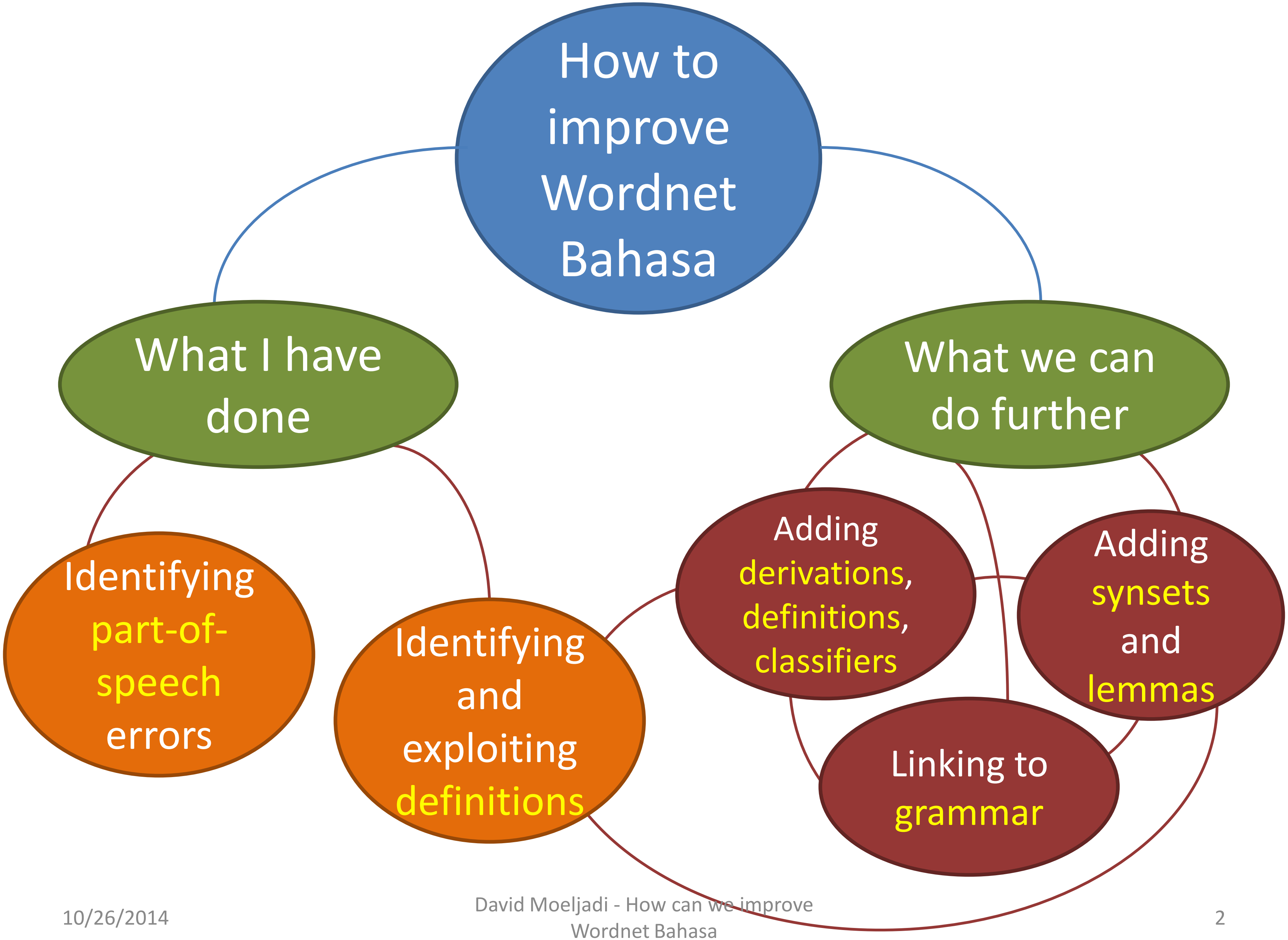
How can we improve Wordnet Bahasa (وردنيت بهاس)?

David Moeljadi
(دافيد موليادي)

Nanyang Technological University

October 26, 2014



How to improve Wordnet Bahasa

What I have done

Identifying part-of-speech errors

Identifying and exploiting definitions

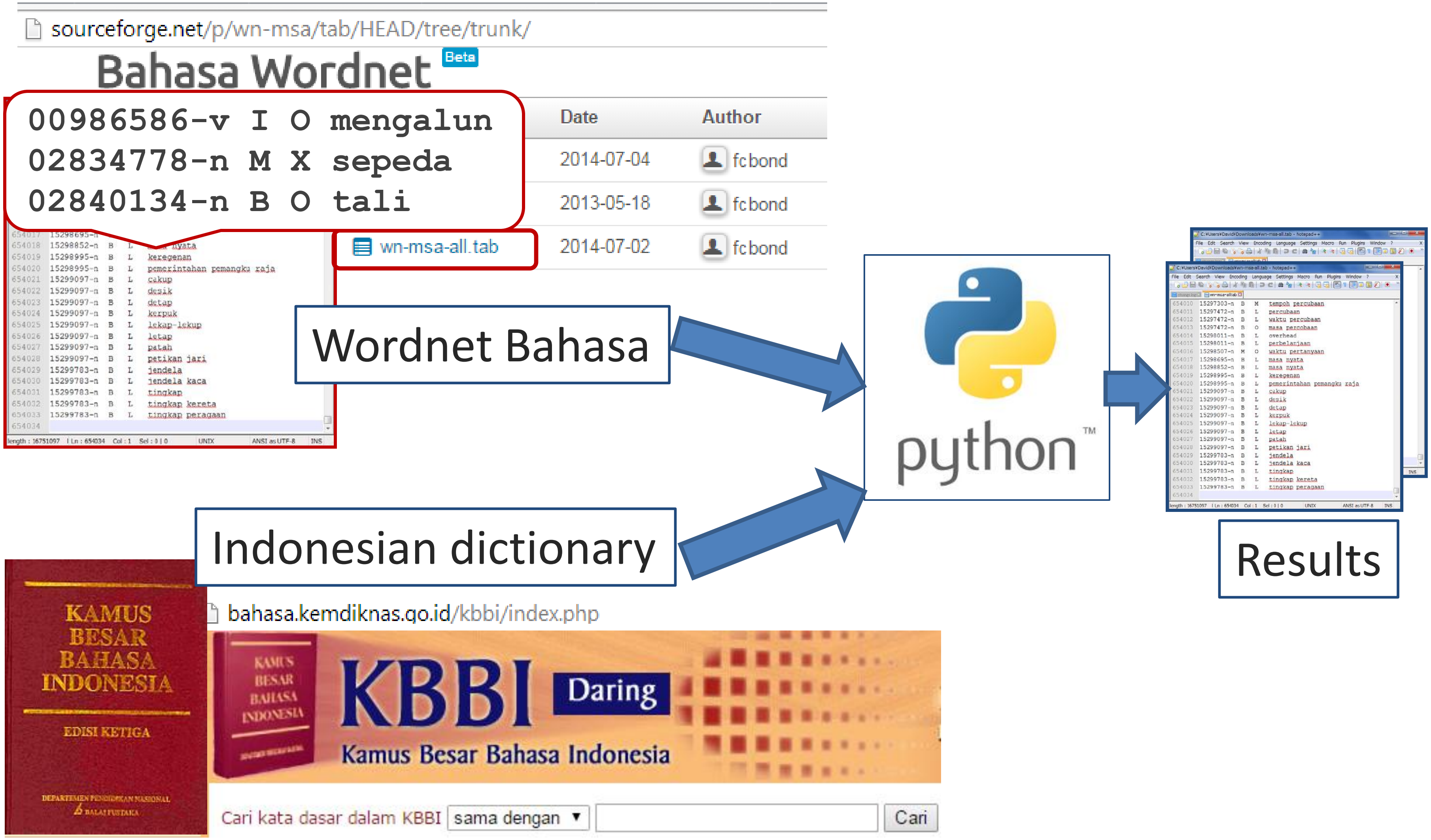
What we can do further

Adding derivations, definitions, classifiers

Adding synsets and lemmas

Linking to grammar

Identifying part-of-speech errors (part 1)



sourceforge.net/p/wn-msa/tab/HEAD/tree/trunk/

Bahasa Wordnet Beta

```
00986586-v I O mengalun
02834778-n M X sepeda
02840134-n B O tali
```

	Date	Author
	2014-07-04	fcbond
	2013-05-18	fcbond
wn-msa-all.tab	2014-07-02	fcbond

```
15298695-n
15298852-n  B  L   nyata
15298995-n  B  L   keregenan
15298995-n  B  L   pemerintahan pemangku raja
15299097-n  B  L   cakup
15299097-n  B  L   desik
15299097-n  B  L   detap
15299097-n  B  L   kerpuk
15299097-n  B  L   lekap-lekup
15299097-n  B  L   letap
15299097-n  B  L   patah
15299097-n  B  L   petikan jari
15299783-n  B  L   jendela
15299783-n  B  L   jendela kaca
15299783-n  B  L   tingkap
15299783-n  B  L   tingkap kereta
15299783-n  B  L   tingkap peragaan

length : 16751097   Ln : 654034   Col : 1   Sel : 0 | 0   UNIX   ANSI as UTF-8   INS
```

Wordnet Bahasa

Indonesian dictionary

C:¥Users¥David¥Downloads¥wn-msa-all.tab - Notepad++

File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
wn-msa-all.tab

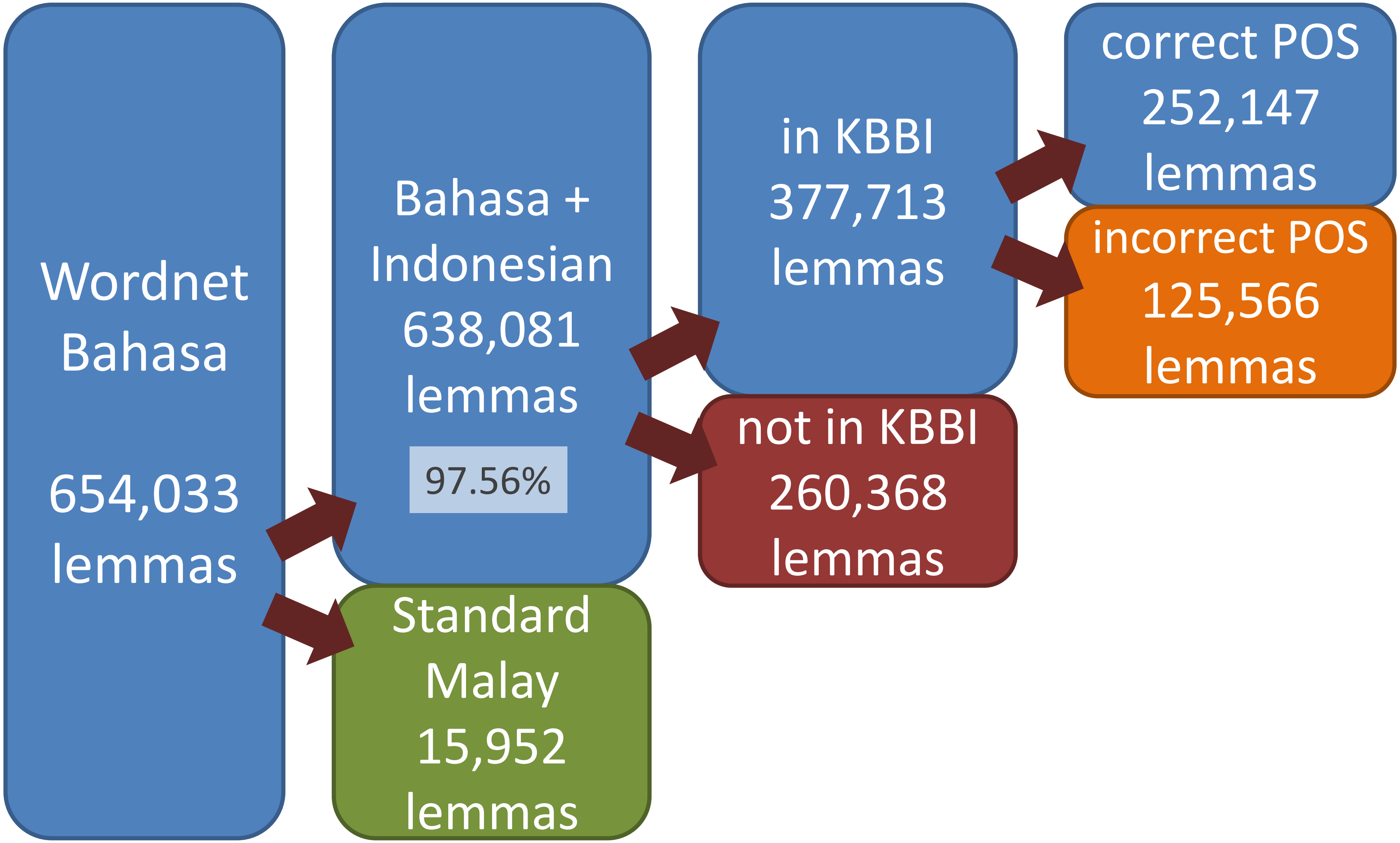
```
15297303-n  B  M   tempoh percubaan
15297472-n  B  L   percubaan
15297472-n  B  L   waktu percubaan
15297472-n  B  O   masa percobaan
15298011-n  B  L   overhead
15298011-n  B  L   perbelanjaan
15298507-n  M  O   waktu pertanyaan
15298695-n  B  L   masa nyata
15298852-n  B  L   masa nyata
15298995-n  B  L   keregenan
15298995-n  B  L   pemerintahan pemangku raja
15299097-n  B  L   cakup
15299097-n  B  L   desik
15299097-n  B  L   detap
15299097-n  B  L   kerpuk
15299097-n  B  L   lekap-lekup
15299097-n  B  L   letap
15299097-n  B  L   patah
15299097-n  B  L   petikan jari
15299783-n  B  L   jendela
15299783-n  B  L   jendela kaca
15299783-n  B  L   tingkap
15299783-n  B  L   tingkap kereta
15299783-n  B  L   tingkap peragaan

length : 16751097   Ln : 654034   Col : 1   Sel : 0 | 0   UNIX   ANSI as UTF-8   INS
```

Results

bahasa.kemdiknas.go.id/kbbi/index.php

Cari kata dasar dalam KBBI [sama dengan ▼] [ ] [Cari]

Identifying part-of-speech errors (part 1)



Wordnet Bahasa

654,033 lemmas

Bahasa + Indonesian 638,081 lemmas

97.56%

Standard Malay 15,952 lemmas

in KBBI 377,713 lemmas

not in KBBI 260,368 lemmas

correct POS 252,147 lemmas

incorrect POS 125,566 lemmas

Identifying part-of-speech errors (part 1)

not in KBBI
260,368
lemmas

komisen “commission”
majoriti “majority”
party
due date
memindahkan organ “
tidak diberi makan “no
dihapuskan “be erased
digunakan “be used”
hentikan “stop it!”
John
Kabul
bernafas “to breathe”
y2k, hb

Standard Malay lemmas

English lemmas

not lexicalized lemmas

passive-form lemmas

imperative-form lemmas

personal/place name lemmas

not standard Indonesian lemmas

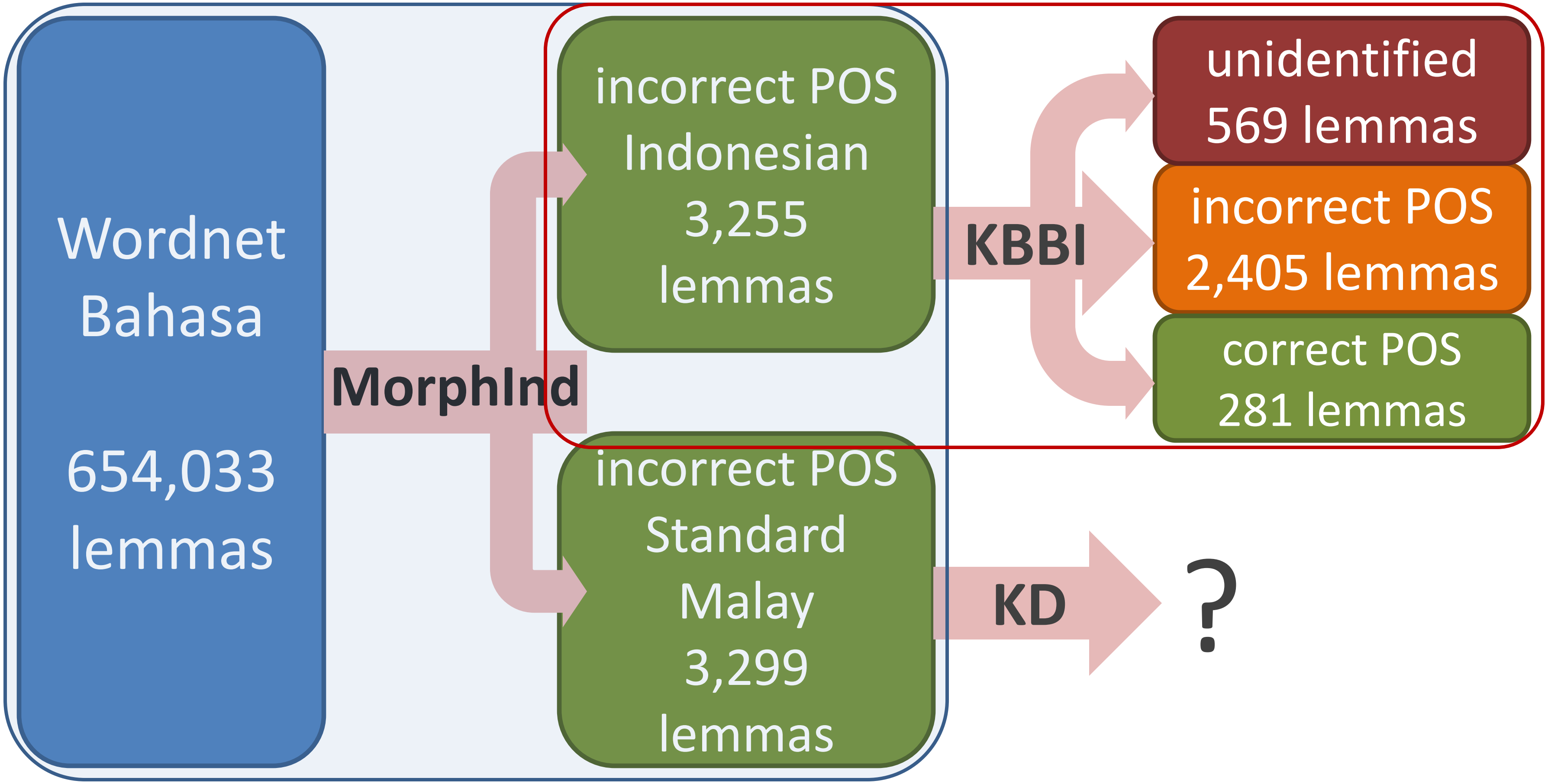
abbreviation lemmas

incorrect POS
125,566
lemmas

Recheck the sense tag

Identifying part-of-speech errors (part 2)

Muhammad Zulhelmy Mohd Rosman (2013)



Wordnet Bahasa

654,033 lemmas

MorphInd

incorrect POS Indonesian 3,255 lemmas

incorrect POS Standard Malay 3,299 lemmas

KBBI

unidentified 569 lemmas

incorrect POS 2,405 lemmas

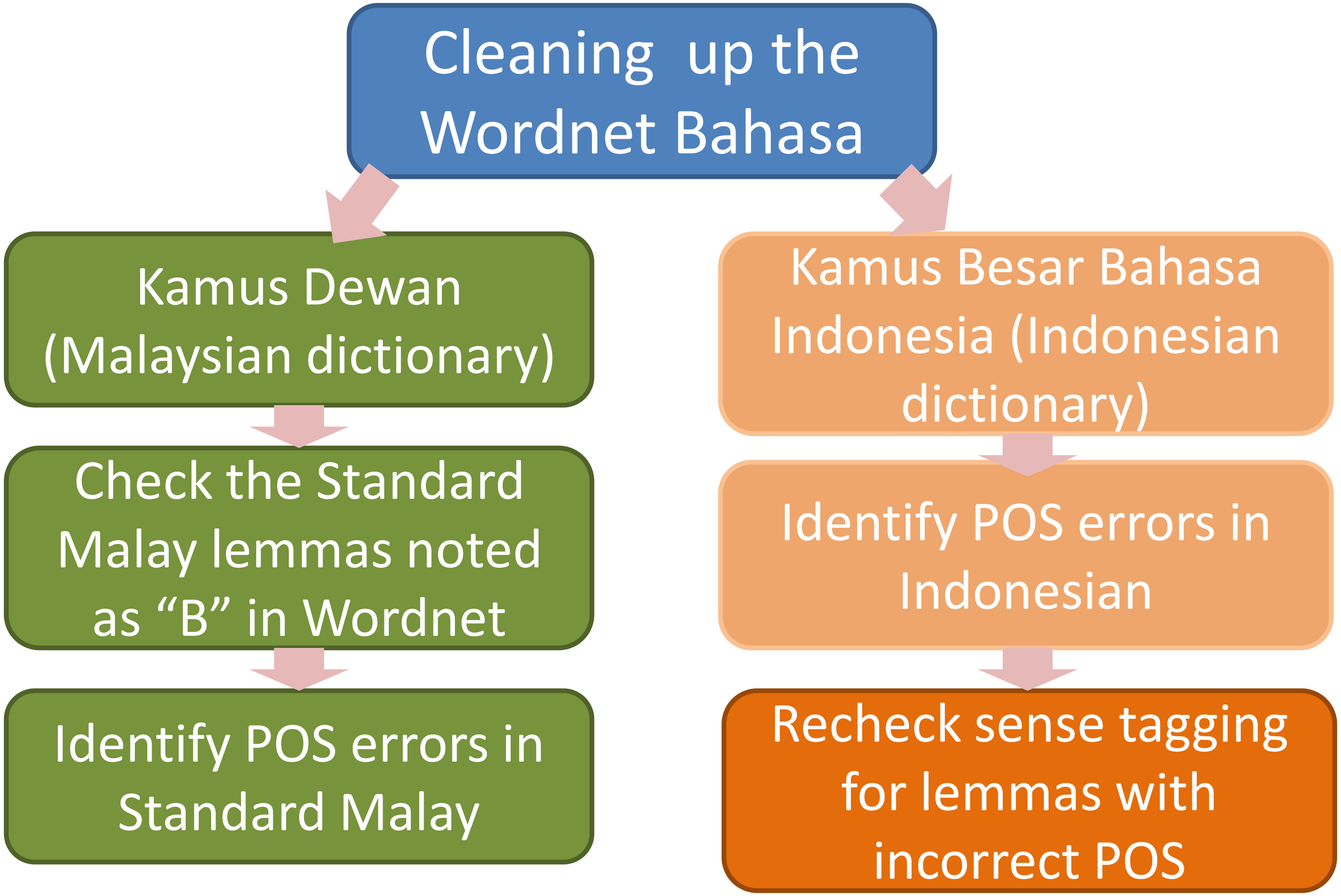
correct POS 281 lemmas

KD

?

MorphInd (Larasati, Kubon, and Zeman 2011)

Identifying part-of-speech errors



Cleaning up the Wordnet Bahasa

Kamus Dewan (Malaysian dictionary)

Check the Standard Malay lemmas noted as “B” in Wordnet

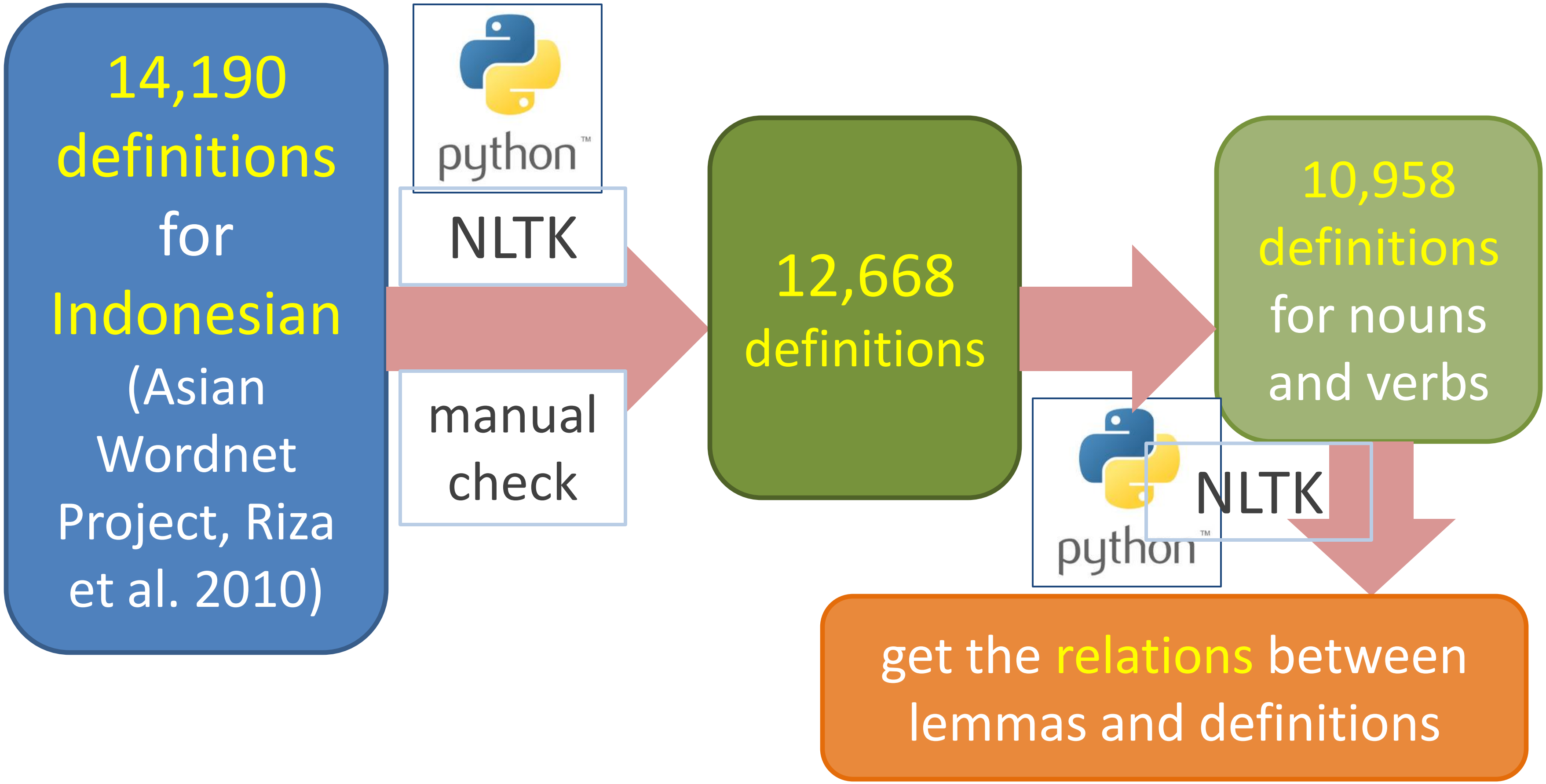
Identify POS errors in Standard Malay

Kamus Besar Bahasa Indonesia (Indonesian dictionary)

Identify POS errors in Indonesian

Recheck sense tagging for lemmas with incorrect POS

Identifying and exploiting definitions



14,190 definitions for Indonesian (Asian Wordnet Project, Riza et al. 2010)

python NLTK manual check

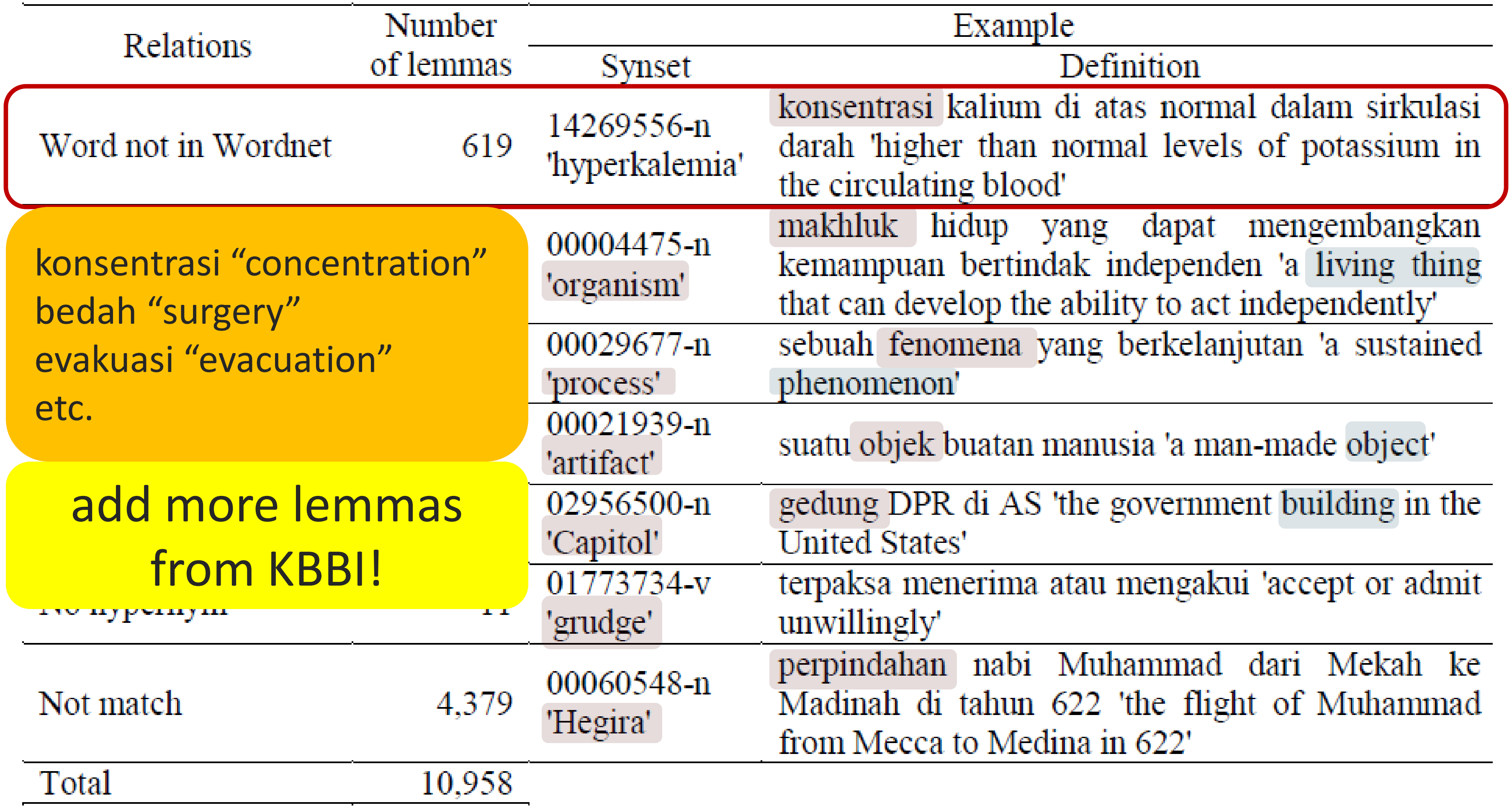
12,668 definitions

python NLTK

10,958 definitions for nouns and verbs

get the relations between lemmas and definitions

Identifying and exploiting definitions



Relations	Number of lemmas	Example	
		Synset	Definition
Word not in Wordnet	619	14269556-n 'hyperkalemia'	konsentrasi kalium di atas normal dalam sirkulasi darah 'higher than normal levels of potassium in the circulating blood'
		00004475-n 'organism'	makhluk hidup yang dapat mengembangkan kemampuan bertindak independen 'a living thing that can develop the ability to act independently'
		00029677-n 'process'	sebuah fenomena yang berkelanjutan 'a sustained phenomenon'
		00021939-n 'artifact'	suatu objek buatan manusia 'a man-made object'
		02956500-n 'Capitol'	gedung DPR di AS 'the government building in the United States'
No hypernym	11	01773734-v 'grudge'	terpaksa menerima atau mengakui 'accept or admit unwillingly'
Not match	4,379	00060548-n 'Hegira'	perpindahan nabi Muhammad dari Mekah ke Madinah di tahun 622 'the flight of Muhammad from Mecca to Medina in 622'
Total	10,958		

konsentrasi "concentration"
bedah "surgery"
evakuasi "evacuation"
etc.

add more lemmas from KBBI!

What we can do further?

1. Adding derivations, definitions, classifiers

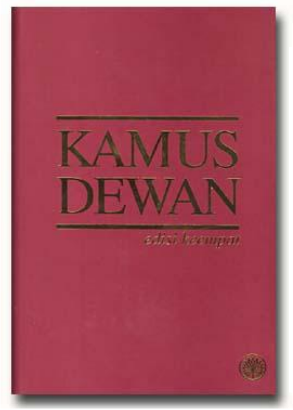
Using MorphInd or other morphological analyser?

Add more (clean) definitions with license

2. Adding synsets and lemmas

78,000 lemmas

90,000 lemmas



49,000 lemmas

Gerbang Maya
Dewan Bahasa dan Pustaka Malaysia
Carian
PRPM@DBPMalaysia
PUSAT RUJUKAN PERSURATAN MELAYU
Carian Spesifik
Cari

3. Linking to grammar (POS, verb frames)

References

- Bird, Steven, Edward Loper, and Ewan Klein (2009) Natural Language Processing with Python. O' Reilly Media Inc. Retrieved April 29, 2014, from http://www.nltk.org/book/
- Bond, Francis and Kyonghee Paik (2012) A survey of wordnets and their licenses. In *Proceedings of the Sixth Global WordNet Conference (GWC 2012)* pp 64-71. Matsue.
- Bond, Francis and Ryan Foster (2013) Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013* pp 1352-1362. Sofia.
- Fellbaum, Christiane (2005) WordNet and wordnets. In: Brown, Keith et al. (eds.) *Encyclopedia of language and linguistics*, second edition, pp 665-670. Oxford: Elsevier. Retrieved May 1, 2014, from http://wordnet.princeton.edu
- Kamus Besar Bahasa Indonesia dalam jaringan (KBBI Daring). Based on *Kamus besar bahasa Indonesia* (The great dictionary of the Indonesian language). Third edition. Retrieved May 1-5, 2014, from http://pusatbahasa.kemdiknas.go.id/kbbi/
- Larasati, S., Kubon, V., & Zeman, D. (2011) Indonesian morphology tool (MorphInd): Towards an Indonesian corpus. In *Systems and frameworks for computational morphology* (Vol. 100, p. 119-129). Springer Berlin Heidelberg.
- Muhammad Zulhelmy Mohd Rosman (2013) *Creating derivational morphology links in Wordnet Bahasa*, Final Year Project, Linguistics and Multilingual Studies, Nanyang Technological University, Singapore
- Nurril Hirfana Mohamed Noor, Suerya Sapuan and Francis Bond (2011) Creating the open Wordnet Bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)* pp 258-267. Singapore.
- *Python v2.7.6 documentation* (2014) 27.1. collections —High-performance container datatypes. Retrieved May 6, 2014, from https://docs.python.org/2/library/collections.html
- Riza, Hammam, Budiono, and Chairil Hakim (2010) Collaborative work on Indonesian WordNet through Asian WordNet (AWN). In *Proceedings of the 8th Workshop on Asian Language Resources*, 9-13. Beijing, China. Asian Federation for Natural Language Processing.

wORDnet
bahAsa

Terima kasih
تريما كاسيه
Thank you