



---

PAN LOCALIZATION PROJECT

---

# DEVELOPING POSTAG FOR BAHASA INDONESIA

Mirna Adriani

Information Retrieval Lab

Faculty of Computer Science

University of Indonesia, Indonesia

13 Januari 2009

Fakultas Ilmu Komputer – Universitas Indonesia



# PAN Localization project

- Join the project in July 2008
- Started in September 2008
  - University of Indonesia (Mirna Adriani)
    - Developing POSTAG
    - Parallel Corpus : English Indonesian (Penn Treebank Corpus)
  - BPPT (Hammam Riza)
    - Parallel Corpus (Indonesian-English)
    - Machine Translation



# Bahasa Indonesia

- Bahasa Indonesia is the national language in Indonesia (used in school, offices etc.)
  - We have 742 local languages (many of them have different characters)
- Use roman characters
  - No tenses
  - Inflection words (prefix, postfix, suffix)



---

# POSTAG for Bahasa Indonesia

---

- No available POSTAG
- We build Tagset based on Penn Treebank tagset
  - Consult Indonesian grammar books
  - Problems : many inconsistency tags
- Our tagset has 29 POStags



# Tagset for Bahasa Indonesia

<b>Number</b>	<b>Category</b>	<b>Post-name</b>	<b>Postag</b>
1	Noun	Countable Common Noun	NNC
2	Noun	Uncountable Common Noun	NNU
3	Noun	Genitive Common Noun	NNG
4	Noun	Proper Common Noun	NNP



# Tagset for Bahasa Indonesia

<b>Number</b>	<b>Category</b>	<b>Post-name</b>	<b>Postag</b>
5	Verb	Transitive	VBT
6	Verb	Intransitive	VBI
7	Verb	Modal	MD
8	Adjective	Adjective	JJ



# Tagset for Bahasa Indonesia

<b>Number</b>	<b>Category</b>	<b>Post-name</b>	<b>Postag</b>
9	Adverb	Adverb	RB
10	Wh-Adverb		WRB
11	Preposition		IN
12	Conjunction	Coordinate Conjunction	CC
13	Conjunction	Subordinate Conjunction	SC



# Tagset for Bahasa Indonesia

<b>Number</b>	<b>Category</b>	<b>Post-name</b>	<b>Postag</b>
14	Pronoun	Personal Pronoun	PRP
15	Pronoun	Wh-Pronoun	WP
16	Pronoun	Number Pronoun	PRN
17	Pronoun	Locative Pronoun	PRL
18	Interjection		UH



# Tagset for Bahasa Indonesia

<b>Number</b>	<b>Category</b>	<b>Post-name</b>	<b>Postag</b>
19	Punctuation		PUN
20	Symbol		SYM
21	Determiner		DT
22	Determiner	Wh-Determiner	WDT
23	Particle		RP



# Tagset for Bahasa Indonesia

<b>Number</b>	<b>Category</b>	<b>Post-name</b>	<b>Postag</b>
24	Cardinal Numeral	Primary Numeral	CDP
25	Cardinal Numeral	Ordinal Numeral	CDO
26	Cardinal Numeral	Irregular Numeral	CDI
27	Cardinal Numeral	Collective Numeral	CDC
28	Negation	Wh-Determiner	NEG
29	Foreign Word		FW



# Example

- Spalletti/**NNP**,/, meski/**CC** belum/**NEG** seterancam/**RB** Ranieri/**NNP** juga/**RB** harus/**MD** mulai/**VBI** memikirkan/**VBT** cara/**NNC** mendongkrak/**VBT** Roma/**NNP** kembali/**RB**./.  
Menang/**NNC** melawan/**VBT** Chelsea/**NNP** akan/**MD** jadi/**VBT** tugas/**NNP** teramat/**RB** berat/**JJ**,/, tapi/**CC** partai/**NNC** away/**FW** melawan/**VBT** Udinese/**NNP** pekan/**NNC** depan/**IN** bisa/**MB** jadi/**VBT** awal/**NNC** baik/**JJ** bagi/**IN** Spalletti/**NNP** untuk/**IN** mengubah/**VBT** peruntungan/**NNU** dirinya/**NNG** dan/**CC** klubnya/**NNG**



# Evaluation

- Corpus : 50 documents
  - 694 sentences (13,465 tokens)
  - The words have been tagged manually
  
- Preliminary result
  - CRF : 70% accuracy
  - Brill tagger : 80% accuracy



---

# Thank You

---