

# COMPILING A MULTILINGUAL BIODIVERSITY NAME REGISTER FROM MONOLINGUAL DICTIONARIES, WORDNETS, CHECKLISTS AND WIKIDATA

---

Dr LIM Lian Tze ([liantze@gmail.com](mailto:liantze@gmail.com))

The Second Wordnet Bahasa Workshop

Malaysia

1. Introduction
2. Aligning Dictionaries and Wordnet via Scientific Names
3. More from Wikidata & WikiMedia
4. Modeling Regional Preferences in WordNet?

## INTRODUCTION

---

- Malay peninsula, Borneo islands → rich biodiversity
- Local languages are rich with names for flora & fauna
- How might locals and visitors know they're referring to the same organism (plant/flower/insect/fruit/animal)?

# COMMON NAMES CAN BE AMBIGUOUS!

Example from (Hussey, Wilkinson & Tweddle, 2006):

North America

Europe

Moose



*Alces alces*  
(Linnaeus, 1758)

Elk

Elk



*Cervus elaphus*  
(Linnaeus, 1758)

Wapiti

- (Checklist = a list of species names)
- Align Malay and English common names of flora and fauna
  - Using scientific names as pivot to avoid ambiguity
  - Potentially other languages later
- So that we can
  - Get more translations for B. Malaysia botanical & zoological names
  - Add more B. Malaysia items in Wordnet Bahasa
  - Mine definition, images, etc from other open data sources

## SOME QUICK NOTES

- Scientific name: from Latin, classic Greek
- a.k.a. binomial name (binomen), trinomial name

<i>Buteo</i>	<i>jamaicensis</i>	<i>borealis</i>
genus	species	subspecies/infraspecies

- Several scientific names may refer to the same organism
  - Nomenclature change as study and understanding advances
  - Nomenclature may differ in different domains/fields
  - *One* **accepted** scientific name
  - Others are **synonyms**

# ALIGNING DICTIONARIES AND WORD- NET VIA SCIENTIFIC NAMES

---



- Kamus Dewan (KD; Hajah Noresah, 2004)
  - Authoritative dictionary for Bahasa Malaysia
  - Published by Dewan Bahasa & Pustaka
- Text Encoding Initiative (TEI; TEI Consortium, 2015)
  - Guidelines for encoding machine-readable texts
  - Humanities, social sciences and linguistics
  - Guidelines for different genres, e.g. dictionaries
- KD macro- and micro-structures annotated with TEI
  - In collaboration with The Name Technology Sdn Bhd in 2014
  - Used for research with permission

## EXAMPLE CONTAINING BINOMEN

```
<entry xml:id="kd_entry.125">
  <form>
    <orth>adas</orth>
  </form>
  <sense xml:id="kd_sense.339" n="1">
    <form type="variant">
      <orth>adas landi</orth>
      <orth>adas pedas</orth>
    </form>
    <def>sj tumbuhan (herba), <name xml:lang="la">Foeniculum
      vulgare</name></def>
  </sense>
  <sense xml:id="kd_sense.340" n="2">
    <form type="variant">
      <orth>adas cina</orth>
      <orth>adas manis</orth>
    </form>
    <def>sj tumbuhan (herba), <name xml:lang="la">Anethum
      graveolens</name></def>
  </sense>
</entry>
```

## SENSE ENTRIES 'FLATTENED' INTO SQL RECORDS FOR EASIER LOOKUP

SenseID	Headword	Orthographic Forms	Def
339	adas	adas; adas landi; adas pedas	sj tumbuhan (herba), _Foeniculum vulgare_
340	adas	adas; adas cina; adas manis	sj tumbuhan (herba), _Anethum graveolens_

- Princeton WordNet contains scientific names!
- *Let's match them!*
  1. Search for scientific name in Wordnet
  2. If unmatched scientific name optional infraspecies e.g. *Morus alba (indica)*, try matching without infraspecies.
- 549 KD sense entries aligned to WordNet using direct matching

## EXAMPLE ALIGNMENTS

Synset ID	B. M'sia (KD)	Scientific name	English (WN)	Definition
11824344-n	bayam duri	<i>Amaranthus spinosus</i>	thorny amaranth	erect annual of tropical central Asia and Africa having a pair of divergent spines at most leaf nodes
01822300-n	bayam lepas	<i>Psittacula krameri</i>	ring-necked parakeet	African parakeet
12180168-n	bebaru	<i>Hibiscus tiliaceus</i>	balibago; mahagua; mahoe; majagua; purau	shrubby tree widely distributed along tropical shores; yields a light tough wood used for canoe outriggers and a fiber used for cordage and caulk; often cultivated for ornament
12399384-n	bebesaran	<i>Morus alba</i>	white mulberry	Asiatic mulberry with white to pale red fruit; leaves used to feed silkworms
01543632-n	belatik	<i>Padda oryzivora</i>	Java finch; Java sparrow; ricebird	small finch-like Indonesian weaverbird that frequents rice fields

## WHEN A SCIENTIFIC NAME ISN'T FOUND IN WORDNET...

- Typo?
- Not accepted name? (WordNet records only accepted names)
- WordNet doesn't have a synset for it?
- There's no English common name (only found in this region)?

- (Checklist: a list of species names)
- By Species 2000, federation of taxonomic database custodians
- Online database of the world's known species of animals, plants, fungi and micro-organisms
- > 1.6M species (84 % coverage) from 154 databases
- Annual checklist available as MySQL  
(Monthly checklist contains *many more* updates, but unavailable for download)
- Common name is not always available!
- **No B. Malaysia nor B. Indonesia**

- If a scientific name doesn't have a match in CoL checklist:
  - Run a full text index match
  - Compute Levenshtein distances for top matches
  - Lowest distance  $\rightarrow$  top candidate
  - Distance  $\leq 2 \rightarrow$  possible typo



## DEALING WITH TYPOS (CONT.)

- Example: 'Puntuis lateristriga'

Candidates	Levenshtein distance
Puntius lateristriga	2
Barbus lateristriga	5
Awaous lateristriga	5
Boulengerella lateristriga	11
Maravichromis lateristriga	11

- 330 possible typos found re CoL

- As study and research progresses, scientific nomenclature change (sometimes changing genus!)
- Varies due to region, domain, etc.
- All synonyms still to be recorded because e.g. legislators need to refer to previous conventions, cases, etc.
- Looking up accepted names from synonyms

- 96 more synsets matched (on top of 549 direct matches earlier)
- 980 other KD sense entries have English translations from CoL
- 758 unique B. Malaysia items mapped to 405 WordNet synsets
- 647 new B. Malaysia items can be added to Wordnet Bahasa

MORE FROM WIKIDATA & WIKIMEDIA

---

- Central storage for the structured data of Wikimedia projects including Wikipedia, Wikivoyage, Wikisource, and others
- MediaWiki action API: HTTP query  
`http://www.wikidata.org/w/api.php`
- Wikidata Query Service: SPARQL API  
`https://wdq.wmflabs.org/api\_documentation.html`

## EXAMPLE VIA QUERY HTTP

Synset ID 01822300-n

English Psittacula krameri; ring-necked parakeet

B. Malaysia bayan lepas

### Look up 'Psittacula krameri'

```
https://www.wikidata.org/w/api.php?  
action=wbgetentities  
&sites=enwiki  
&titles=Psittacula krameri  
&normalize&format=xml  
&props=datatype|labels|aliases  
&languages=en|zh|ms|id|ja|fr|es
```

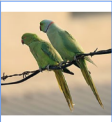
- Also available as JSON, PHP array
- Many other information available

```

<?xml version="1.0"?>
<api success="1">
  <normalized>
    <n from="Psittacula krameri" to="Rose-ringed parakeet" />
  </normalized>
  <entities>
    <entity type="item" id="Q208060">
      <labels>
        <label language="fr" value="Perruche à collier" />
        <label language="es" value="Psittacula krameri" />
        <label language="ja" value=" ワカケホンセイインコ" />
        <label language="en" value="Rose-ringed Parakeet" />
        <label language="zh" value=" 红领绿鹦鹉" />
      </labels>
      <aliases>
        <language id="fr">
          <alias language="fr" value="Perruche à collier rose" />
          <alias language="fr" value="Psittacula krameri" />
        </language>
        <language id="es">
          <alias language="es" value="Cotorra de Kramer" />
        </language>
        <language id="en">
          <alias language="en" value="Psittacula krameri" />
        </language>
      </aliases>
    </entity>
  </entities>
</api>

```

# NAMES FROM KD, PWN, WIKIDATA, COL2015

	A	B	C	D	E	F	G	H	I	J	K	L
	Thumbnail	Synset ID	Scientific Name	English	B. Malaysia	B. Indonesia	Chinese	Japanese	Korean	Thai	Arabic	French
44		01822300-n	Ptilinopus krameri	Rose-ringed Parakeet; ring-necked parakeet	bayan lepas		紅領綠鸚鵡	ワカケホンセイインコ			بالرأس الأخضر: ببس مطرد: البازكيت الأخضر: ببس مطرد Indian Ringneck بالرأس الأخضر: Parrots	Perruche à collier; Perruche à collier rose
45		12399384-n	Morus alba	white mulberry	bebesaran; besaran; kertau; mulberi		白桑; 白桑椹		홍나무; 오디; 홍나무꽃		توت أبيض	Mûrier blanc; Murier blanc
46		01543632-n	Padda oryzivora	Java Sparrow; Java finch; Java rice sparrow; Java rice bird; ricebird	Burung Cak Jawa; belatik	Gelatik Jawa	禾雀; 爪哇禾雀; 爪哇雀; 灰文鳥; 灰文鳥; 文鳥; 文鳥	ブンチョウ; 手乗りブンチョウ; 手乗り文鳥; 文鳥	문조	นกกระจกนา; Java finch; Java sparrow; นกนา	جرار سارو	Padda de Java; Moineau de Java; Calfat de Java; Padda oryzivore
47		00000000-n	Averrhoa	Carambola;	Belimbing besi; Pokok Belimbing Besi; belimbing batu; belimbing			スターフルーツ; 五郎子; カラ	카람 불라; 스타 프루트; 스타 후르츠			

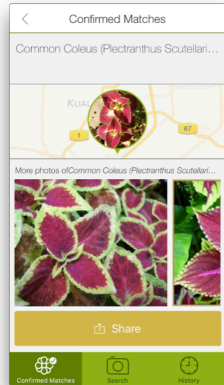
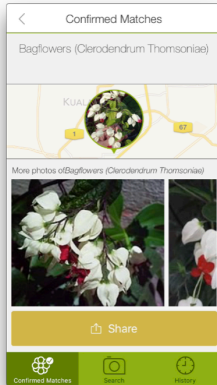
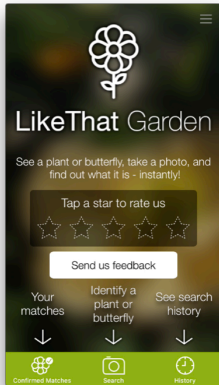
+ Sheet1



- Extract/generate at will based on the topic/area/theme



# GARDEN: IOS/ANDROID APP TO RECOGNISE PLANTS FROM CAMERA







# MODELING REGIONAL PREFERENCES IN WORDNET?

---

- *Euthynnus alleteratus* (most common tuna) = ikan aya, tongkol, kayu
  - 'Aya' commonly used in the East Coast of West Malaysia (Kelantan & Terengganu)
  - Rest of Malaysia use 'tongkol' or 'ikan kayu'
- Also recall 'elk' example earlier
- Also applies to regionally-preferred words e.g. petrol vs gas



## REFERENCES

-  Hajah Noresah, b. B. (Ed.). (2004). *Kamus Dewan*. Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka.
-  Hussey, C., Wilkinson, S. & Tweddle, J. (2006). Delivering a name-server for biodiversity information. *Data Science Journal*, 5, 18–28.
-  Roskov, Y., Abucay, L., Orrell, T., Nicolson, D., Kunze, T., Culham, A., ... De Wever, A. (Eds.). (2015). Species 2000 & itis catalogue of life, 2015 annual checklist, *Leiden, the Netherlands*. Retrieved from <http://www.catalogueoflife.org/annual-checklist/2015>
-  TEI P5: Guidelines for Electronic Text Encoding and Interchange. (2015). In TEI Consortium (Ed.), (Chap. Dictionaries). Retrieved November 25, 2014, from <http://www.tei-c.org/release/doc/tei-p5-doc/html/DS.html#DSFLT>