

A DAY (OR TWO) WITH LEXICOGRAPHERS

COLLABORATING WITH DBP ON WORDNET CHECKING

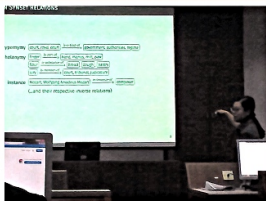
Dr LIM Lian Tze (liantze@gmail.com)

The Second Wordnet Bahasa Workshop

Malaysia

- Institute of Language and Literature
- Established as Balai Pustaka in 1956
- Autonomous powers to:
 - formulate specific policies (regarding use of B. Malaysia)
 - organise language and literature construction and development programmes
 - undertake the publishing and sale of books

WORDNET WORKSHOP AT DEWAN BAHASA & PUSTAKA (DBP)



Calangan Entri Baru untuk Wordnet Bahasa (30 M)

* Required

Rekod Syaset Princeton WordNet

Syaset ID *

Salin syaset ID dari Open Multilingual Wordnet

12345678-n

Lemma Bahasa Inggeris dari Princeton WordNet untuk rujukan

English lemma from Princeton WordNet -- for reference only

english lemma

Rekod Syaset Wordnet Bahasa

Sila edit bahagian ini sekiranya terdapat rekod yang kurang memuaskan

Hubungan dengan syaset Princeton WordNet

Relation to Princeton WordNet syaset

setara dengan (equivalent to)

Bidang

Domain

Lemma dalam Bahasa Malaysia (xsm) *

Sekiranya lebih daripada satu lemma, pisahkan lemma dengan tanda koma (,)

Malay lemma


- 2-hour seminar: Introduction to Wordnets
- 2-hour hands-on: Editing/verifying Wordnet Bahasa contents

WE WANTED THE PROCESS TO BE

- Easy to set up
- Easy to collect
- Easy for validators
- (Hopefully) not too confusing

- Easy to design and set up
- Responses saved to a spreadsheet on Google Docs
- Can pre-fill fields:
`https://docs.google.com/forms/<formID>/viewform?entry.<entry1ID>=Answer_1&entry.<entry2ID>=Answer_2&entry.<entry3ID>=Answer_3`

PRE-POPULATING FROM WORDNET BAHASA

03001627-n  'a seat for one person, with a support for the back';

Search WN

English

Malaysian

English

chair₃₅

Indonesian

kursi , bangku

Malaysian

kerusi , kursi , bangku

[Edit Synset \(Experimental\)](#)

Definitions

English

a seat for one person, with a support for the back — *he put his coat over the back of the chair and sat down*

Relations

Hyponym:

[armchair](#) [barber_chair](#) [chair_of_state](#) [chaise_longue](#) [eames_chair](#) [fighting_chair](#)
[folding_chair](#) [highchair](#) [ladder-back](#) [lawn_chair](#) [rocking_chair](#) [straight_chair](#) [swivel_chair](#)
[tablet-armed_chair](#) [wheelchair](#)

Hypernym:

[seat](#)

Meronym-Part: [back](#) [leg](#)

Semantic Field: artifact_n

External Links

SUMO: = [Chair](#)

TempoWN: ◀ ◻ ▶ (Past: 0.000; Present: 0.000; Future: 0.022)

SentiWN: ◀ ◻ (+0.00 -0.00) ML SentiCon: ▲ ▼ (+0.25 -0.00)

Thanks to Francis, Luis and David



Cadangan Entri Baru untuk Wordnet Bahasa (30 Mac 2015)

* Required

Rekod Synset Princeton WordNet

Synset ID: *

Salin synset ID dari Open Multilingual Wordnet



Lemma Bahasa Inggeris dari Princeton WordNet untuk rujukan

English lemma from Princeton WordNet -- for reference only

B. MALAYSIA MEMBERS

Rekod Synset Wordnet Bahasa

Sila edit bahagian ini sekiranya terdapat rekod yang kurang memuaskan

Hubungan dengan synset Princeton WordNet

Relation to Princeton WordNet synset

setara dengan (equivalent to)



Bidang

Domain

Lemma dalam Bahasa Malaysia (zsm) *

Sekiranya lebih daripada satu lemma, pisahkan lemma dengan tanda koma (,)

kerusi, kursi, bangku

Definisi dalam Bahasa Malaysia

Definition text in Malay

OPTIONAL: DEFINITION AND COMMENTS

Definition text in Malay

Komen untuk lemma Bahasa Malaysia

Comments. Apa-apa komen yang ingin diisi tentang cadangan lemma ini

Lemma dalam Bahasa Indonesia (ind)

Sekiranya lebih daripada satu lemma, pisahkan lemma dengan tanda koma (,)

kursi, bangku

WHICH SYNSETS TO CHECK?

- Most frequent concepts/synsets with no B. Malaysia equivalents (based on counts from `ntu-mc` and `semcor` combined)
- Most frequent concepts/synsets (almost certainly containing bad B. Malaysia entries)
- Adding entries from DBP's own resources to synsets
- Concentrating on "groups of words" e.g. colours; dishes; artistic forms...
- Checking on "false friends" e.g. http://en.wikipedia.org/wiki/Comparison_of_Malaysian_and_Indonesian#False_friends

WHICH SYNSETS TO CHECK?

- Most frequent concepts/synsets with no B. Malaysia equivalents (based on counts from `ntu-mc` and `semcor` combined)
- Most frequent concepts/synsets (almost certainly containing bad B. Malaysia entries)
- Adding entries from DBP's own resources to synsets
- Concentrating on "groups of words" e.g. colours; dishes; artistic forms...
- Checking on "false friends" e.g. http://en.wikipedia.org/wiki/Comparison_of_Malaysian_and_Indonesian#False_friends

- \simeq 20 staff from Bahagian Perancangan Bahasa & Perkamusan
- 138 English–B. Malaysia synsets checked and edited in 2 hours
- Great job!! 👍

COLLECTED RESPONSES

fx meminta utk berkahwin; meminang										
	B	C	D	E	F	G	H	I	J	K
1	Synset ID:	Lemma dalam Bahasa Malaysia (zsm)	Hubungan dengan synset Princeton WordNet	Tindakan (Action proposed)	Definisi dalam Bahasa Malaysia	Komen untuk lemma Bahasa Indonesia	Bidang	Lemma Bahasa Inggeris dari Princeton WordNet untuk rujukan	Lemma dalam Bahasa Indonesia (ind)	Definisi dalam Bahasa Indonesia
17	02186338-a	salu			angka pertama ketika membilang			one, 1, l, ane	salah, satu	
18	14331873-n	kepedihan, keperlan, rasa pedih, bergaya						smartness, smart, smarting	bergaya, smart, rasa pedih, kepedihan	sejenis nyeri yang disebabkan oleh lul bakar/sakit
19	02534761-v	melamar			meminta utk berkahwin; meminang			court	melamar, memacar	
20	01391351-a	ketot, katik, kecil,						small, little	katik, kecil, ketot,	
21	00947719-n	salah guna dadah						misuse, abuse		
22	00115777-a	marah, berang						irate, ireful	marah, berang, menunjukkan kemarahan ekatrim	
23	01078783-v	berhadapan, menentang, bersemuka dgn, mendatangi, beragah, menghadap, bertentangan, menghadapi, mempertemukan						face, confront	beragah, mengarahkan, mempertemukan, menghadap, menentang, berhadapan, bersemuka dgn, bertentangan, berhadapan muka, menghadap, mendatangi, nebagak	
24	02138766-v	membina, membangunkan, memajukan						develop	memajukan, membina, jadi, kembang, berkembang maju dgn pesat, meluaskan, membangun dgn pesat, membangunkan, mencuci	berkembang maju pesat
					sangat menarik apabila			great, really, good, fast		



Form Responses 1 -



CAUSES OF ERROR/CONFUSION

- Getting used to wordnet

CAUSES OF ERROR/CONFUSION

- Getting used to wordnet
- Getting used to the interface

CAUSES OF ERROR/CONFUSION

- Getting used to wordnet
- Getting used to the interface
- Distracted by the English or (existing) B. Malaysia members

CAUSES OF ERROR/CONFUSION

- Getting used to wordnet
- Getting used to the interface
- Distracted by the English or (existing) B. Malaysia members
- (English) synset definitions is important

CAUSES OF ERROR/CONFUSION

- Getting used to wordnet
- Getting used to the interface
- Distracted by the English or (existing) B. Malaysia members
- (English) synset definitions is important
- Lack of B. Malaysia definition

CAUSES OF ERROR/CONFUSION

- Getting used to wordnet
- Getting used to the interface
- Distracted by the English or (existing) B. Malaysia members
- (English) synset definitions is important
- Lack of B. Malaysia definition
- Frequently used words → highly polysemous, ambiguous

CAUSES OF ERROR/CONFUSION

- Getting used to wordnet
- Getting used to the interface
- Distracted by the English or (existing) B. Malaysia members
- (English) synset definitions is important
- Lack of B. Malaysia definition
- Frequently used words → highly polysemous, ambiguous
- English Wordnet sense granularity is very (too?) fine
(*re* Christiane's discussion on alternatives)

CAUSES OF ERROR/CONFUSION

- Getting used to wordnet
- Getting used to the interface
- Distracted by the English or (existing) B. Malaysia members
- (English) synset definitions is important
- Lack of B. Malaysia definition
- Frequently used words → highly polysemous, ambiguous
- English Wordnet sense granularity is very (too?) fine
(*re* Christiane's discussion on alternatives)
- Underspecified

CAUSES OF ERROR/CONFUSION

- Getting used to wordnet
- Getting used to the interface
- Distracted by the English or (existing) B. Malaysia members
- (English) synset definitions is important
- Lack of B. Malaysia definition
- Frequently used words → highly polysemous, ambiguous
- English Wordnet sense granularity is very (too?) fine
(*re* Christiane's discussion on alternatives)
- Underspecified
 - Different ways to 'cut', 'carry'

CAUSES OF ERROR/CONFUSION

- Getting used to wordnet
- Getting used to the interface
- Distracted by the English or (existing) B. Malaysia members
- (English) synset definitions is important
- Lack of B. Malaysia definition
- Frequently used words → highly polysemous, ambiguous
- English Wordnet sense granularity is very (too?) fine
(*re* Christiane's discussion on alternatives)
- Underspecified
 - Different ways to 'cut', 'carry'
- **Better guidelines needed**

- Wordnet Bahasa marks lemma–synset mappings which are only used in
 - Malaysia (*zsm*)
 - Indonesia (*ind*)
 - both (*msa*)
- Also useful to check for ‘false friends’
- **Brunei (*kxd*)??**
e.g. ‘tambak’ means ‘kitchen’ only in Bahasa Melayu Brunei (Nasruddin Abdullah, 1988)
- From Ethnologue:
 - *zsm, ind*: > 80 % lexical similarity
 - *zsm, kxd*: 80–82 % lexical similarity


POSSIBLE COLLABORATION ON WORDNET BAHASA?

- Data
 - Checking entries (focus on needs/areas)
 - Writing definitions
 - B. Malaysia/B. Indonesian/B. Brunei distinctions
- Checking/Verification (semi-automatic)

OTHER POSSIBLE COLLABORATIONS

- Resources or expertise for verifying/adding synset members
 - Most common words/concepts
 - 'Groups of friends'
 - Domain-specific terminologies (train experts?)
- Definitions: data/resources, guidelines
 - for human consumption (readers' understanding)?
 - for machine consumption (easier NLP parsing/analysis)?
- Culture-specific synsets ('expand' approach) e.g. clothing apparel items, food, etc
 - Tie-in with other projects
 - May be easier to attract interest
- Others!



-  Nasruddin Abdullah. (1988). *Kamus kata dan ungkapan am : Bahasa Indonesia, Bahasa Malaysia, Bahasa Melayu Brunei*. Kuala Lumpur: Dewan Bahasa dan Pustaka.