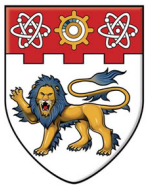
Extending Wordnet:
The Never-ending Story…

- **Motivation (NTUMC)**
- **Our Previous Efforts**
- **New Extensions**
  - Classifiers
  - Interjections
- **Future Work**



NANYANG TECHNOLOGICAL UNIVERSITY

NTU-Multilingual Corpus (NTUMC)

- **Parallel Corpus of English + Asian Languages (CMN, JPN, IND)**
  (Multi-genre, Multilingual, +Italian, +The Spider's Thread, etc.)

- **Full Suite of Annotation Tools (OMWEdit, IMI)**
  (Revamped, and now taking the first steps into Sentiment Analysis)

- **Public Corpus Browser** (http://compling.hss.ntu.edu.sg/ntumc/cgi-bin/showcorpus.cgi)
  (Asynchronous)

- **Gold:** Tokenisation; POS tagging; MWE; Sense annotation; cross-lingual sense alignments; sense and sentence level sentiment analysis; structural semantics; etc.

Previously, on Extending WN…

- **Chengyu (Chinese Classical Idioms)**

  207 Chengyu Concepts (Still ongoing work…)

- **Pronouns, Determiners & Quantifiers (come back to them later)**

  140 Types, 333 Tokens in 4 languages — marked for 40+ features

  (e.g. usage, person, number, gender, politeness, etc.)

- **∗ Classifiers ∗**

  Top-down propagation

  External resource to mark concepts

  Low coverage (human work)

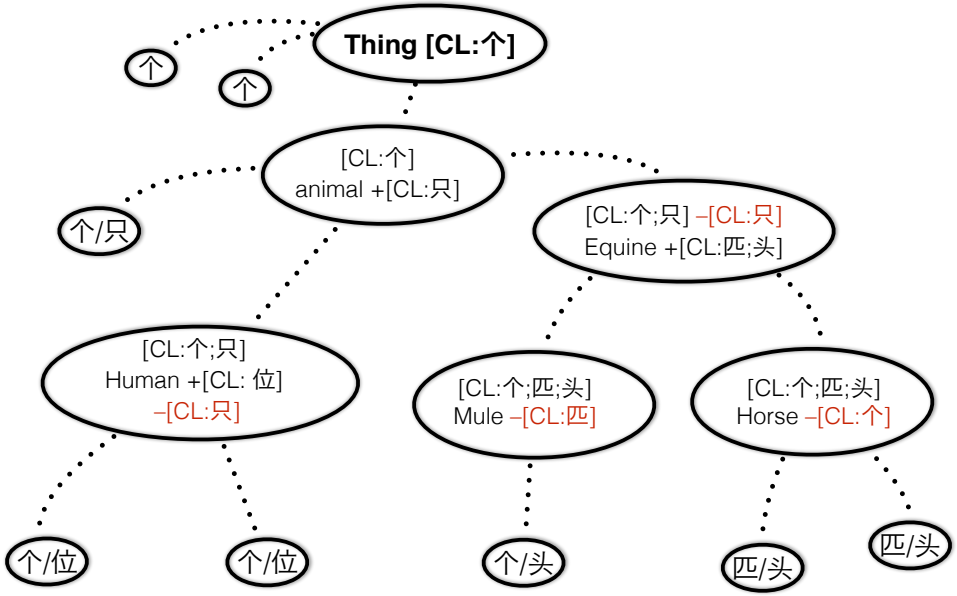


Figure 1: Classifier Propagation.

NANYANG TECHNOLOGICAL UNIVERSITY

Classifiers as Wordnet Concepts

sān zhī láo shǔ

三 只 老 鼠

(three mice)

(a slice of cake)

yí piàn dàn gāo

一 片 蛋 糕

Classifiers (I)

- **There are many types of classifiers:**

sortal (which classify the kind of the noun phrase they quantify);

event (which are used to quantify events);

mensural (which are used to measure the amount of some property);

group (which refer to a collection of members);

taxonomic (which force the noun phrase to be interpreted as a generic kind)

- **Most languages make use of some/different types of classifiers**

a) a kilo of coffee (mensural classifier)
satu kilo kopi
sekilo kopi (?)

b) a school of fish (group classifiers)
satu kawanan ikan
sekawanan ikan (?)

Classifiers (II)

- **Sortal Classifiers (S-CLs) are interesting for many reasons:**
  - S-CLs don't currently have an adequate representation in Wordnet (mensural or group classifiers are considered nouns)
  - S-CLs usage is licensed by a number of semantic features (e.g. physical, functional, etc.)

(1) 两 只 狗
liǎng zhǐ gǒu
2 CL dog
"two dogs"

(2) 两 条 狗
liǎng tiáo gǒu
2 CL dog
"two dogs"

(3) 两 条 路
liǎng tiáo lù
2 CL road
"two roads"

(4) 三 台 电脑
sān tái diànnǎo
3 CL computer
"three computers"

(5) *三 只 电脑
sān zhǐ diànnǎo
3 CL computer
"three computers"

Classifiers (III)

- **Classifiers comprise about 2.5% of our CMN corpus**

  (we expect similar numbers for Japanese, and slightly lower for Indonesian)

- **Classifiers have some kind of semantics!**

  - 一 个 木头 (general classifier)

    yī ge mùtou

    1 CL log (of wood) / blockhead

    "a log / blockhead"

  - 一 位 木头 (human, formal classifier)

    yī wèi mùtou

    1 CL blockhead

    "a blockhead"

  - 一 根 木头 (long, slender objects classifier)

    yī gēn mùtou

    1 CL log (of wood)

    "a log"

Classifiers (IV)

- **CLs in Wordnet**
  - 'x' as part-of-speech
  - definition with the form *"a … classifier used ..., such as ..."*
  - domain usage: **classifier** (06308436-n)

- **87 Chinese S-CLs in COW**
- **30 Indonesian S-CLs in WN Bahasa**

80000003-x	
lemmas	把 (bǎ)
definition	a sortal classifier used with tools and objects with a handle, such as a hammer, a broom, a guitar or a teapot
domain usage	06308436-n (classifier)

80000004-x	
lemmas	根 (gēn)
definition	a sortal classifier used for long slender objects, such as a banana, a pillar, a sausage or a needle
domain usage	06308436-n (classifier)

30 Indonesian S-Classifiers

batang	a sortal classifier used with long objects, such as a pencil, a tree or a river
batu	a sortal classifier used with teeth
bengkawan	a sortal classifier used with roofs
bentuk	a sortal classifier used with curvy objects, such as a ring, a bracelet or a key
bidang	a sortal classifier used with wide objects, such as a rice field, a land or a farm
biji	a sortal classifier used with small objects, such as an eyeball, a mango or a seed
bilah	a sortal classifier used with long and thin objects, such as a sword, a board or a machete
buah	a sortal classifier used with inanimate objects, such as a ship, a country or a plan
bulir	a sortal classifier used with ears, such as a paddy or a barley
butir	a sortal classifier used with round and small objects, such as a bullet, an egg or a pearl
carik	a sortal classifier used with wide and thin paper objects, such as a paper, a letter or a musical score
ekor	a sortal classifier used with animals, such as a tiger, a fish or a bird
eksemplar	a sortal classifier used with printed materials, such as a book, a magazine or a newspaper
helai	a sortal classifier used with thin or soft objects, such as a paper, a cloth, a hair or a thread
kaki	a sortal classifier used with umbrellas

30 Indonesian S-Classifiers

keping	a sortal classifier used with flat objects, such as a board or a coin
kuntum	a sortal classifier used with flowers, such as a rose, a hibiscus or a jasmine
labuh	a sortal classifier used with curtains
laras	a sortal classifier used with rifles
lembar	a sortal classifier used with wide and thin objects, such as a board, a paper or a cloth
lonjor	a sortal classifier used with long objects, such as a bamboo, a pipe or a log
orang	a sortal classifier used with human being, such as a man, a child or a farmer
patah	a sortal classifier used with words, such as a word
pintu	a sortal classifier used with apartments or rooms for dwelling
pucuk	a sortal classifier used with tools and objects with a pointed end, such as a needle, a letter or a rifle
siung	a sortal classifier used with onions and garlics
unit	a sortal classifier used with objects complete with their parts, such as a vehicle, a computer or a house
untai	a sortal classifier used with chains or strings, such as a necklace, a firecracker or a bracelet
urat	a sortal classifier used with rattans
utas	a sortal classifier used with threadlike objects, such as a thread, a rope or a wire

Interjections as Wordnet Concepts

Interjections (I)

- **What are Interjections?** (They are hard to define!)
  - words or phrases
  - constitute a whole linguistic act

    (do not combine in integrated syntactic constructions)
  - do not refer to events / do not have referents, but instead carry expressive meaning

    (Huddleston and Pullum, 2002)

- **We follow Jovanovíc (2004) and Ameka (1999), and use the term broadly, covering plain interjections, greetings and many more…**

Interjections (II)

- **Interjections are quite frequent in our corpus!**

(This is expected to be true in any corpus that contains direct speech)

a. "Ah! That is suggestive. Now, on the other side of this narrow wing runs the corridor from which these three rooms open. There are windows in it, of course?"

b. "Yes, but very small ones. Too narrow for anyone to pass through."

c. "Thank you. That is quite settled" said he, rising and putting his lens in his pocket.

d. "Hullo! Here is something interesting"

The Adventure of the Speckled Band (Conan Doyle, 1892)

Interjections (III)

synset	lemmas	definition
00049758-r	**now**	indicates a change of subject or activity
15119919-n	**now**	the momentary present
00049220-r	**now**, at present	at the present moment
00049102-r	**now**	used to preface a command or reproof or request
00048475-r	**now**, today, nowadays	in these times
00048739-r	immediately, at once, right away, **now**, (. . .)	without delay or hesitation; with no time intervening
00049685-r	**now**	in the immediate past
00049433-r	**now**	in the historical present; at this point in the narration of a series of past events
07203900-n	**yes**	an affirmative
07229245-n	**thank you**	a conversational expression of gratitude
06632511-n	hello, hi, **hullo**, howdy, how-do-you-do	an expression of greeting

- **The available senses to tag these words are, arguably, inadequate!**

E.g. (vs) *I was hoping for a yes.* *Few job candidates send thank yous.*

Interjections (IV)

- **There are at least two interjections in PWN!**

synset	lemmas	definition
00150351-r	right, right on	an interjection expressing agreement
00049889-r	now now	interjection of rebuke

- **And some places where the lexicographer's intentions are unclear!**

synset	lemmas	definition
06632671-n	morning, good morning	a conventional expression of greeting or farewell, used to wish someone a good morning
06632807-n	afternoon, good afternoon	a conventional expression of greeting or farewell, used to wish someone a good afternoon
06632947-n	good night	a conventional expression of farewell

All three hyponyms of the concept 06629392-n, defined as acknowledgment or expression of goodwill at parting.

Her good-mornings/good-afternoons/good-nights were hasty and mumbled. **(?!)**

Interjections (V)

- **What does our broad sense of interjections include?**
  - expressions of emotion, such as surprise, disgust, etc.
    (e.g. *wow*, *ugh*, *yuk*, *gosh*, …)
  - expressions used in greetings, leave-taking, thanking, apologizing, etc.
    (e.g. *hello, thank you*, *goodbye*, …)
  - expressions used for swearing
    (e.g. *damn, shit, bite me*, …)
  - expressions used in responding
    (e.g. *yes, no, OK, yeah, you bet*, …)
  - and a long range of onomatopoeias
    (e.g. *hush, boo, meow, oink,* …)

Interjections (VI)

- **Interjections in Wordnet**
  - 21 Classes of Expressive Interjections (Jovanovic, 2004) merged into 12 Classes
  - Defined other major classes, enforcing only near synonymy between senses
  - ‘x’ as part-of-speech
  - definition with the form “*an expression that is uttered …*”
  - domain usage: utterance (07109847-n)
  - enrich this flatter hierarchy with links to other existing concepts (when possible)

Interjections (VII)

80000001-x (general greeting)	
lemmas	aloha, ciao, g'day, good day, hallo, halloa, halloo, hallow, hello, hi, howdy, hullo, 'sup
definition	an expression that is uttered as a general greeting, regardless of the time of day
domain usage	15167474-n (utterance)
see also	06630017-n (greeting)

80000002-x (checkmate)	
lemmas	checkmate, mate
definition	an expression that is uttered during a game of chess to declare that the final winning move has taken place
domain usage	15167474-n (utterance)
see also	00167764-n (checkmate)

Interjections (VIII)

Concept	Senses
Surprise, Wonder	58
Pity, Sorrow	19
Joy, Pleasure	17
Anger, Annoyance, Irritation	41
Approval, Triumph, Enthusiasm	10
Contempt, Disgust, Impatience	59
Pain	7
Sympathy	2
Delight	11
Fear	3
Relief	2
Encouragement	16
Attention-Seeking	36
Toasting	10
General Greetings	13
Morning Greetings	2
Afternoon Greetings	2
Night Greetings	2
General Farewells	21
Night Farewells	5
Checkmate	2
Number of senses	336

There is a big long tail of specific synsets…
(e.g. **checkmate** or **tally-ho**)

And the same for onomatopoeias …

Our current data is mainly for English

Chinese, Japanese and Bahasa soon!

Multilingual sources:

omniglot.com

Wiktionary.

Next time you're surprised…

ah!
alack!
blimey!
boy!
caramba!
coo!
cor!
crazy!
dear!
dear me!
deuce!
doggone!
gad!
gee!
gee-whiz!
golly!
good!
goodness!
gosh!
gracious!
ha!
heck!
heigh!
heigh-ho!
hey!
heyday!
ho!
hollo!
hoo-ha!
huh!
humph!
indeed!
jiminy!
lord!
man!
mercy!
my!
nu!
od!
oh!
oh no!
oho!
phew!
say!
shit!
so!
son of a bitch!
the devil!
upon my soul!
upon my word!
well!
what!
whoof!
whoosh!
why!
wow!
yow!
zounds!

Future Work / Discussion

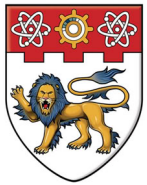
- **Future Work** (pondering)
  - Titles
  - Modals
  - Prepositions
  - Conjunctions

- **Sharing resources**
  - Do you have/know of resources to could help us add interjections, titles, prepositions or conjunctions to Wordnet Bahasa?

- **Wish-lists / Reticences ?**
  - What other classes of concepts you like to see added to WN Bahasa? (or Wordnets in general?)
  - Do any of these expansions worry you in any way?

Extending Wordnet:
The Never-ending Story…

Thank You!



NANYANG
TECHNOLOGICAL
UNIVERSITY