.2.2)

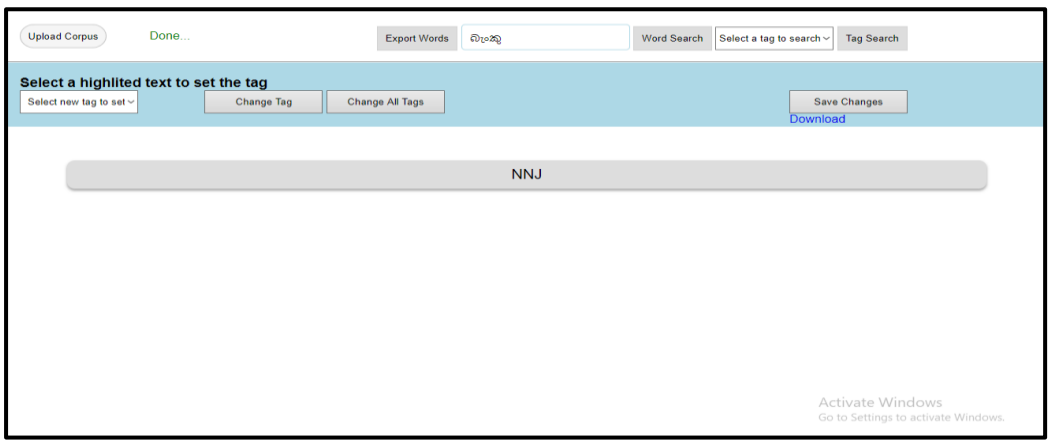

Figure 4.2.2 – Tags related to the searched word

- Click on the <TAG> button
- It explores all the data relevant to the searched word and highlights the relevant word (Figure 4.2.3)

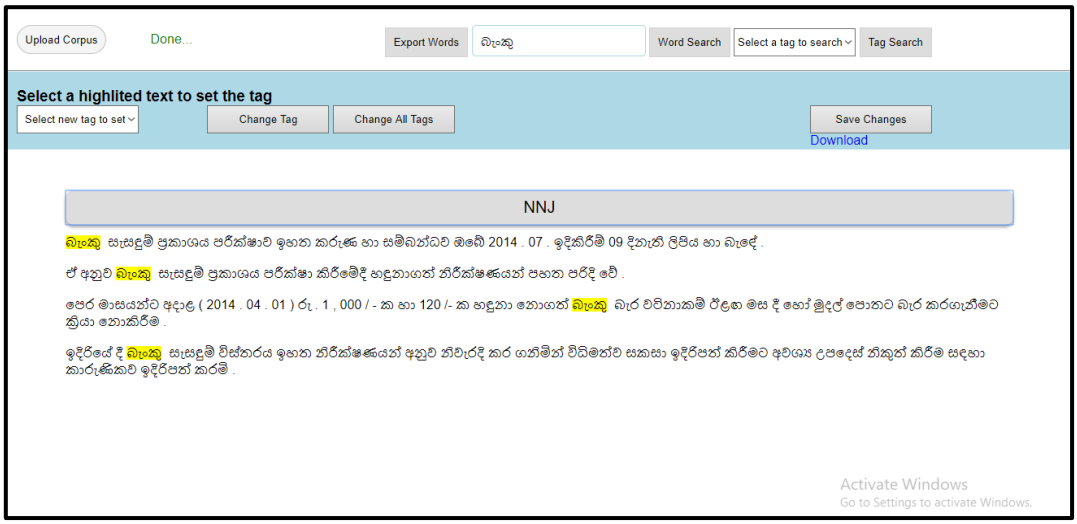

Figure 4.2.3 – Data related to the searched word

<u>Changing the tag of a word</u>

- If the given tag is incorrect, click on the yellow color highlighted word to set the tag
- Select the correct tag using the drop down list of tags (SELECT A TAG)in the left side
- Click on CHANGE TAG button
- Click on OK button after receiving the message "Are you sure you want to update?"
- All the tags relevant to the changed word is shown

<u>Changing the tag of all words at once</u>

- If the given tag is incorrect, select the correct tag using the drop down list of tags (SELECT A TAG)in the left side
- Click on CHANGE ALL TAGS button (Figure 4.2.4)
- Click on OK button after receiving the message "Are you sure you want to update all tags?"
- The changed tag is shown

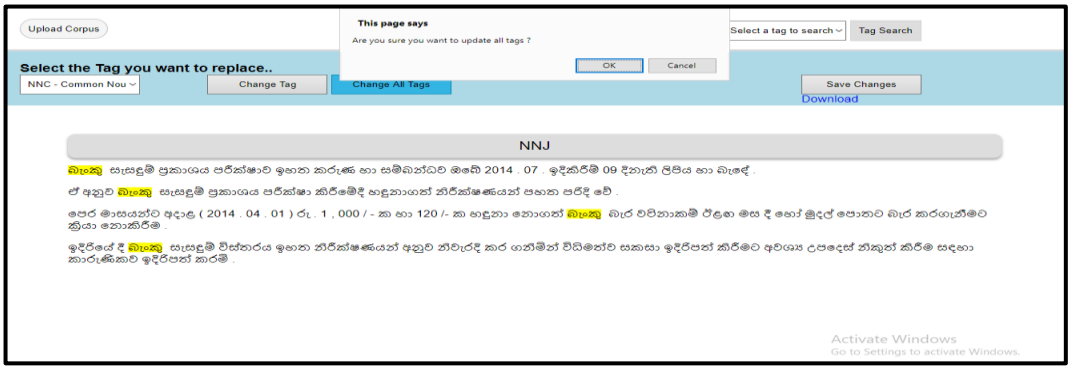

Figure 4.2.4 – changing all tags of a given word

STEP 3

Downloading the edited file

- After completing the editing, click on SAVE CHANGES button in the right side
- Click on DOWNLOAD link to download the file
- Double click on the downloaded file and open the file

STEP 4

To change the tag of a word using TAG SEARCH,

- Select a tag from the dropdown list next to the WORD SEARCH button
- Click on TAG SEARCH button
- It shows all the words related to the given tag. (Figure 4.2.5)

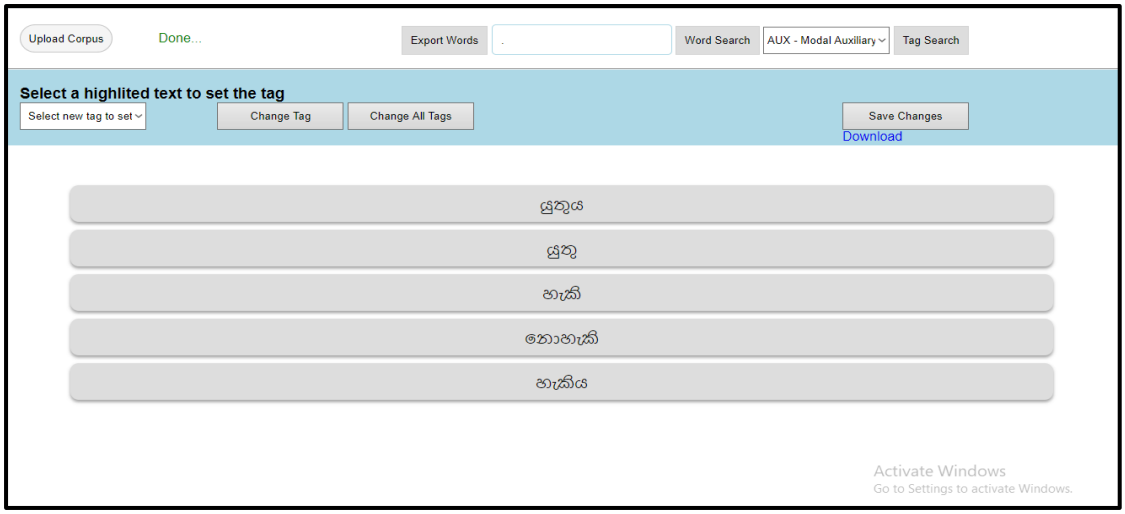

Figure 4.2.5 – selecting the words using tag search

- Follow the same steps mentioned above afterwards for editing

4.3 Undefined Tags

The words with some errors such as words without tags and words without space among word and tag are identified as undefined tags. These tokens should be edited into the standard format.

4.4 Editing the text

As this application is not supported for typing, the following steps should be followed.

- Open the downloaded edited file from the application using NotePad++
- Open the unique word list generated in Section 4.1 step 1. This is an ordered list of words with tags per each token. If you go through it, you will see that misspelled words are close to each other.
- Search the word to be edited in NotePad++
- Replace the correct word in the text (open the REPLACE box using Ctrl+ h/ search>find>replace tab ⟶ insert the incorrect word in the box called FIIND WHAT→ Insert the correct word in the box called REPLACE WITH) (Figure 4.3.1) / Use plugins compare→ compare→commands for comparing and editing (Figure 4.3.2).
- With all the changes, save the file

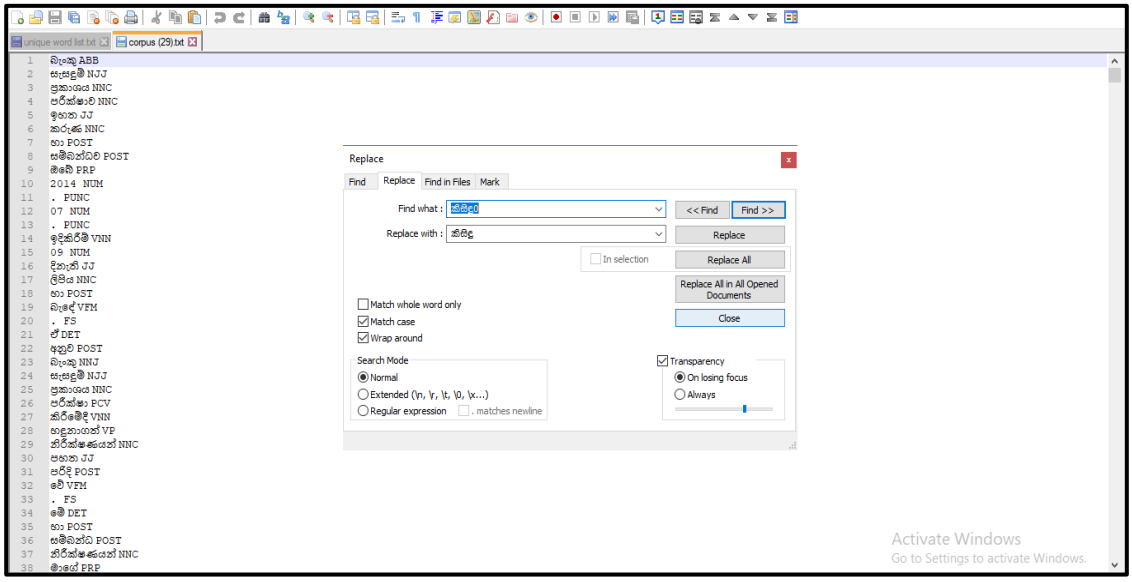

Figure 4.3.1 – Text editing using NotePad++

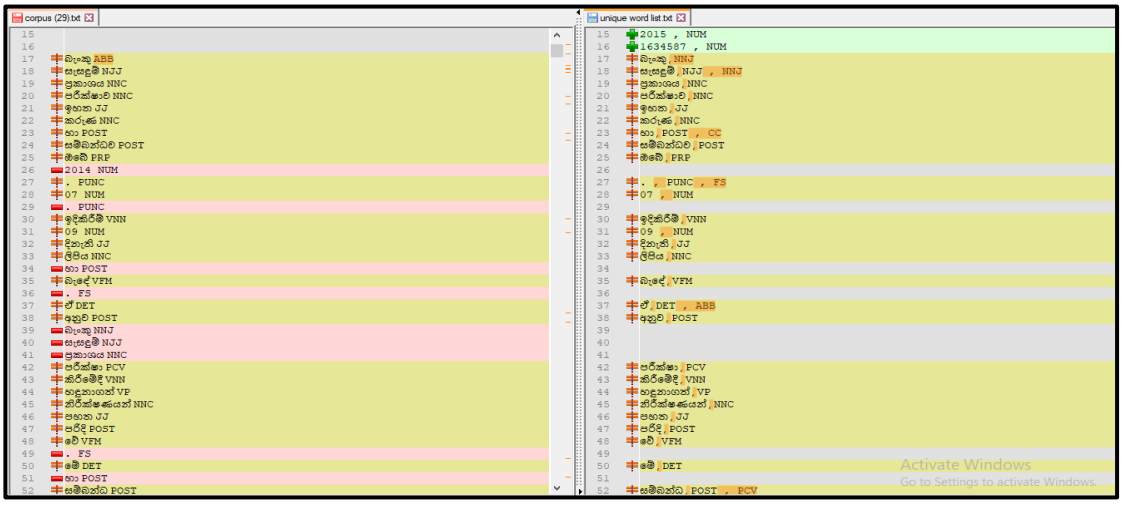

Figure 4.3.2 – Text editing using NotePad+