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. Thank you!

--Prihantoro--

SANTI-morf Tagset

1. Formal morphological criteria analytic labels

	Classification	Tag	Description
1	Root	ADJ	Adjective
2	POS	ADV	Adverb
3		ART	Article
4		BOU	Precategorial
5		CLS	Classifier
6		CNJ	Conjunction
7		ITJ	Interjection
8		NOU	Noun
9		NUM	Numeral
10		PCL	Particle
11		PRE	Preposition
12		PRO	Pronoun
13		VER	Verb
14		FRG	Foreign
15		PUNC	Punctuation
16		DGT	Numeric digit

2. Functional analytic labels

17	Outcome	R_ADJ	Adjective
18	POS	R_ADV	Adverb
19		R_CNJ	Conjunction
20		R_NOU	Noun
21		R_NUM	Numeral
22		R_VER	Verb
23	Others	ACV	Active
24		PSV	Passive
25		RECP	Reciprocal
26		RFLX	Reflexive
27		APPL	Applicative
28		CAUS	Causative
29		EQTV	Equative adjective
30		SPV	Superlative adjective
31		ITRV	Iterative aspect
32		RAND	Random unordered event
33		DEF	Definite
34		NYA	nya (depending on how it functions)

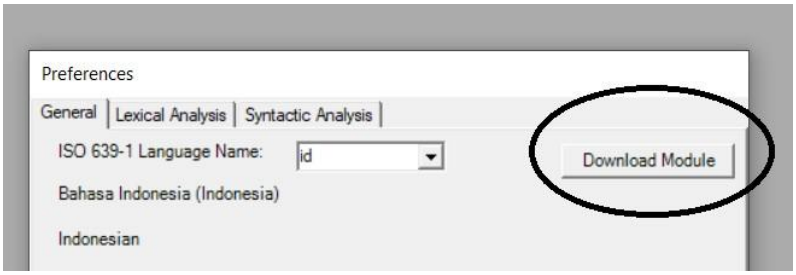
Requirement

Nooj (http://nooj4nlp.org/)

How to use

1. Download Indonesian language resources via Nooj Preferences

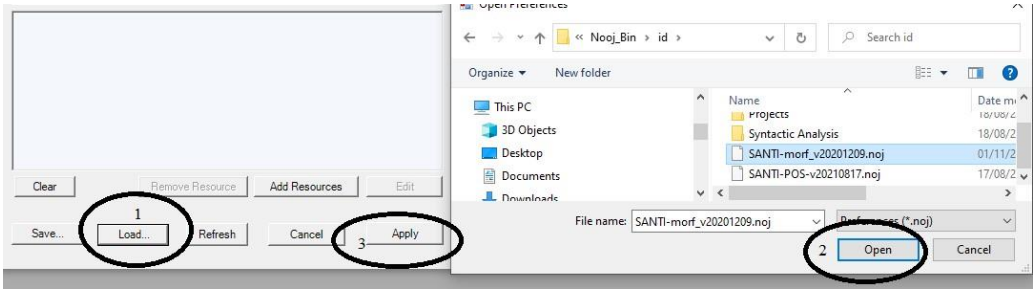
Info > Preferences > Next to language name, choose ‘id’ > Download Module



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

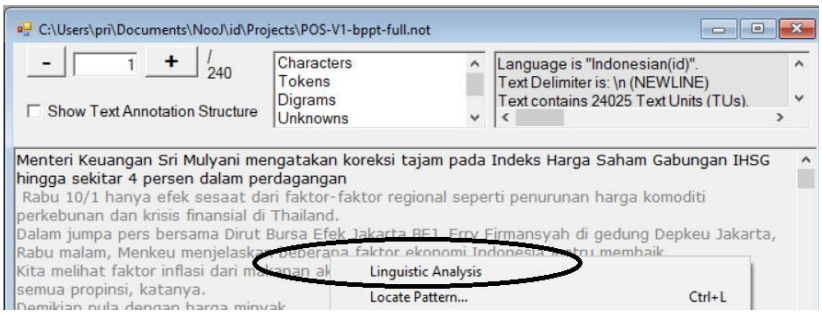
2. Load SANTI-POS configuration file

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



3. Annotate your corpus

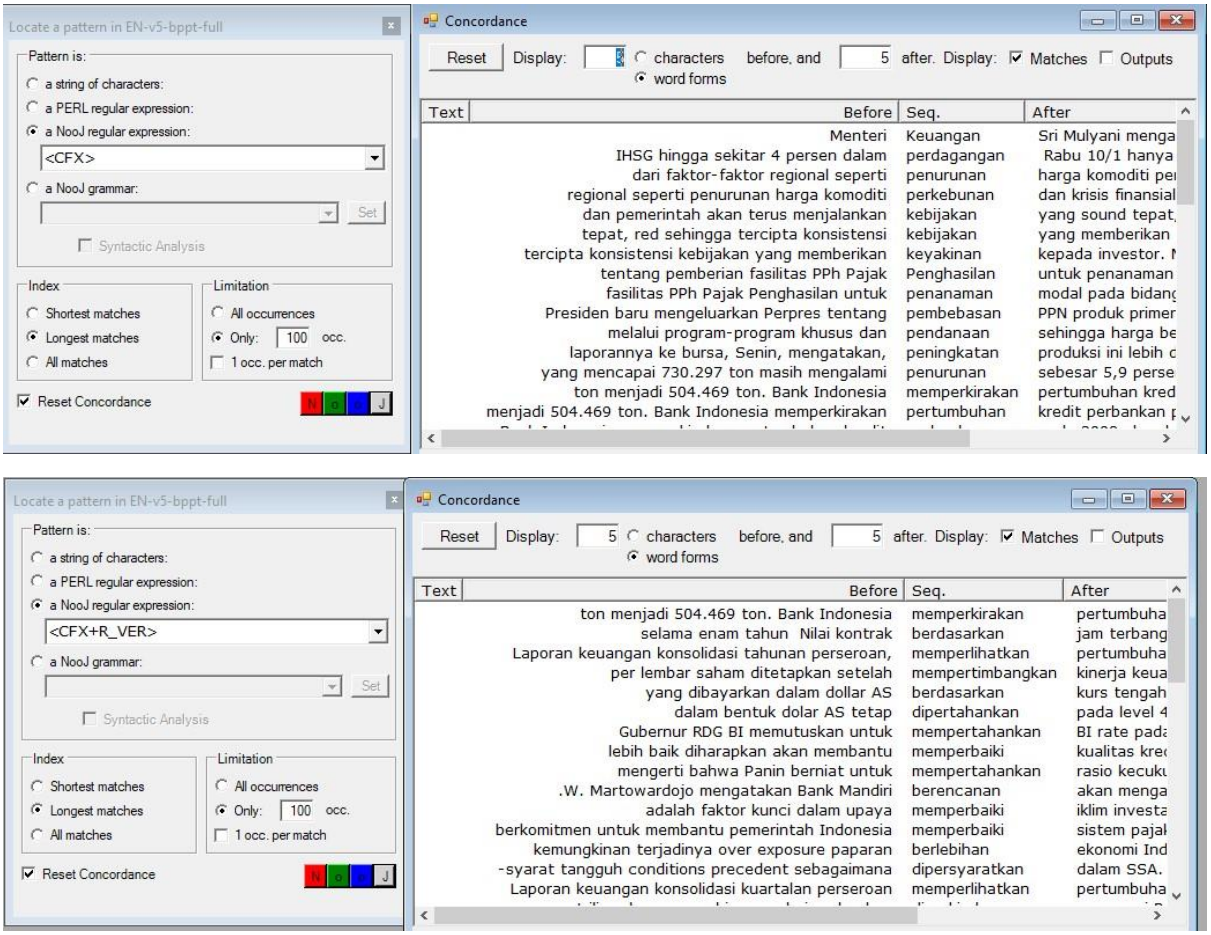
File > Open > Text/Corpus > choose text/corpus file(s) > Open > right click on Text/Corpus > Linguistic Analysis



4. 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.

Exporting to .txt document

Open your Text/Corpus > Right Click > Export annotated text as XML format > Rename the resulting not.xml.txt text file into input.txt > run convert.php > see output.txt for the final result

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).