Cross-Lingual Disambiguation and the Open Multilingual Wordnet

Francis Bond

Linguistics and Multilingual Studies
Nanyang Technological University
National Institute for Information and Commmunications, Japan
bond@ieee.org

2012-06-21

Coli-2012

NANYANG TECHNOLOGICAL UNIVERSITY

Overview

- Research on Computational Linguistics at Nanyang
- Cross-lingual Disambiguation
- Toward an Open Multi-lingual WordNet
- A survey of released wordnets and their licenses
- The effects of license choice
- Erasmus Mundus Multi

NANYANG TECHNOLOGICAL UNIVERSITY

Bond Lab Members

- Petter Haugereid (MT and HPSG)
- Matthieu Morey (MRS comparison and Graph Matching) Erasmus Post Doc visitor
- Fan Zhenzhen (Chinese HPSG)
- Tan Liling (Multilingual Corpus Construction, WSD)
- Lea Frermann (Cross-lingual Disambiguation) Erasmus MA visitor
- Dominikus Wetzel (rephrasing for SMT: negation) Erasmus MA visitor
- Suerya, Nurril, Helmy (Wordnet Bahasa)
- Huini, ShuWen, Jeanette, Yui Ting (Chinese/Japanese Wordnets)

The Core Problem of MT (& NLU)

(1) 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 do we teach a computer?

NANYANG TECHNOLOGICAL UNIVERSITY

Languages Mark Different Things

- 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)
- Shared level somewhere beyond syntax
  This is the level that interests me: **semantics**

NANYANG TECHNOLOGICAL UNIVERSITY

How to get to meaning?

- Careful hand analysis: HPSG, LFG, Wordnet, Framenet, Penn Treebank, Negra, Goitaikei
- Supervised learning on annotated data
- Unsupervised learning on vast amounts of data

Can we combine these?

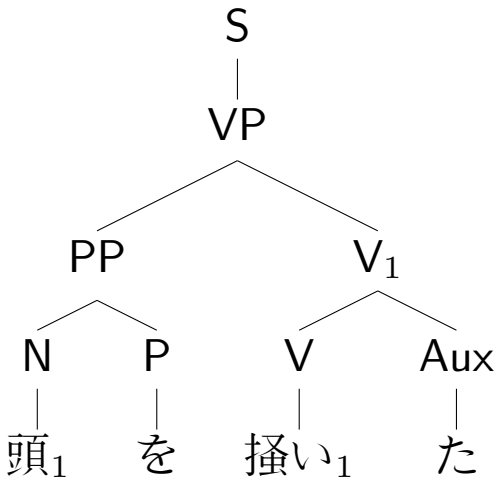
NANYANG TECHNOLOGICAL UNIVERSITY

Exploit Cross-lingual Differences

- Build on a core of hand written monolingual and cross-lingual resources
- Learn as much as we can from other languages
  - Structural Semantics
  - Lexical Semantics

Start: Rich Annotation

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



atama$_1$	is-a	bodypart
kaku$_1$	is-a	itch
kaku	ARG1	zero-pronoun
kaku	ARG2	atama
kaku	TENSE	past

Syntax & Lexical Semantics　　Structural Semantics

NANYANG TECHNOLOGICAL UNIVERSITY

Why both Lexical and Structural?

- As a scientist — semantics motivates syntax
- As an engineer
  - We want semantic output and we want higher accuracy
  - Syntax helps with WSD (Stevenson)
  - Semantics helps with parse ranking (Fujita et al, Egirre et al)
  - We want to look at joint models

NANYANG TECHNOLOGICAL UNIVERSITY

Rich Annotation

- Empirical NLP needs data
- There is no corpus marked up with what we want (Ontonotes is close)
  - Structural Semantics
    Predicate-Arguments, Deep Case, Quantification
  - Lexical Semantics
    Context dependent sense information from an ontology
- With this data we could start to answer some of the questions
  - Still need more discourse/context

- So we add the information
  - HPSG Tree banking
  - WordNet Sense annotation
  - Open data
- And build the grammars and wordnets
  - Jacy, MCG with DFKI, ERG with CSLI
  - Japanese wordnet with NICT, . . .

Too much hard work

Bootstrapping Annotation Bilingually

- Languages are ambiguous in different ways

  (3) ϕ 頭 を 掻い た

  I scratched my head

- We can use the two sentences to disambiguate each other:
  - Zero-pronoun is 1st person (I$\leftrightarrow \phi$)
  - kaku$_1\leftrightarrow$scratch$_3$
  - atama$\leftrightarrow$head could be `boss` or `body part` or `top`, not `grammatical head`

- This also works for some syntactic disambiguation

NANYANG TECHNOLOGICAL UNIVERSITY

Open Research Questions

- How much can be disambiguated in this way?
  depends on language pair, translatation quality
  Ja↔En, Tanaka Corpus reduce ambiguity to $\approx 30\%$

- What is the best level for alignment?
  - deeper processing aligns better
  - less robust, so less cover

- Currently experimenting with Japanese and English
  - HPSG grammars for parsing — DELPH-IN
  - WordNet for the sense inventory — Global WordNet Association
  - MT system for deep alignment — DELPH-IN
  - n-grams for shallow alignment — MOSES
  - MRS n-grams for hybrid alignment — DELPH-IN/MOSES

NANYANG TECHNOLOGICAL UNIVERSITY

Concrete Projects

- Wordnets (Global Wordnet association)
  - **Japanese WordNet** (with Toyohashi, NICT)
  - **Wordnet Bahasa** (with MMU, BPPT)
  - Chinese WordNet (with South-Eastern)
  - English Wordnet (with Princeton)
- Grammars (HPSG: DELPH-IN)
  - Japanese grammar: NTU + Darmstadt Poly + Washington
  - Chinese: DFKI + Shanghai + NTU (Fan Zhenzhen)
  - Eng: Stanford

NANYANG TECHNOLOGICAL UNIVERSITY

- NTU Multilingual Corpus (Tan Liling)
  - JEC sentence aligned word-net tagged
  - JECIVK sentence aligned (JECI will be tagged)
- Cross-lingual disambiguation (sense, structure)
- Japanese-English MT
- Korean-Japanese-English MT (with KHU in Korea)

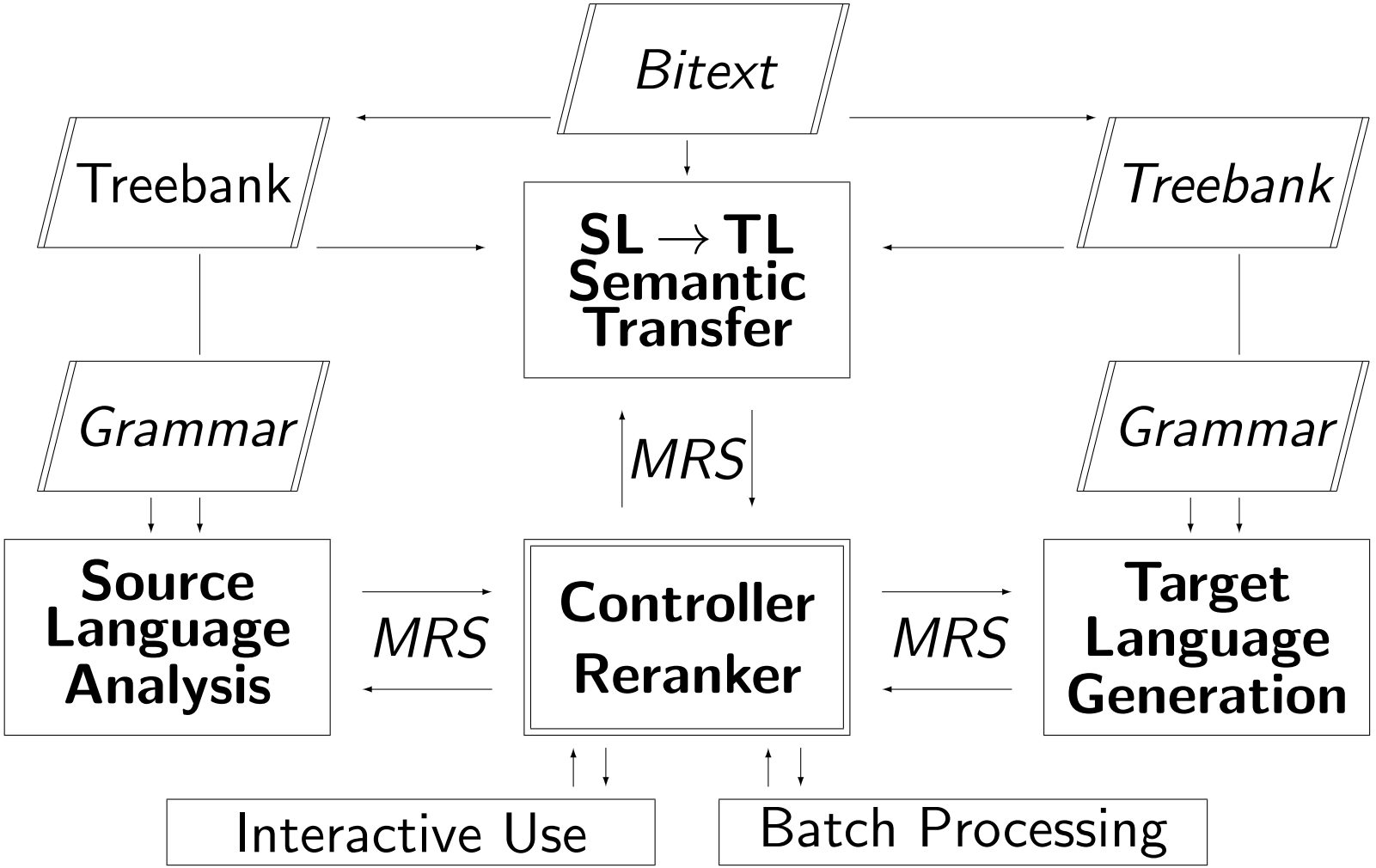
NANYANG TECHNOLOGICAL UNIVERSITY

Semantic Transfer-based MT

- HPSG grammars for parsing and generation
- stochastic models at every step
  (parsing, transfer, generation, end-to-end)
- core of hand-written rules
- open rules (some quite complex) learned from corpora
- limited coverage, high quality
  beating MOSES for 58/100 sentences
- Now J-E, soon to add K-J-E, like to add C-E, M-E

NANYANG TECHNOLOGICAL UNIVERSITY

MT System Architecture



NANYANG TECHNOLOGICAL UNIVERSITY

Jaen Latest results

	BLEU	METEOR	HUMAN
JaEn First	16.77	28.02	**58**
MOSES	**30.19**	**31.98**	42

BLEU Comparison of Jaen and MOSES (1194 items)
Transfer rules learned from the bitext

- S: 偏見 は 持つ べき で は ない 。
  R: We shouldn't have any prejudice.
  M: You should have a bias.
  J: I shouldn't have prejudice.

Cross-lingual sense disambiguation

- Tag multiple languages with a common ontology and take the intersection
  - Only possible if synsets can be linked
  - Precision: $\approx 80\%$ (jpn$\cap$eng)
  - Recall: $\approx 60\%$
  - Easy to do with limited resources: just segmentation and a WN

Revealing Meaning via Translation

➢ Crosslingual links narrow the interpretations

		委員長 として 党の 結束を 大切に したい			
As the chairperson,	A				作为
I		B			委员长，
would like to					我
regard				C	希望
the unity of			E		维护
the party$_i$		F			党内
as important.				G	团结。

➢ We can make the implicit meaning become explicit

➢ party$_{i=1}$ “an organization to gain political power” $\equiv$ 党$_1$

NANYANG TECHNOLOGICAL UNIVERSITY

Revealing Meaning (2)

- English
  - party$_1$ "an organization to gain political power"
  - party$_2$ "a group of people gathered together for pleasure"
  - party$_3$ "a band of people associated temporarily in some activity"
  - party$_4$ "an occasion on which people can assemble for social interaction"
- Japanese
  - 党$_1$ "an organization to gain political power"

NANYANG TECHNOLOGICAL UNIVERSITY

Basic Approach

1. Take a bitext

2. Tag each side with WordNet senses
   (including unknown)

3. Align the texts (by word or synset)
   `giza++`

4. Take the intersection of the synset candidates
   (if unknown, suggest to add to the synsets)

5. Prefer MWEs if compatible

NANYANG TECHNOLOGICAL UNIVERSITY

Corpus Ambiguity

- If each word has one sense, then ambiguity is 1
  this would be true for a manually tagged corpus
- Ambiguity in English text is 9.0
  common words are more ambiguous (WordNet is 1.9)
- MWEs add ambiguity

machine	*translation*	*is*	*hard*
n:6 v:2	n:7	n:1 v:13	r:10 a:12
n:1			

- We say that ***machine*** is 9 (6+2+1)

NANYANG TECHNOLOGICAL UNIVERSITY

Ambiguity reduction

Text	Eng	Jpn	Eng∩Jpn
The Adventure of the Dancing Men	9.4	5.7	3.1
The Adventure of the Speckled Band	9.3	5.8	3.2
The Cathedral and the Bazaar	8.2	5.1	2.6

Preliminary results

- Redo after refining (expect further ambiguity reduction)
  - Adding unknown words and senses to WordNet
  - Improving alignment
- Chinese tagging now underway: can we halve again?

NANYANG TECHNOLOGICAL UNIVERSITY

Discussion

- These results are surprisingly good
  - Can use for discriminative WSD training
  - Can use as pre-processing for manual tagging
    $\Rightarrow$ tag twice as fast
- Very cheap — requires only
  - Bitext
  - POS tagger/segmenter
  - POS to WordNet POS mapping
  - Wordnets — expensive
  - Word Aligner (`GIZA++`)

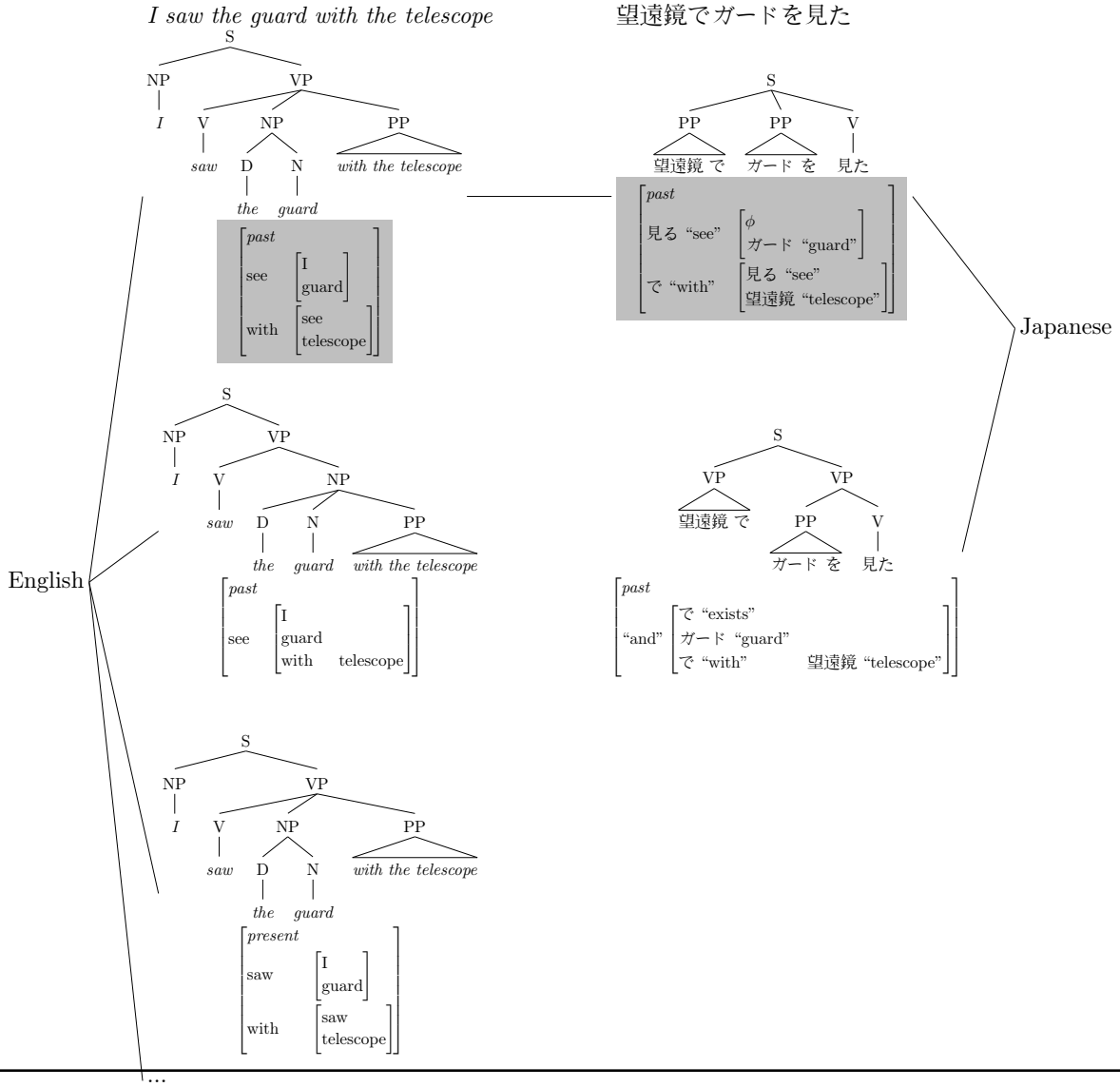
NANYANG TECHNOLOGICAL UNIVERSITY

CL WSD: Going further

- Alignment between languages is not just word-to-word
  離党 ≡ *leave the party* ≡ 党を　離れる
  butter ≡ *spread butter on* ≡ バターを 塗る
- We can identify units at various levels (constructions)
  - need to enrich the WordNets (or external grammars)
  - bilingual data leads to monolingual paraphrases
- Systematic Ambiguity (Universal ?)
  - Plant/Produce
    * *apple*: fruit of the apple tree
    * *apple*: the apple tree

Cross-lingual Syntactic Disambiguation

- We can use translations to disambiguate syntax
  - ITG, DOP, syntax-based MT, . . . directly match trees
  - But translations match on the semantic level
- Exploit MT systems to match meaning
  - Consider Japanese and English Text
    - parse Japanese to J_M (meaning)
    - translate J_M to E'_M
    - parse English to E_M
    - match E'_M to E_M to find the correct interpretation
  - Preliminary results show 69% accuracy on full sentences


I saw the guard with the telescope
望遠鏡でガードを見た
English
Japanese
S
NP
VP
I
V
NP
PP
saw
D
N
with the telescope
the
guard
past
see
I
guard
with
see
telescope
望遠鏡 で
ガード を
見た
見る “see”
ϕ
ガード “guard”
で “with”
望遠鏡 “telescope”
VP
で “exists”
“and”
present
saw
...

NANYANG TECHNOLOGICAL UNIVERSITY

Cross Lingual Disambiguation

Reduced ambiguity to 30%

for the 71% of sentences that parse and partially translate

	English		Japanese	
	Prec	**F**	**Prec**	**F**
Included	0.820	0.897	0.804	0.887
First Rank	0.659	0.791	0.676	0.803
MRR	0.713	0.829	0.725	0.837

Accuracy and F-scores for disambiguation performance of our system. Recall was 99% in every case. 'Included': inclusion of the gold parse in the reduced set of parses or not. 'First Rank': ranking of the preferred parse as top in the reduced list. 'MRR': mean reciprocal rank of the gold parse.

NANYANG TECHNOLOGICAL UNIVERSITY

Positive Feedback

- Better parsing $\Rightarrow$ Better rule acquisition
- Better rule acquisition $\Rightarrow$ Better machine translation
- Better machine translation $\Rightarrow$ Better treebanks
- Better treebanks $\Rightarrow$ Better parsing

  And there is much more that can be learned

- MWEs, senses, . . .

NANYANG TECHNOLOGICAL UNIVERSITY

Dynamic Updating

- Improve the model (grammar/wordnet)
- (Semi-)Automatically update the parallel corpus
- Select parses with semantic models
  - You can train on different languages!
  - Although of course, meaning will always be changed a little

WordNets

NANYANG TECHNOLOGICAL UNIVERSITY

Motivation

- Would like to compare meaning across languages quantitatively if possible
- Experiments tend to end up measuring resource maturity e.g., semantic correlation of countability, classifiers, valence
- Want a collection of resources of comparative completeness
  - Build our own (Japanese, Malay)
  - Contribute to others
- Resources need to be accessible and usable
  - legally OK to use
  - of sufficient quality, size and with a documented interface

NANYANG TECHNOLOGICAL UNIVERSITY

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

NANYANG TECHNOLOGICAL UNIVERSITY

Composition

- Lexical items are categorised into ∼115,000 glossed **synsets** (= synonym sets)

```
1. enrichment -- (act of making fuller or more
   meaningful or rewarding)
2. enrichment -- (a gift that significantly increases
   the recipient's wealth)
```

- Lexical relations at either the synset level or sense (= combination of lexical item and synset) level
- Strongly lexicalist (orginally):
  - synsets only where words exist
  - but many multiword expressions ($\approx 50\%$)

NANYANG TECHNOLOGICAL UNIVERSITY

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

An Accessible Multilingual Wordnet

- To help us in disambiguation when making the Japanese and Bahasa wordnets we needed to link various wordnets
- There were many small idiosyncrasies ☹
- To make it easier for others we have released our combined database + scripts
  only for those resources whose license allows it
- Hope to be superseded by a more flexible framework (ILI)
  - That allows new (especially) non-English synsets
  - That allows variants

NANYANG TECHNOLOGICAL UNIVERSITY

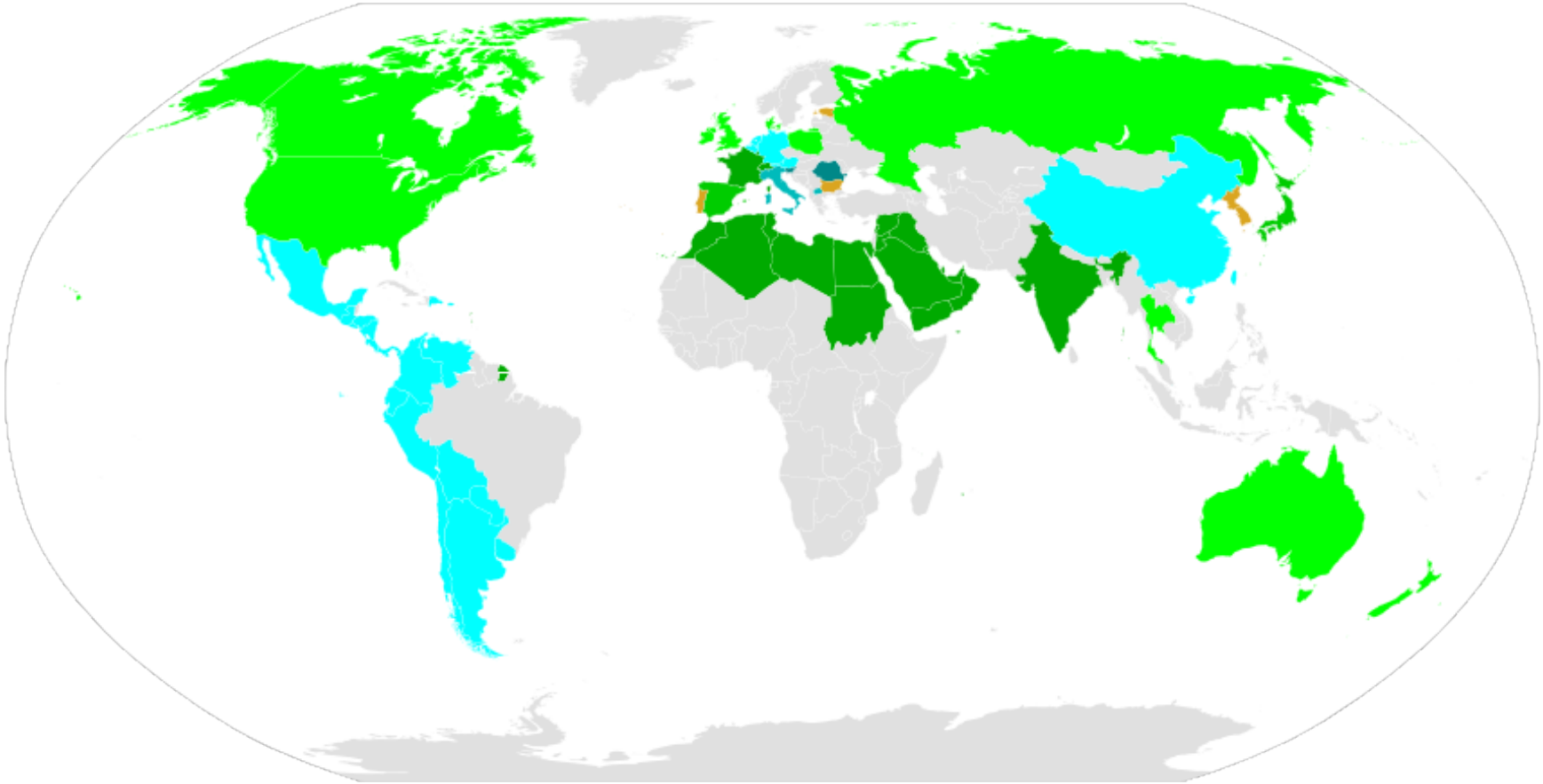
Current State (last week)

Wordnet	Lang	Synsets	Words	Senses	Core	Licence
Arabic WordNet	arb	10,165	14,595	21,751	48%	CC BY SA 3.0
Princeton WordNet	eng	117,659	148,730	206,978	100%	wordnet
Persian Wordnet	fas	17,759	17,560	30,461	41%	Free to use
FinnWordNet	fin	116,763	129,839	189,227	100%	CC BY 3.0
WOLF	fre	32,466	37,996	46,188	48%	CeCILL-C
Hebrew Wordnet	heb	5,448	5,325	6,872	27%	GPL
Japanese Wordnet	jpn	57,178	91,959	158,062	95%	wordnet
Wordnet Bahasa	ind	19,260	19,659	48,317	98%	MIT
Wordnet Bahasa	zsm	19,267	19,638	48,321	98%	MIT
OpenWN-PT	por	34,087	35,811	51,471	77%	CC by SA 3.0
Thai Wordnet	tha	73,350	82,504	95,517	81%	wordnet

- `http://casta-net.jp/~kuribayashi/multi/`
- Just got: Italian; Spanish, Catalan, Galician, Basque Danish, Norwegian (Bokmal/Nynorsk)

NANYANG TECHNOLOGICAL UNIVERSITY

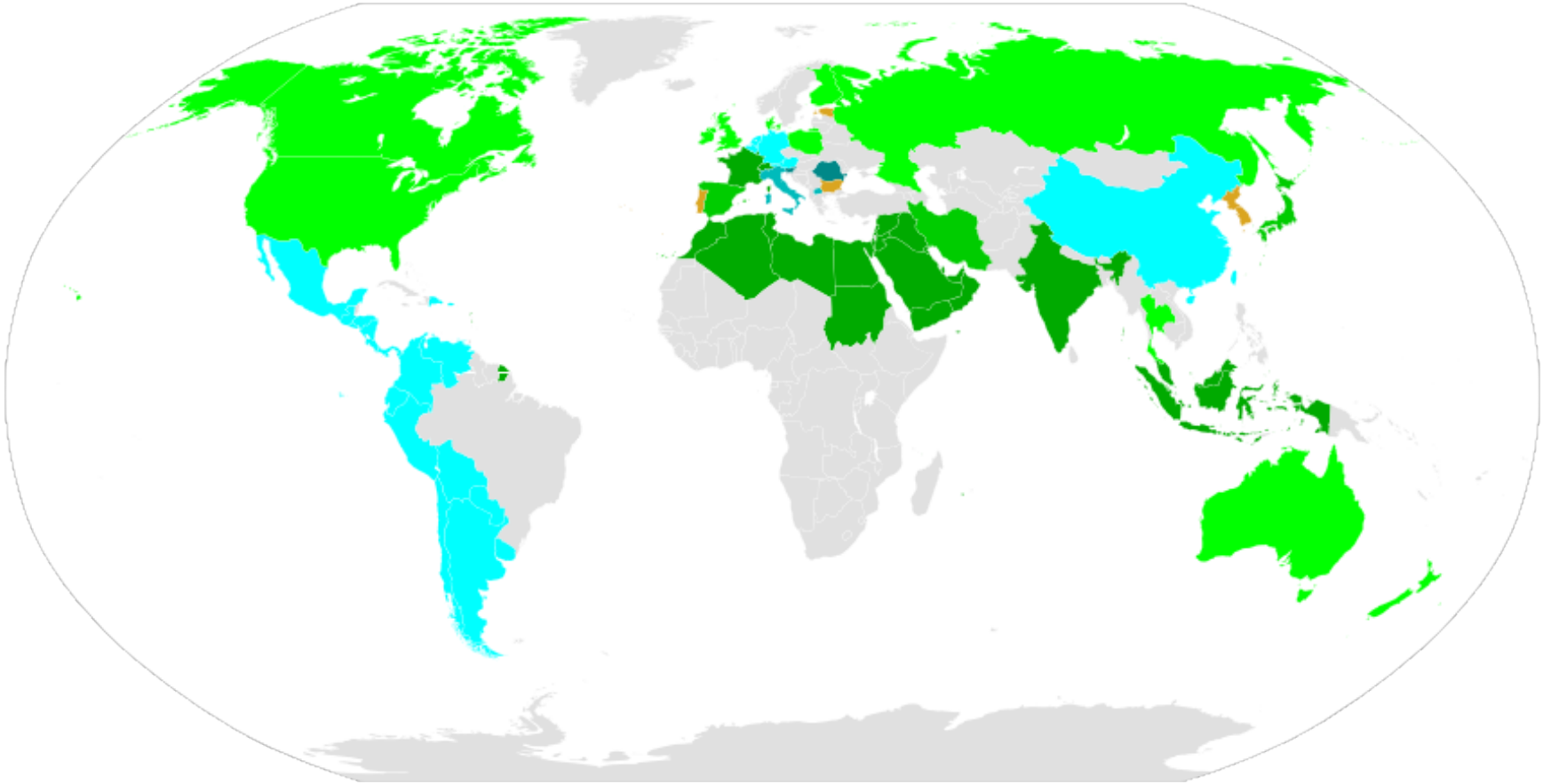
Wordnets in the world 2011-06



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

NANYANG TECHNOLOGICAL UNIVERSITY

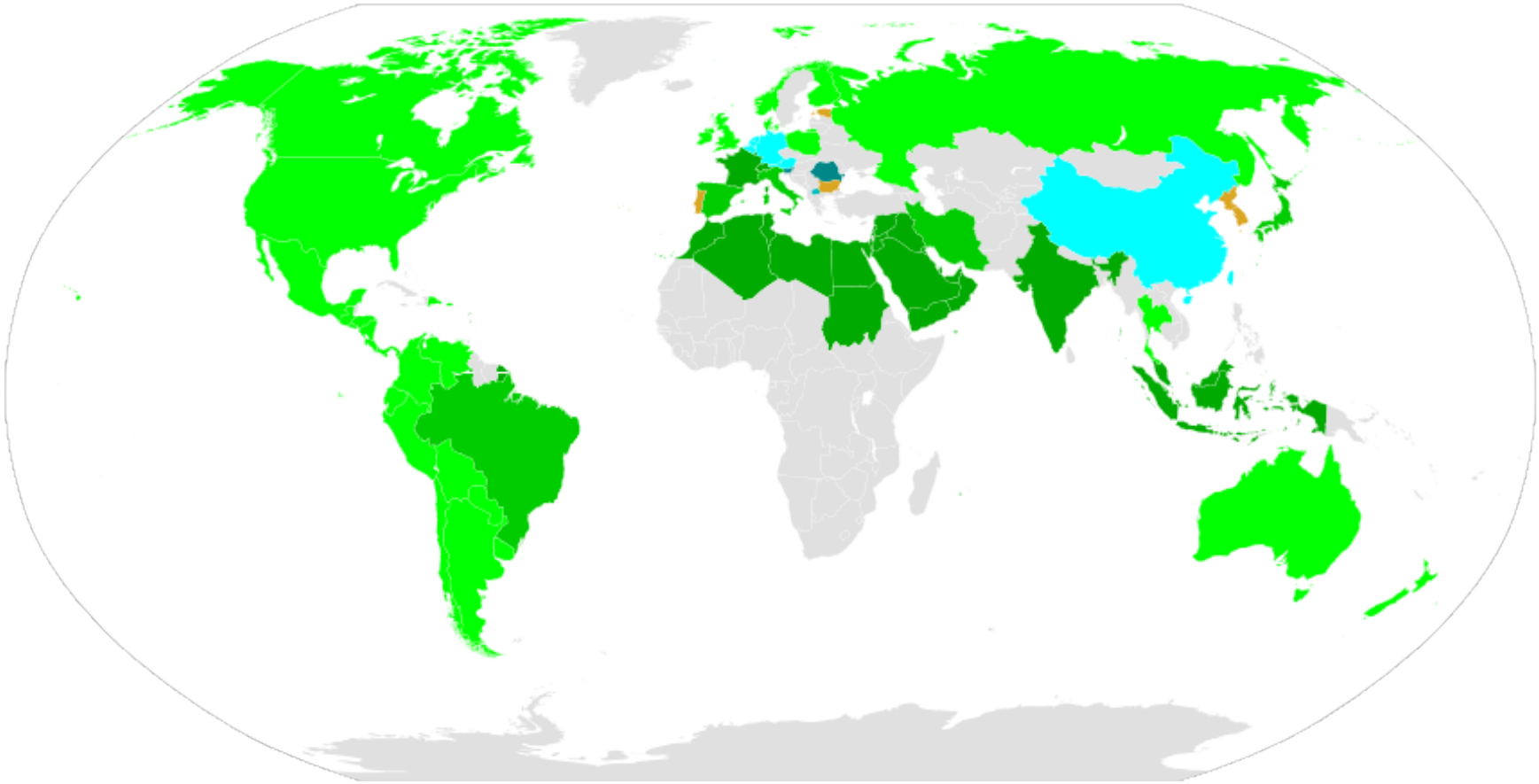
Wordnets in the world 2012-01



Added: Finnish, Persian, Bahasa

NANYANG TECHNOLOGICAL UNIVERSITY

Wordnets in the world 2012-06



Added: Norwegian; Freed: Italian, Portuguese, Spanish

NANYANG TECHNOLOGICAL UNIVERSITY

What is lacking?

- German and Chinese ☹
- Proper handling of orthographic variants
  - Japanese: 桧, 檜, ひのき, ヒノキ, 火の木 *hinoki*
  - Hebrew, Arabic: with and without diacritics
  - English: color, colour; data base, data-base, database
- Richer morphological information (not just v,a,n,r)
- Substructure for MWEs
- Sense specific frequencies (cross-lingually annotate)

NANYANG TECHNOLOGICAL UNIVERSITY

Multi-lingual issues

- Representation of lexical gaps
  - some concepts are only lexicalized for some languages
- Relations may vary for different languages
- Names for synsets would be really useful

NANYANG TECHNOLOGICAL UNIVERSITY

Next steps

- Standard data formats
  - Sharing tools and corpora
  - Sharing best practice
- Opening up data
- Cooperative Interlingual Index
  - *last year*
  - *tsumetai* vs *samui*
  - Different ILIs for Italian and Spanish
- Language Independent WSD + language dependent WSD

What is a free license?

The Open Source Definition

1. Free Redistribution
2. Source Code
3. Derived Works

4. Integrity of The Author's Source Code
5. No Discrimination Against Persons or Groups
6. No Discrimination Against Fields of Endeavor
7. Distribution of License
8. License Must Not Be Specific to a Product
9. License Must Not Restrict Other Software
10. License Must Be Technology-Neutral

NANYANG TECHNOLOGICAL UNIVERSITY

Why Open?

- Research should be open in principle:

  . . . the principle of openness in research - the principle of freedom of access by all interested persons to the underlying data, to the processes, and to the final results of research - is one of overriding importance.

- Making resources Open Source removes difficulties in distributing work or in continuing work at another institution.

- CL needs serious resources
  - They cannot be built and maintained by a single group
  - Open Source is the only practical way of achieving flexible multi-group collaboration

- CL needs standards and historically the successful ones have been created bottom-up.
- Seeing one's work used by other groups is very rewarding.
- People are generally enthusiastic about contributing to widely used work.
- Researchers are evaluated by the impact that their work has. Open Source work generally has more impact.
- Licensing attempts are extremely time-consuming.
- In my personal opinion, Restricting access to information about human language, the great achievement of our species, is **immoral.**

NANYANG TECHNOLOGICAL UNIVERSITY

Free Wordnets

Name	Language	# Synsets	Release	License
Princeton WN*€	English	155,000	1991	WordNet
	Fellbaum (1998): 6,821 citations			
FinnWordNet	Finnish	117,700	2010	WordNet
	Lindén and Carlson. (2010): 0 citations			
Russian WN	Russian	117,000	2004	Wordnet
	Balkova et al. (2008): 15 citations			
Thai Wordnet	Thai	73,593	2007	WordNet
	Thoongsup et al. (2009): 4 citations			
DanNet*	Danish	65,000	2008	WordNet
	Pedersen et al. (2009): 8 citations			
Japanese WN*	Japanese	57,000	2009	WordNet
	Isahara et al. (2008): 24 citations			
Catalan WN*	Catalan	42,000	1999	GPL
	Benítez et al. (1998): 17 citations			

⊙ A subset released under a less restrictive license; € A version from EuroWordNet is also available from ELRA

NANYANG TECHNOLOGICAL UNIVERSITY

Name	Language	# Synsets	Release	License
MultiWN$^{*€}$	Italian	38,877	2008	CC BY
	Pianta et al. (2002): 143 citations			
LSG	Irish Gaelic	32,742	2003	GNU FDL
	Scannell (2003): 11 citations			
Hindi WN	Hindi	28,687	?	GNU FDL
	Jha et al. (2001): 10 citations			
WOLF	French	22,000	2009	Cecill-C†
	Sagot and Fišer (2008): 22 citations			
Bahasa*	Malay/Indonesian	20,000	2011	CC BY
	Nurril *et al.* 2011: —			
Spanish WN$^{*⊙€}$	Spanish	15,556	2006	LGPL
	Farreres et al. (1998): 65 citations			
Catalan WN$^{*⊙€}$	Catalan	15,556	2006	LGPL
	Benítez et al. (1998): 17 citations			

$^{⊙}$ A subset released under a less restrictive license; $^{€}$ A version from EuroWordNet is also available from ELRA

NANYANG TECHNOLOGICAL UNIVERSITY

Name	Language	# Synsets	Release	License
Arabic WN*	Arabic	11,269	2008	CC BY SA
	Black et al. (2006): 28 citations			
Hebrew WN*	Hebrew	5000	2006	GPL
	Ordan and Wintner (2007): 0 citations			

*Results from our survey; Citation counts from Google Scholar (accessed on 2011-09-23)

† A variant of the LGPL

⊙A subset released under a less restrictive license

€A version from EuroWordNet is also available from ELRA

Release is the first release under this license

NANYANG TECHNOLOGICAL UNIVERSITY

Free for Research Wordnets

Name	Language	# Synsets	Release	License
Chinese WN*	Chinese	115,424	2008	research/com.
	Xu et al. (2008): 0 citations			
KorLex*	Korean	90,000	2007	research/com.
	subset of Yoon et al. (2009), no reference for nouns only			
Spanish WN*€	Spanish	62,000	1999	research/com.
	Farreres et al. (1998): 65 citations			
Cornetto*€	Dutch	70,371	2009	research/com.
	Vossen et al. (2008): 19 citations			
GermaNet*€	German	69,594	2011	research/com.
	Kunze and Lemnitzer (2002): 52 citations			
MWN*	Macedonian	33,276	2010	CC BY NC
	Saveski and Trajkovski (2010): 0 citations			

⊙ A subset released under a less restrictive license; € A version from EuroWordNet is also available from ELRA

NANYANG TECHNOLOGICAL UNIVERSITY

Name	Language	# Synsets	Release	License
Ro-WordNet*	Romanian	30,000	soon	no-derive.
	Tufiș et al. (2008): 9 citations			
Czech WN *€	Czech	29,000	1999	research/com.
	Pala and Smrž (2004): 34 citations			
SloWnet*	Slovene	20,000	2010	CC BY NC SA
	Fišer and Sagot (2008): 13 citations			

⊙ A subset released under a less restrictive license; € A version from EuroWordNet is also available from ELRA

NANYANG TECHNOLOGICAL UNIVERSITY

Available for Research Wordnets

Name	Language	# Synsets	Release	License
KorLex*	Korean	130,878	2007	research/com.
	Yoon et al. (2009): 5 citations			
Estonian*€	Estonian	47,000	—	ELRA
	Kerner et al. (2010): 0 citation			
BasqWN	Basque	30,281	?	ELRA
	Pociello et al. (2011): 0 citation			
BulNet*⊙	Bulgarian	23,715	2004	ELRA
	Koeva (2008): 3 citations			

⊙ A subset released under a less restrictive license; € A version from EuroWordNet is also available from ELRA

NANYANG TECHNOLOGICAL UNIVERSITY

Euro Wordnet

Name	Language	# Synsets	Release	License
EuroWordNet				
English Addition	English	16,361	1999	ELRA
Dutch	Dutch	44015	1999	ELRA
Spanish	Spanish	23370	1999	ELRA
Italian	Italian	48529	1999	ELRA
German	German	15,132	1999	ELRA
French	French	22,745	1999	ELRA
Czech	Czech	22,745	1999	ELRA
Estonian	Estonian	9,317	1999	ELRA
ItalWordNet	Italian	49,360	1999	ELRA
	Vossen (1998): 728 citations			

- These are sold by ELRA regardless of the availability of newer free versions.

⊙ A subset released under a less restrictive license; € A version from EuroWordNet is also available from ELRA

NANYANG TECHNOLOGICAL UNIVERSITY

Comments

- Most researchers very forward about sharing their data
  - Will share on request
  - Can often download (even for ELRA licensed Wordnets)!
  - Data occasionally redistributed under different license
- Some reasons given for opening data
  - Desire to make language resources publicly available
  - To encourage use and avoid license issuing overhead
- Some reasons given for not opening data
  - Financial — get money to support further construction
  - Quality Control — want to stop low quality forks

NANYANG TECHNOLOGICAL UNIVERSITY

Effects of different licenses

Size	Date	Open	Free	Non free
Large	2009	Danish/Thai		Korean
		8/4		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.

NANYANG TECHNOLOGICAL UNIVERSITY

Exceptions

- Chinese wordnet (Xu et al., 2008) has no citations, despite it being large and **free for research**. We suspect that this is because you must email and ask for permission to use it, which is a substantial barrier to use.

- The Italian wordnet (Pianta et al., 2002) has a very high number of citations. In this case it was developed as part of a multilingual wordnet with several other languages, thus giving it a large citation group.

- Finally, the Thai wordnet (Thoongsup et al., 2009), has relatively few citations, in this case it is also a part of a large project (the Asian Wordnet: Sornlertlamvanich et al. (2008)) which gets more citations (10).

NANYANG TECHNOLOGICAL UNIVERSITY

Reasons for not releasing

- Quality Control
- Non-free resources used in construction
- Sustainability (covering construction costs)

NANYANG TECHNOLOGICAL UNIVERSITY

Best Practice

- Open, standard license
  MIT, Wordnet, CC, lgpl
- Direct download
  no registration
- Standard, documented format
  Wordnet-LMF or tsv

NANYANG TECHNOLOGICAL UNIVERSITY

Thanks

- JSPS/NTU joint project on *Revealing Meaning Using Multiple Languages*
- The compilers of the Global WordNet Association's *Wordnets in the World* page
- The Creative Commons Catalyst Grant: *Assessing the effect of license choice on the use of lexical resources*
- Erasmus Mundus Multi Exchange program
  - `http://www.em-multi.eu/`
  - Exchanges still possible (MA/PhD/PostDoc/Staff)
    NTU_S, NTU_T, HK Poly, Macau, Borneo
  - Talk to me or Daniele Moretti (A2.2 Room 1.23)

*

References

Balkova, V., Sukhonogov, A., and Yablonsky, S. (2008). Russian wordnet: From UML-notation to internet/intranet database implementatiom. In Sojka, P., Pala, K., Smrž, P., Fellbaum, C., and Vossen, P., editors, *Proceedings of the Second International WordNet Conference — GWC 2004*, pages 31–38, Brno.

Benítez, L., Cervell, S., Escudero, G., López, M., Rigau, G.,

and Taulé, M. (1998). Methods and tools for building the Catalan wordnet. In *ELRA Workshop on Language Resources for European Minority Languages*, Granada, Spain.

Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., Bertran, M., and Fellbaum, C. (2006). The Arabic wordnet project. In *Proceedings of LREC 2006*.

Farreres, X., Rigau, G., and Rodríguez, H. (1998). Using wordnet for building wordnets. In *COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal, Canada.

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

Fišer, D. and Sagot, B. (2008). Combining multiple resources to build reliable wordnets. *Text, Speech and Dialogue*, LNCS 2546:61–68.

Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., and Kanzaki, K. (2008). Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.

Jha, S., Narayan, D., Pande, P., and Bhattacharyya, P. (2001).

A wordnet for Hindi. In *International Workshop on Lexical Resources in Natural Language Processing*, Hyderabad.

Kerner, K., Orav, H., and Parm, S. (2010). Growth and revision of Estonian wordnet. In Bhattacharyya, P., Fellbaum, C., and Vossen, P., editors, *5th Global Wordnet Conference: GWC-2010*, pages 198–202, Mumbai.

Koeva, S. (2008). Derivational and morphosemantic relations in Bulgarian wordnet. In *Intelligent Information Systems, XVI*, pages 359–389, Warsaw. Academic Publishing House.

Kunze, C. and Lemnitzer, L. (2002). Germanet —

representation, visualization, application. In *LREC*, pages 1485–1491.

Lindén, K. and Carlson., L. (2010). Finnwordnet — wordnet påfinska via översättning. *LexicoNordica — Nordic Journal of Lexicography*, 17:119–140. In Swedish with an English abstract.

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

Mohamed Noor, N., Sapuan, S., and Bond, F. (2011). Creating the open Wordnet Bahasa. In *Proceedings of the 25th Pacific*

Asia Conference on Language, Information and Computation (PACLIC 25), pages 258–267, Singapore.

Ordan, N. and Wintner, S. (2007). Hebrew wordnet: a test case of aligning lexical databases across languages. *International Journal of Translation*, 19(1):39–58.

Pala, K. and Smrž, P. (2004). Building Czech wordnet. *Romanian Journal of Information Science*, 7:79–88.

Pedersen, B., Nimb, S., Asmussen, J., Sørensen, N., Trap-Jensen, L., and Lorentzen, H. (2009). DanNet — the

challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). Multiwordnet: Developing an aligned multilingual database. In *In Proceedings of the First International Conference on Global WordNet*, pages 293–302, Mysore, India.

Pociello, E., Agirre, E., and Aldezabal, I. (2011). Methodology and construction of the Basque wordnet. *Language Resources and Evaluation*, 45(2):121–142.

Sagot, B. and Fišer, D. (2008). Building a free French wordnet

from multilingual resources. In (ELRA), E. L. R. A., editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Saveski, M. and Trajkovski, I. (2010). Automatic construction of wordnets by using machine translation and language modeling. In Tomaz Erjavec, J. Z. G. L., editor, *In Proceedings of Seventh Language Technologies Conference, 13th International Multiconference Information Society*, volume C.

Scannell, K. P. (2003). Automatic thesaurus generation for minority languages: an Irish example,. In *TALN Workshop*

on Traitement Automatique des Langues Minoritaires et des Petites Langues, volume 2, pages 203–212, Batz-sur-Mer.

Sornlertlamvanich, V., Charoenporn, T., Robkop, K., and Isahara, H. (2008). KUI: Self-organizing multi-lingual wordnet construction tool. In Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., and Vossen, P., editors, *4th Global Wordnet Conference: GWC-2008*, pages 417–427, Szeged, Hungary.

Thoongsup, S., Charoenporn, T., Robkop, K., Sinthurahat, T., Mokarat, C., Sornlertlamvanich, V., and Isahara, H. (2009). Thai wordnet construction. In *Proceedings of*

The 7th Workshop on Asian Language Resources (ALR7), Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing (IJCNLP),, Suntec, Singapore.

Tufiș, D., Ion, R., Bozianu, L., Ceaușu, A., and Ștefănescu, D. (2008). Romanian wordnet: Current state, new applications and prospects. In *Proceedings of the 4th Global WordNet Association Conference*, pages 441–452, Szeged.

Vossen, P., editor (1998). *Euro WordNet*. Kluwer.

Vossen, P., Maks, I., Segers, R., and Van der Vliet, H. (2008). Integrating lexical units, synsets and ontology in the Cornetto database. In *LREC 2008*, Marrakech, Morocco. European Language Resources Association (ELRA).

Xu, R., Gao, Z., Qu, Y., and Huang, Z. (2008). An integrated approach for automatic construction of bilingual Chinese-English WordNet. In *3rd Asian Semantic Web Conference (ASWC 2008)*, pages 302–341.

Yoon, A., Hwang, S., Lee, E., and Kwon, H.-C. (2009). Construction of Korean wordnet KorLex 1.5. *Journal of KIISE: Software and Applications*, 36(1):92–108.