Verbs
in the
Open Multilingual Wordnet

Francis Bond
Linguistics and Multilingual Studies,
Nanyang Technological University

Affectedness Workshop 2014, NTU

Overview

- What do we do?
- What is a wordnet?
  - How are verbs represented?
- What is the Open Multilingual Wordnet?
  and the NTU Multilingual Corpus
- How should affectedness be represented?

NANYANG TECHNOLOGICAL UNIVERSITY

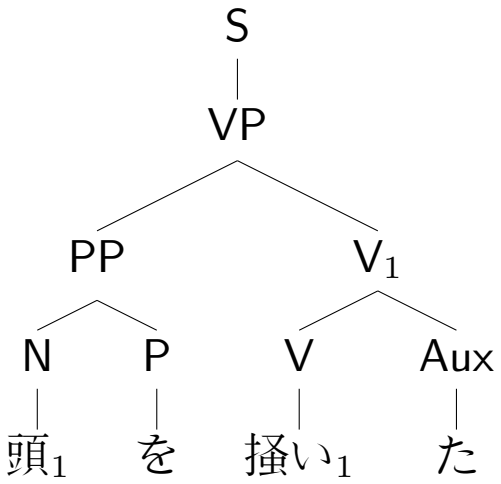
Our Vision

- We want to understand language
- We want computers to understand language:
  assign an interpretation to an utterance
  - model words as concepts (predicates)
  - link predicates together (structural semantics)
  - link predicates to the world (lexical semantics)
  - for any language
- Our approach is incremental
  - model what we can: so that we can produce descriptions
  - improve the model: more coverage/richer description
  - repeat

NANYANG TECHNOLOGICAL UNIVERSITY

Rich Representation

(1) 頭　　を　掻いた
atama wo kaita
head ACC scratched
"I scratched my head."



Syntax

atama$_1$(y)	is-a	bodypart
kaku$_1$(e,x,y)	is-a	change
kaku	ARG1	zero-pronoun (?speaker)
kaku	ARG2	atama
kaku	TENSE	past

Semantics

NANYANG TECHNOLOGICAL UNIVERSITY

Why multiple languages?

- to be able to make knowledge available in any language
  - machine translation
  - cross-lingual information retrieval
- to exploit translations to bootstrap learning
  - translation sets can pinpoint concepts
  - translations can disambiguate structure
  - different languages pick out different things
- aim for a uniform semantic representation
  - roughly the same across languages
  - roughly the same level of detail for all phenomena

The Core Problem of MT (& NLU)

(2) 頭　　を　掻いた
atama wo kaita
head ACC scratched
"I scratched my head."

➢ The Japanese text doesn't say

1. That 掻く should be *scratch*, not *shovel*, *row*, ...
2. Who scratched
3. That 頭 should be *head*, not *boss*, *top*, ...
4. That *head* needs a possessive pronoun
5. Whose head it is

➢ A native speaker of Japanese would know (2,5), could deduce (1,3)

➢ A native speaker of English knows (4)

? How can we learn these things?

NANYANG TECHNOLOGICAL UNIVERSITY

Languages Mark Things Differently

- E.g., most languages care about possession
  - English: pronouns
    my head
  - Japanese: politeness, evidentiality
    your honorable head vs *my head*
    I itch vs *you seem to itch*
  - Russian: reflexives
    I scratch self head
  - Swedish: definiteness
    I scratch the head (head-et)
  - German: Ich habe mich am Kopf gekratzt.
    I have me at+the head scratched

NANYANG TECHNOLOGICAL UNIVERSITY

But translation is AI-complete

> Translation, you know, is not a matter of substituting words in one language for words in another language. Translation is a matter of saying in one language, for a particular situation, what a native speaker of the other language would say in the *same* situation. The more unlikely that situation is in one of the languages, the harder it is to find a corresponding utterance in the other.
>
> Suzette Haden Elgin
> *Earthsong: Native Tongue II* (1994: 9)

Wordnets

WordNet

- Princeton WordNet (PWN) is an open-source electronic lexical database of English, developed at Princeton University

  `http://wordnet.princeton.edu/`

- Made up of four linked semantic nets, for each of nouns, verbs, adjectives and adverbs

- Wordnets exist for many, many languages

- None are as mature as PWN

Psycholinguistic Foundations

- Strong foundation on hypo/hypernymy (lexical inheritance) based on
  - response times to sentences such as:

    a canary {can sing/fly,has skin}

    a bird {can sing/fly,has skin}

    an animal {can sing/fly,has skin}
  - analysis of anaphora:

    I gave Kim a novel but the {book,?product,...} bored her

    Kim got a new car. It has shiny {wheels,?wheel nuts,...}
  - selectional restrictions

Major Relations (WordNet)

hypernyms: Y is a hypernym of X if every X is a (kind of) Y

instances: X is an instance of Y if X is a member of Y

holonym: Y is a holonym of X if X is a part of Y

troponym: the verb Y is a troponym of the verb X if the activity Y is doing X in some manner ($lisp$ to $talk$)

entailment: the verb Y is entailed by X if by doing X you must be doing Y ($sleeping$ by $snoring$)

antonymy (hot vs $cold$)

related nouns (hot vs $heat$)

Verb Relations (WordNet)

hypernym the verb Y is a hypernym of the verb X if the activity X is a (kind of) Y (travel to movement)

troponym the verb Y is a troponym of the verb X if the activity Y is doing X in some manner ($lisp$ to $talk$)

entailment the verb Y is entailed by X if by doing X you must be doing Y ($sleeping$ entails $snoring$)

cause the verb Y causes X if by doing X Y is caused (A $heats$ B causes B $heats$ up)

derivation ($\texttt{driver}_{n:1}$ to $drive$v2)

Sentence Frames

1 Something ----s
2 Somebody ----s
3 It is ----ing
4 Something is ----ing PP
5 Something ----s something Adjective/Noun
6 Something ----s Adjective/Noun
7 Somebody ----s Adjective
8 Somebody ----s something
9 Somebody ----s somebody
10 Something ----s somebody
11 Something ----s something
12 Something ----s to somebody

```
13    Somebody ----s on something
14    Somebody ----s somebody something
15    Somebody ----s something to somebody
16    Somebody ----s something from somebody
17    Somebody ----s somebody with something
18    Somebody ----s somebody of something
19    Somebody ----s something on somebody
20    Somebody ----s somebody PP
21    Somebody ----s something PP
22    Somebody ----s PP
23    Somebody's (body part) ----s
24    Somebody ----s somebody to INFINITIVE
```

```
25    Somebody ----s somebody INFINITIVE
26    Somebody ----s that CLAUSE
27    Somebody ----s to somebody
28    Somebody ----s to INFINITIVE
29    Somebody ----s whether INFINITIVE
30    Somebody ----s somebody into V-ing something
31    Somebody ----s something with something
32    Somebody ----s INFINITIVE
33    Somebody ----s VERB-ing
34    It ----s that CLAUSE
35    Something ----s INFINITIVE
```

Very English specific — not done for other languages

NANYANG TECHNOLOGICAL UNIVERSITY

Many Enhancements

- Corpus annotation and sense frequency
- Links to pictures, geo-coordinates, sentiments, temporal . . .
- Synset names
- Glosses (disambiguated)
- Many similarity measures
  - path based
  - information based
- Many software tools

Wordnets in Translation

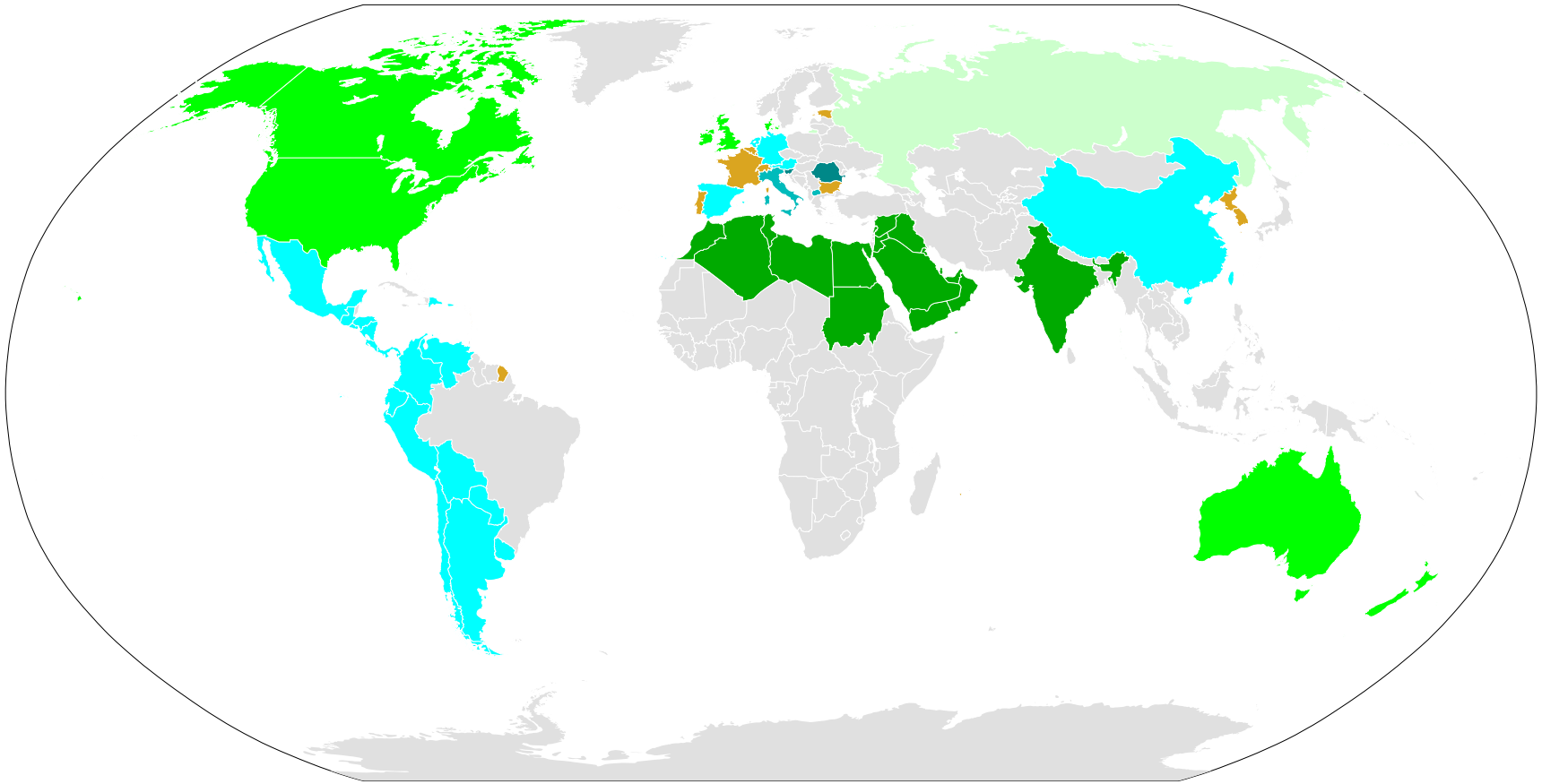
- A wide variety of new wordnets built (over 25 released)
- Typically by translating PWN
  - most have less cover
  - typically have few non-English synsets
    - Exceptions: Chinese, Korean, Arabic, Dutch, Polish Japanese, Malay
  - We are trying to fix this with the ILI
    - Add synsets (concepts) not lexicalized in English
    - Add or remove relations for different languages
    - prototype by early August with Piek Vossen (VU)

Toward a Multilingual Wordnet

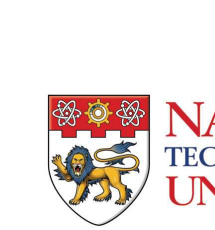
- Needed to link different language's wordnets to exploit the cross-lingual discriminating power:
  - *table*: テーブル $\subset$ $\texttt{furniture}_{n:1}$
  - *table*: 表 $\subset$ $\texttt{diagram}_{n:1}$
- Turned out to be un-necessarily time-consuming
  - Many idiosyncrasies in formats
  - Licensing often left unclear
- We want to save other people this pain
  - So that we can move onto the interesting problems

NANYANG TECHNOLOGICAL UNIVERSITY

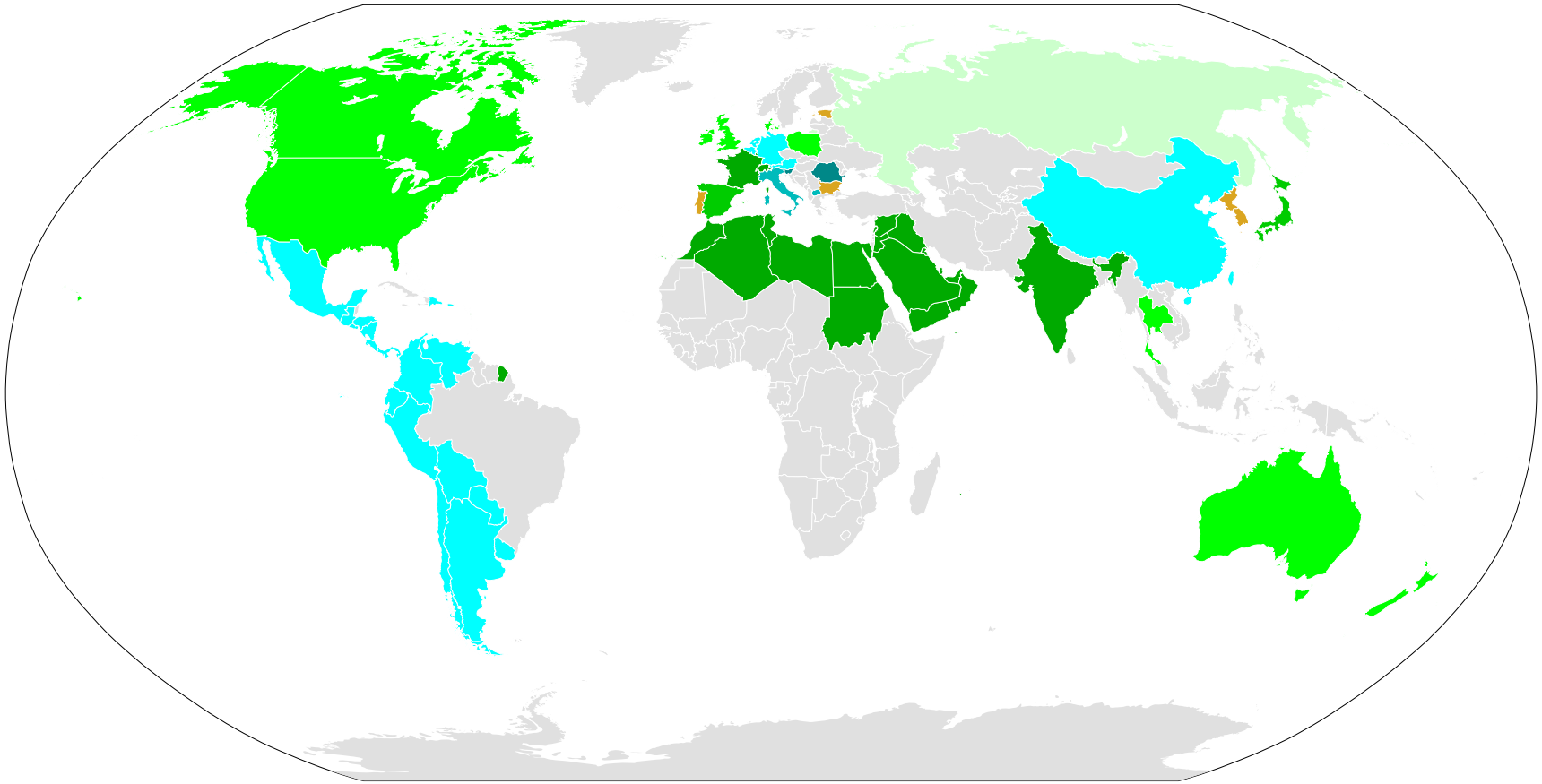
Wordnets in the world 2008



Green is free; Blue is research only; Brown costs money



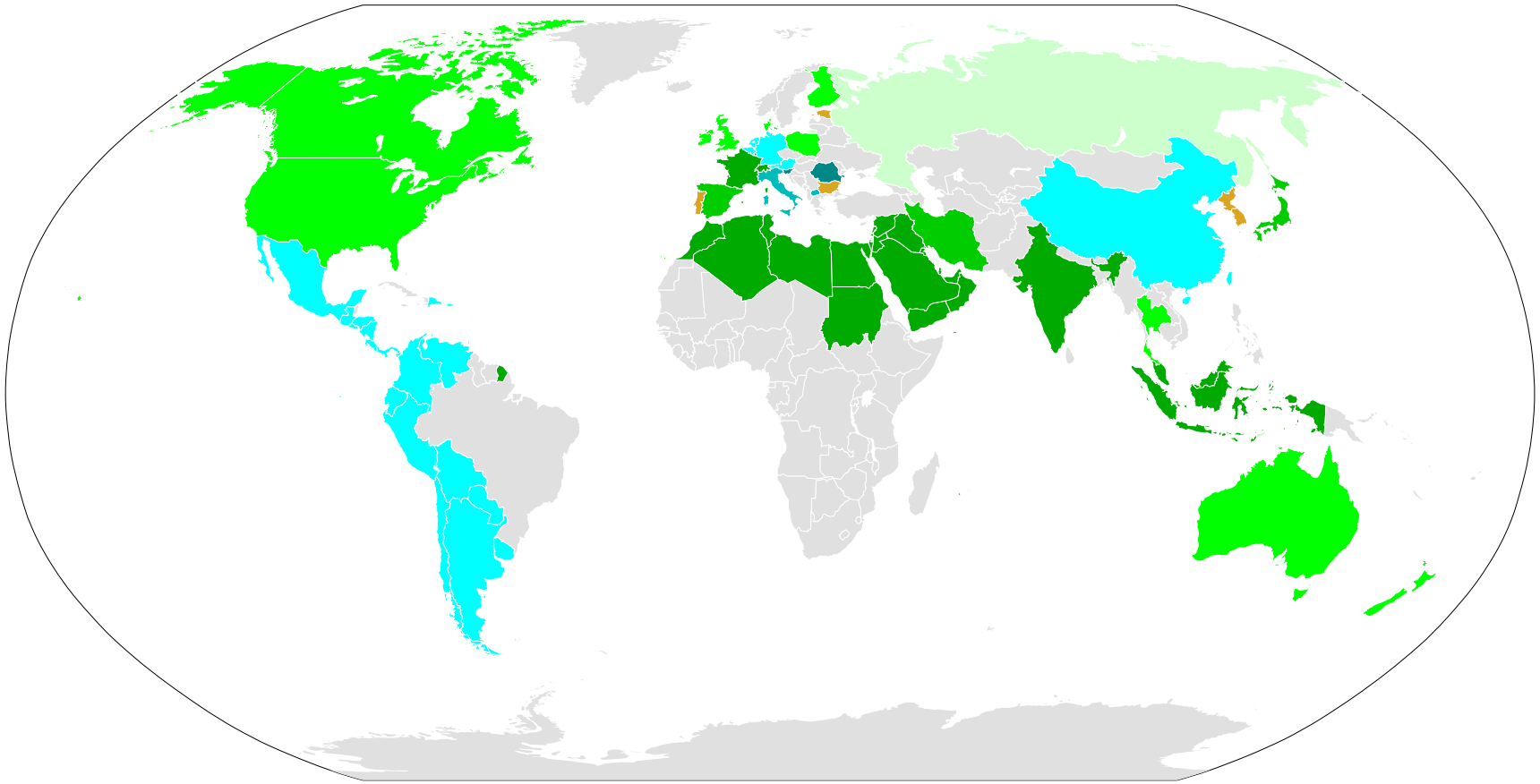
Wordnets in the world 2011-06



Green is free; Blue is research only; Brown costs money

NANYANG TECHNOLOGICAL UNIVERSITY

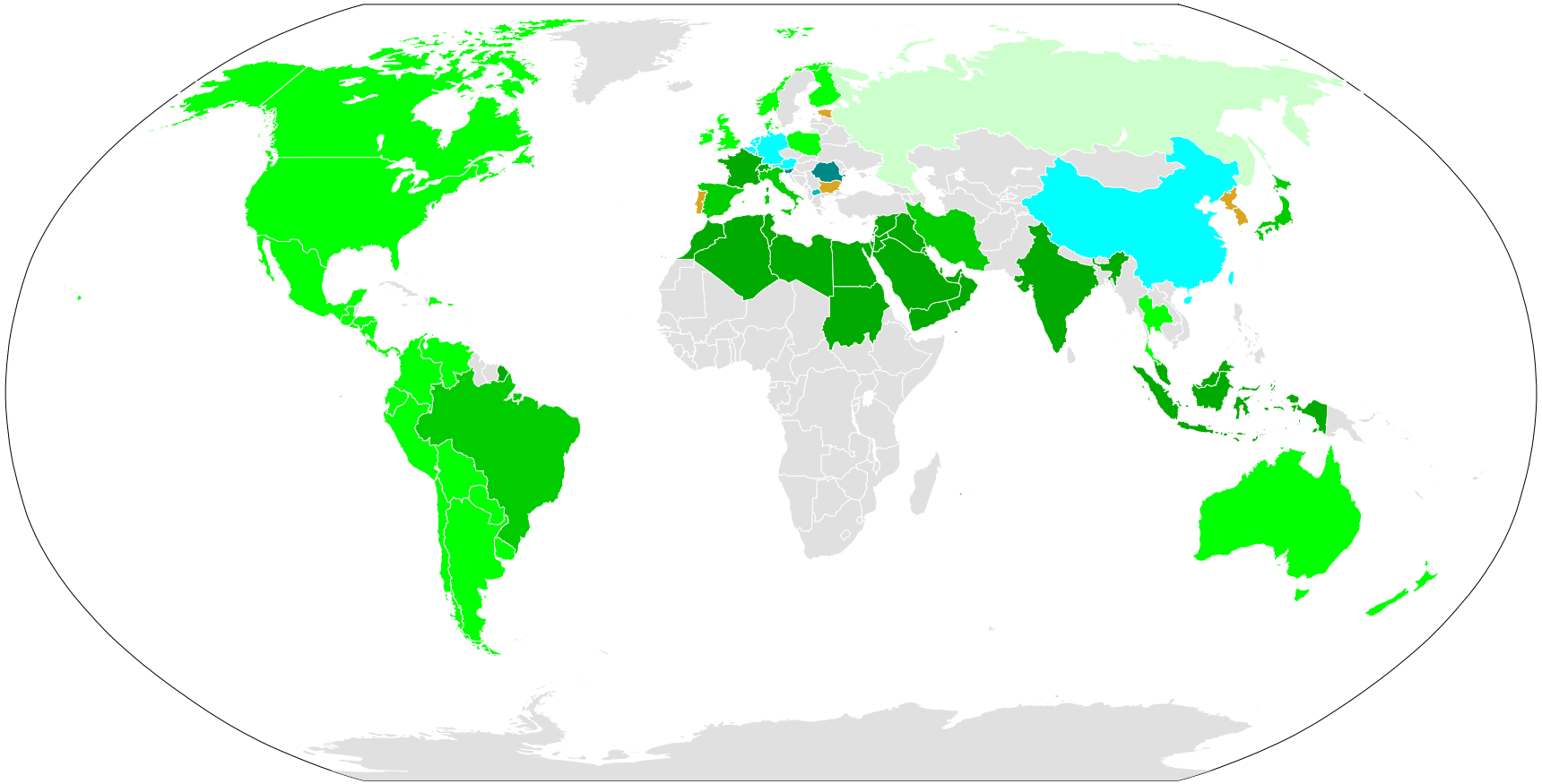
Wordnets in the world 2012-01



Added: Finnish, Persian, Bahasa

NANYANG TECHNOLOGICAL UNIVERSITY

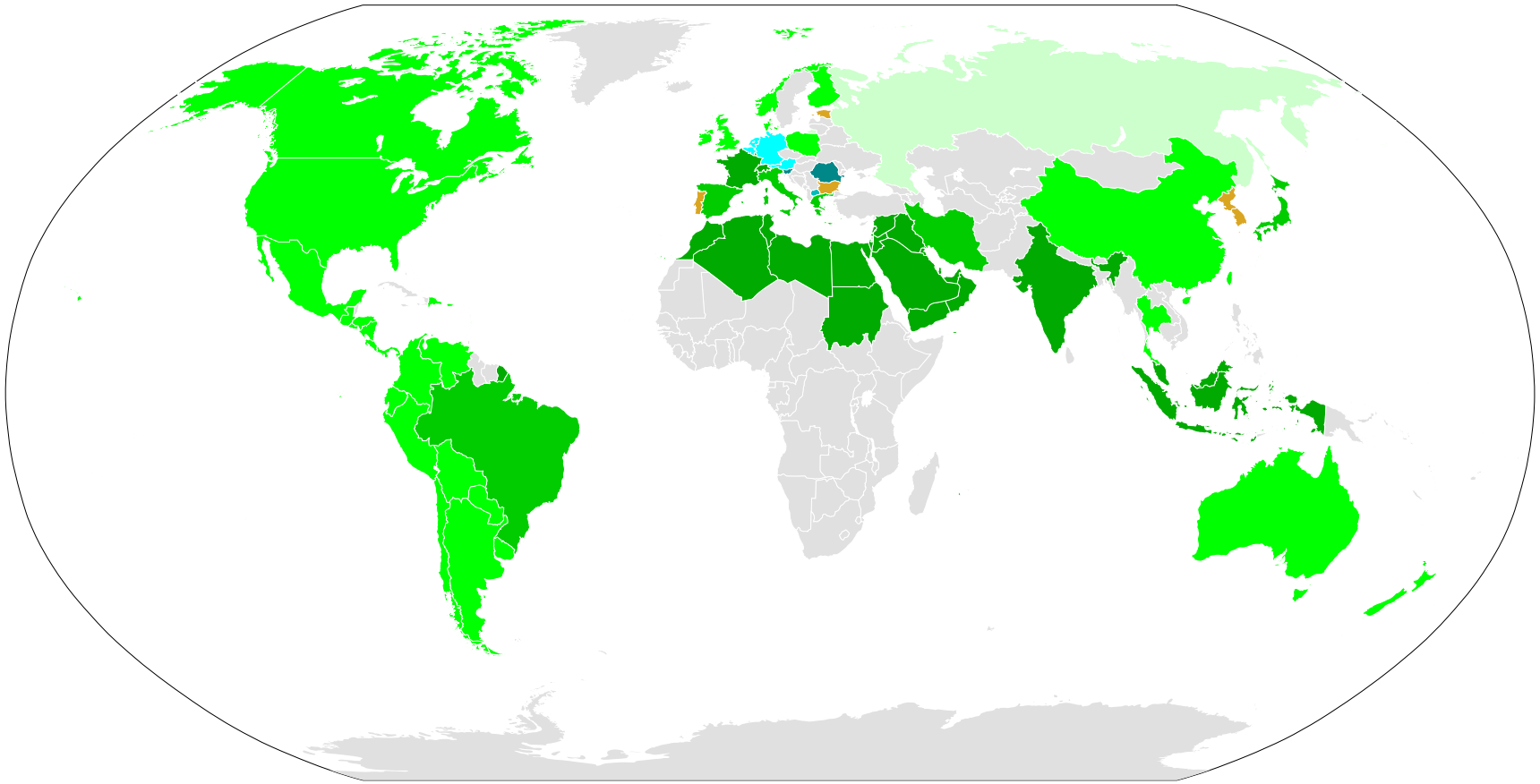
Wordnets in the world 2012-06



Added: Norwegian; Freed: Italian, Portuguese, Spanish

NANYANG TECHNOLOGICAL UNIVERSITY

Wordnets in the world 2013-06



Added: Greek; Freed: Chinese

NANYANG TECHNOLOGICAL UNIVERSITY

Wordnets in the world 2014-06

- Added: Swedish, Slovenian, Romanian
- Freed: Dutch
- Added 150 automatically built wordnets (> 500 synsets)
- Linked sentiment and temporal analyses
- Play with it here: `compling.hss.ntu.edu.sg:/omw/`

NANYANG TECHNOLOGICAL UNIVERSITY

Methodological Aside

- Studying language is hard: linguistic description and analysis is labor intensive and time consuming (although often fun)

- There is a lot to study
  - It is inefficient to have to redo this analysis
  - We don't really gain from having multiple dictionaries

$\Rightarrow$ we should make our data as easy to use as possible

  - share it as open data (open source license)
    corpora, lexicons, stimuli, programs, grammars, . . .

Disclaimer: this research was partially funded by Creative Commons

NANYANG TECHNOLOGICAL UNIVERSITY

Effects of different licenses

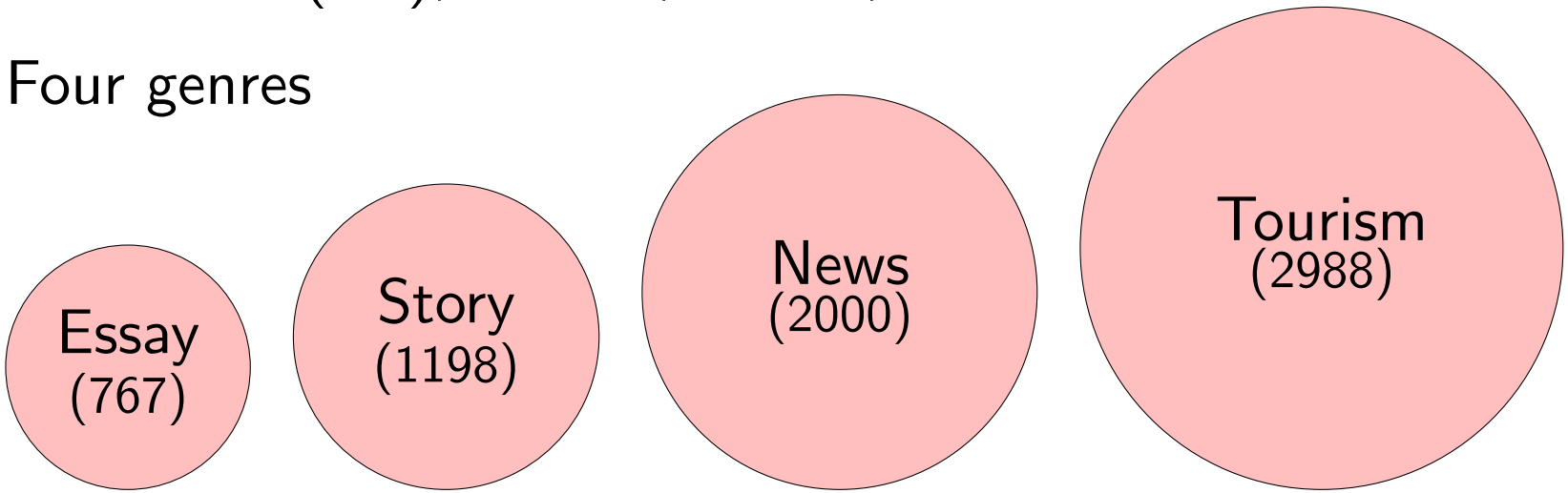
Size	Date	Open	Free	Non free
Large	2009	Danish/Thai		Korean
		8/10		5
Large	2008	Japanese	Dutch	
		24	19	
Small	2008	French	Slovenian	Bulgarian
		22	13	3

- Uptake of a resource partially depends on how usable (legally accesible) the resource is (and many other factors)
- Open licenses may still be incompatible: CC-BY $\not\leftrightarrow$ GPL, CC-BY-SA $\not\leftrightarrow$ CC-BY-SA-NC

NANYANG TECHNOLOGICAL UNIVERSITY

NTU Multilingual Corpus

- Parallel data
- Opportunistically collected from translated texts we could redistribute
- English (eng), Mandarin Chinese (cmn), Japanese (jpn), Indonesian (ind), Korean, Arabic, Vietnamese and Thai
- Four genres



NANYANG TECHNOLOGICAL UNIVERSITY

Now checking the annotation

- Essay (CEJ:many)
  - *The Cathedral and the Bazaar*
- Story (CEJ:many)
  - *The Adventure of the Dancing Men*
  - *The Adventure of the Speckled Band*
- Tourism (CEI:JVKA)
  - *Your Singapore*
- News (CEJ): *Mainichi Daily News*

NANYANG TECHNOLOGICAL UNIVERSITY

Monolingual Tagging

Genre	English				
	Concepts	in WN	%	Tagged	%
Essay	10,435	9,588	91.9	8,607	82.5
Story	11,340	10,761	94.9	9,550	84.2
Tourism	40,844	35,979	88.1	32,990	80.8

	Chinese				
	Concepts	in WN	%	Tagged	%
Essay	11,365	8,620	75.8	8,773	77.2
Story	12,630	9,521	75.4	8,737	69.2
Tourism	43,164	23,699	54.9	24,663	73.2

Multilingual Tagging

- Attempt to link concepts across languages
- Can link many-to-many

NANYANG TECHNOLOGICAL UNIVERSITY

How are meanings linked?

	Type	Example
$=$	same concept	*say* ↔言う *iu* "say"
$\supset$	hypernym	*wash* ↔洗い落とす *araiotosu* "wash out"
$\supset^2$	2nd level	*dog* ↔ 動物 *doubutsu* "animal"
$\subset$	hyponym	*sunlight* ↔光 *hikari* "light"
$\subset^n$	nth level	
$\sim$	similar	*notebook* ↔メモ帳 *memochou* "notepad"
		$dull_a$ ↔くすむ *kusumu* "darken"
$\approx$	equivalent	*be content with my word* ↔
		わたくし の 言葉 を 信じ-て "believe in my words"
!	antonym	*hot* ↔寒く=ない *samu=ku nai* "not cold"
#	weak ant.	*not propose to invest* ↔
		思いとどまる *omoi=todomaru* "hold back"

NANYANG TECHNOLOGICAL UNIVERSITY

Numbers of Links

Link	Story		Essay	
	#	%	#	%
=	2,642	41.7	2,155	48.9
<	107	1.7	31	0.7
>	205	3.2	123	2.8
∼	2184	34.5	1464	33.2
d	166	2.6	72	1.6
D	1,149	18.1	624	14.2
m	16	0.3	1	0.0
M	15	0.2	5	0.1
#	23	0.4	7	0.2
Total	6,336	100.0	4,407	100.0
Concepts	10,435		11,340	

NANYANG TECHNOLOGICAL UNIVERSITY

Very much not one-to-one

(3) Put$_a$ that way,$_B$ the question$_c$ answers$_D$ itself.

这样$_B$ 一 问$_e$, 答案$_D$ 自明$_f$。

zhèyàng yī wèn, dá'àn zìmíng.

like this one ask, answer self-evident

"Asking like this, the answer is self-evident."

NANYANG TECHNOLOGICAL UNIVERSITY

(4) The bullet had passed through the front of her brain.

子弹	是	从	她的	前额	打	进去	的。
Zǐdàn	shì	cóng	tāde	qián'é	dǎ	jìnqù	de.
bullet	is	from	her	forehead	shoot	enter	

"The bullet was shot in from her forehead"

NANYANG TECHNOLOGICAL UNIVERSITY

Pronomilization

(5) She$_i$ shot him$_j$ and then herself$_i$

a. 奥-さん が 旦那-さん を
oku-san ga danna-san wo
wife-HON NOM husband-HON ACC
撃って 、 それから 自分 も 撃った
utte , sorekara jibun mo utta
shoot-CONJ , and+then self too shoo-PST
Wife$_i$ shot husband$_j$ and then shot self$_i$ too

NANYANG TECHNOLOGICAL UNIVERSITY

Pronomilization

(6) She$_i$ shot him$_j$ and then herself$_i$

a. 她 拿 枪 先 打 丈夫 , 然后
tā ná qiāng xiān dǎ zhàngfū , ránhòu
3SG take gun first shoot husband , and+then
打 自己
dǎ zìjǐ
shoot self

She$_i$ took the gun to first shoot husband$_j$, and then shot self$_i$

NANYANG TECHNOLOGICAL UNIVERSITY

Ongoing and Future Work

- Improving the tagging guidelines
  will share on-line
- Improving matching (many minor variations)
  add variants to Japanese wordnet
  like to do so for English $tool\ kit \rightarrow \boldsymbol{toolkit}$.
  improve lemmatization (use a real parser)
- Finish tagging
- Look at some individual phenomena
  - Pronouns
  - Chinese Idioms (成语 *chéngyǔ*)
  - English possessive idioms ($X\ looses\ X's\ head$)

NANYANG TECHNOLOGICAL UNIVERSITY

Affectedness

What can we do?

- For things that are lexicalized (conventionally)
  - such as
    - Czech markers
    - ? affected arguments and telic classes
    - Beaver's classes?
  - Mark them (with a new feature?, through inheritance)
  - Link related senses (*throw in*, *throw out*)
  - Polish does this for e.g. PERFECTIVE/IMPERFECTIVE (and introduces the great relation FUZZYNYMY)
- Can we leverage cross-linguistic differences to do this semi-automatically

- For things that are not lexicalized
  - Investigate their distribution in a corpus
  - See how the same phenomenon is expressed in different languages
  - See if it correlates with other phenomena
    - verb class
    - semantic class of arguments
    - ...
  - Is affectedness marked as often in different languages?
    - if not, why not?

*

References

Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue. 64–71.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Miller, G. (1998). Foreword. In Fellbaum (1998), pages xv–xxii.

Tan, L. and Bond, F. (2012). Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). *International Journal of Asian Language Processing*, 22(4):161–174.