**SANTI-morf DOCUMENTATION v20201209**

SANTI (*Sistem Analisis Teks Indonesia*) is an ambitious project aiming to provide users with an integrated text analysis system for Indonesian. This documentation describes one of SANTI sub systems, SANTI-morf, an Indonesian morphological analysis system. SANTI-morf tokenises words into morphemes and assigns formal and functional morphological analytic labels to each morpheme. It is still a work in progress! Feel free to drop your suggestions, questions, criticisms, here: [prihantoro@live.undip.ac.id](mailto:prihantoro@live.undip.ac.id). Thank you!

--Prihantoro--

**SANTI-morf Tagset**

1. Formal morphological criteria analytic labels

Table

Description automatically generated

1. Functional analytic labels

A picture containing text, crossword puzzle, receipt

Description automatically generated

Table

Description automatically generated

**Requirement**

Nooj (<http://nooj4nlp.org/>)

**How to use**

1. Download Indonesian language resources via Nooj Preferences

Info > Preferences > Next to language name, choose ‘id’ > Download Module

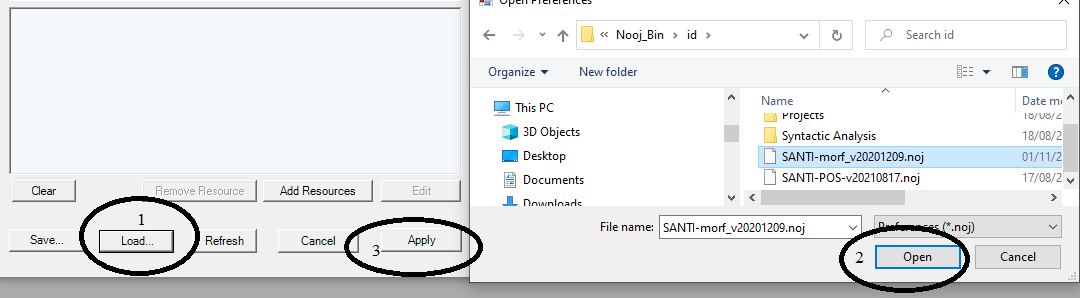
Graphical user interface

Description automatically generated

Or manually download them from here <https://drive.google.com/drive/folders/10XtOlLyW3tgX5SWaVLdFU_8inejbKi2G?usp=sharing>

1. Load SANTI-POS configuration file

Preferences> Load > Choose SANTI-morf-v20201209.noj > Open > Apply



1. Annotate your corpus

Import/Load Text/Corpus > right click > Linguistic Analysis

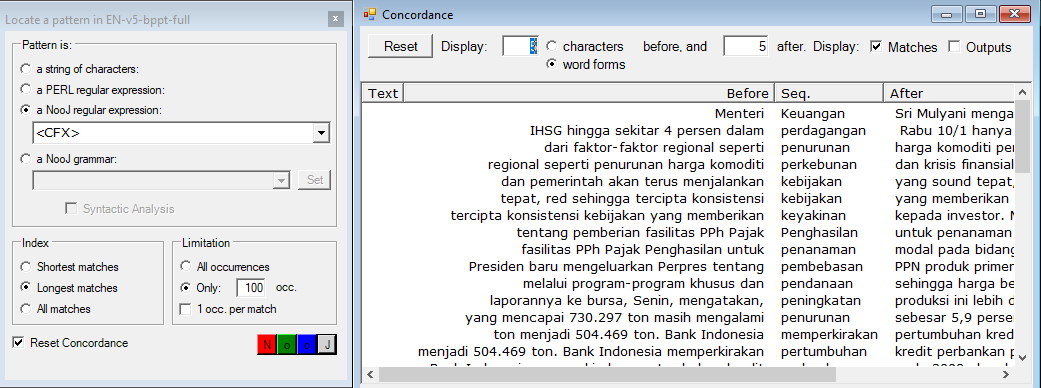
Graphical user interface, text

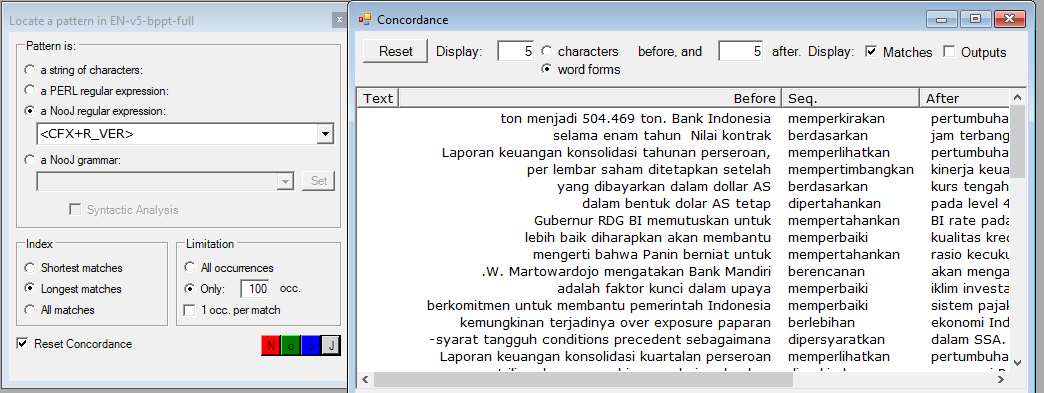
Description automatically generated

1. Searching with SANTI-morf tags

Right click on text > Locate Pattern > supply query

* A tag is a combination of a formal analytic label, or a formal + one/some functional analytic labels. In the result, you will get all words that contain your query. So, if you query with <CFX>, to retrieve words with circumfix, you will get, for example, *ke-uan-gan* and *mem-per-kira-kan.* This is because *ke—an* , and *peR—kan* contain tags which include CFX label.





**Corpus**

The corpus that comes with the resources is BPPT-PAN Corpus (EN1-v4-50Lines.not). But you can use your own corpus if you want.

**Citation**

Prihantoro. (2021). *SANTI-morf: A new morphological annotation system for Indonesian: Unpublished thesis (in progress)*. Lancaster: Lancaster University Press

Riza, H., & Hakim, C. (2009). Resource report: building parallel text corpora for multi-domain translation system. In *Proceedings of the 7th Workshop on Asian Language Resources (ALR7)* (pp. 92-95).