NTU, NTT Site Reports

Francis Bond and Sanae Fujita^{NTT}

Petter Haugereid, Mathieu Morey, Fan Zhenzhen, Tan Liling

Lea Frermann, Dominkus Wetzel

Division of Linguistics and Multilingual Studies

Nanyang Technological University

^{NTT}**Nippon Telegraph and Telephone Corporation**

<bond@ieee.org>

2012

DELPH-IN

Overview

- NTU
  - Machine Translation (Jaen, SMT)
  - Grammars: Jacy, Norsyg, MCG
  - Cross-lingual parse selection and rephrasing
  - Wordnets: Japanese, English, Chinese, Malay, Multi
  - NTU Multilingual corpus
  - Classifiers
- NTT Report (Sanae Fujita & Takaaki Tanaka)
  Release of GoiTaikei — A Japanese Lexicon (NC) almost
  Joint work with NTU on corpus annotation and WSD

Jaen

- Japanese-English MT system using LOGON transfer
- core of hand-written rules
- open rules (some quite complex) learned from corpora
  - 10 million word J-E parallel corpus
  - learn rules from phrase table based on lemmas
    - ∗ learn from all sentences — high cover
  - learn rules from phrase table based on predicates
    - ∗ learn from parsed sentences (1/3)— high precision

Results

	Parsing	Transfer	Generation	Overall	NEVA	Oracle	F1
Lemm	79.8%	46.6%	56.0%	20.8%	18.65	22.99	19.69
Pred	79.8%	49.7%	52.6%	20.8%	**21.11**	**25.75**	20.96
All	79.8%	60.9%	54.7%	**26.5%**	19.77	24.00	**22.66**

Table 1: Evaluation of the Tanaka Corpus Test Data

	BLEU	METEOR	HUMAN
JaEn (All)	16.77	28.02	**58**
MOSES	**30.19**	**31.98**	42

Table 2: Comparison of Jaen and MOSES (1194 items)

Rule extraction machinery is being prepared for release

Translation examples

(1)	Source:	我々 は 魚 を 生 で 食べる 。
	Ref.:	We eat fish raw.
	Moses:	We eat fish raw.
	Jaen:	We eat fish in the camcorder.
(2)	Source:	カーテン が ゆっくり 引か れ た 。
	Ref.:	The curtains were drawn slowly.
	Moses:	The curtain was slowly.
	Jaen:	The curtain was drawn slowly.
(3)	Source:	偏見 は 持つ べき で は ない 。
	Ref.:	We shouldn't have any prejudice.
	Moses:	You should have a bias.
	Jaen:	I shouldn't have prejudice.

Moses loses the negation 2/3 of the time!

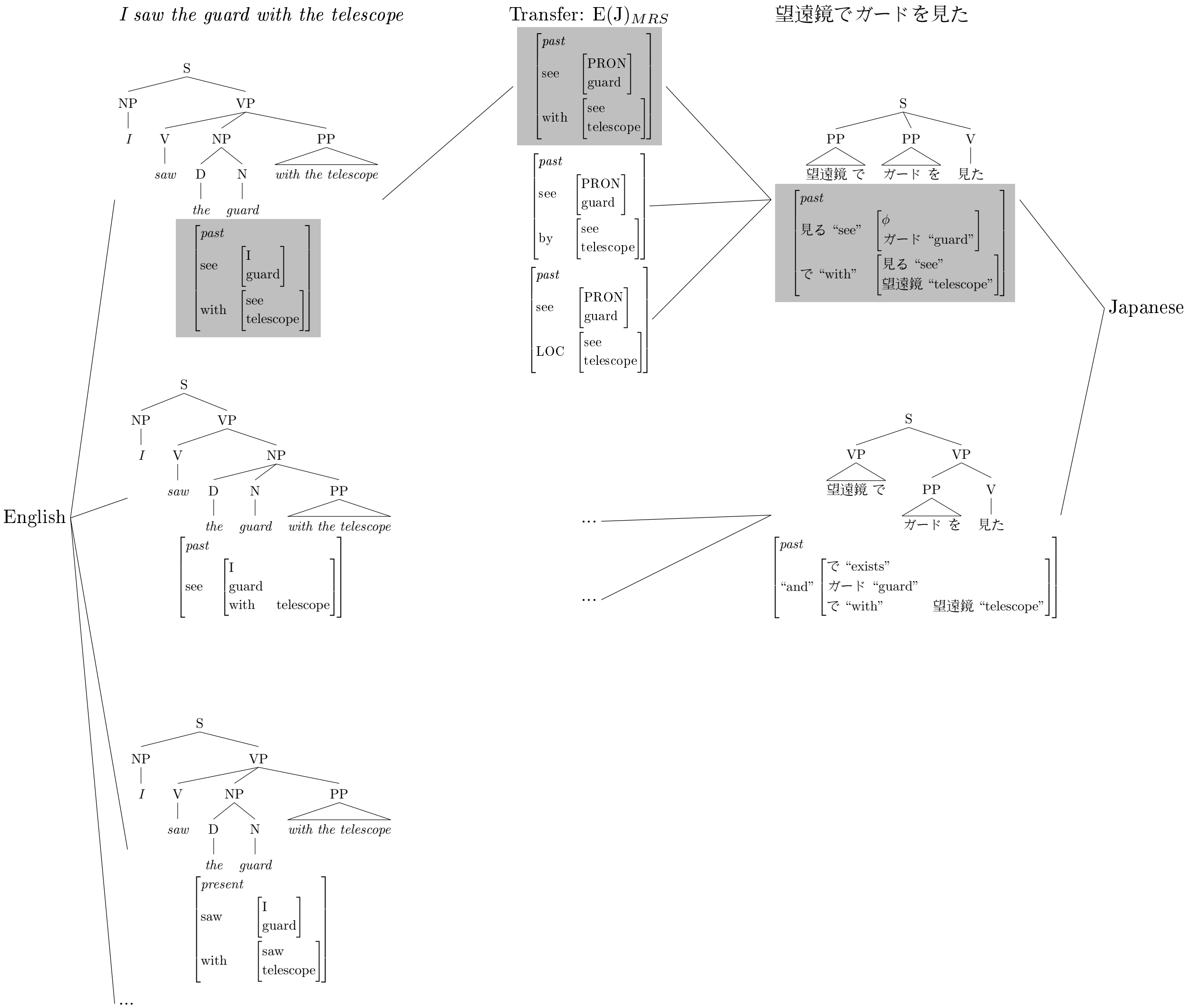
Improve by making negative training data by rephrasing (+3.24 BLEU)

NANYANG TECHNOLOGICAL UNIVERSITY

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_{MRS_i} (meaning)
    - ∗ translate J_{MRS_i} to E(J)_{MRS_j}
    - ∗ parse English to E_{MRS_k}
    - ∗ best parse(s) $= \arg\max_{(i,k)}(\text{sim}(\text{E(J)}_{MRS_k},\text{E}_{MRS_i}))$

Matching Semantics



Cross Lingual Disambiguation

Generally about 3 sentences have the same similarity: reduced ambiguity to 30% (11 →3). We can do better.

	English		**Japanese**	
	Prec	**F**	**Prec**	**F**
First Rank	0.659	0.791	0.676	0.803
Included	0.820	0.897	0.804	0.887

for the 71% of sentences that parse and partially translate

MRS comparison as graph matching

- MRSs are (directed acyclic) graphs
  ⇒*inexact graph matching problem*

- Differences between MRSs can be formulated in terms of graph edit operations, with associated costs:
  - insertion/deletion of EPs,
  - insertion/deletion of ARG links,
  - substitution of a relation following the type hierarchy. . .

- Transfer rules then correspond to sequences of graph edit operations.

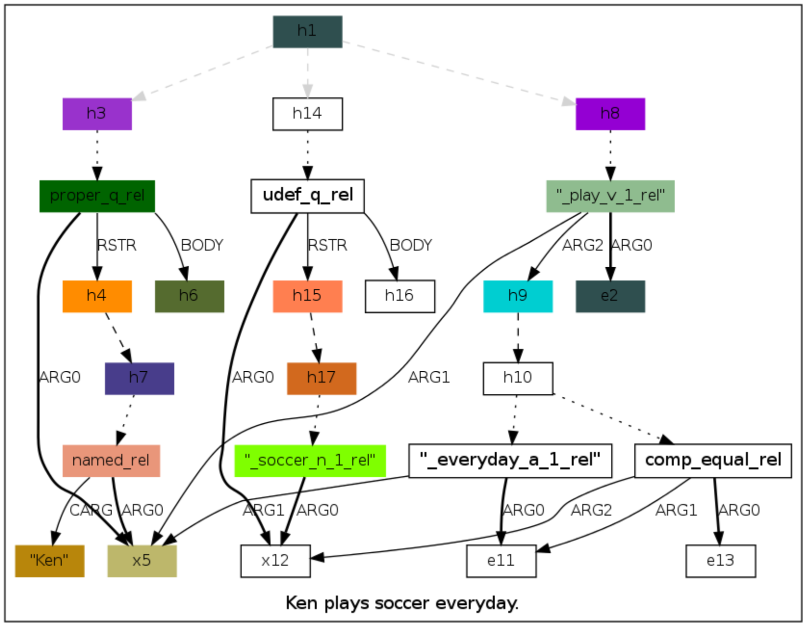
MRS comparison as graph matching

- Pros
  - graph matching is more robust and flexible than comparing n-grams of Elementary Dependencies,
  - graph edit operations directly describe transfer rules,
- Cons:
  - finding optimally interesting/useful edit costs is not trivial,
  - automatically partitioning the set of edit operations (between two big MRSs) into linguistically meaningful transfer rules is tricky
    (guide with patterns e.g. N+ADJ → N+N)

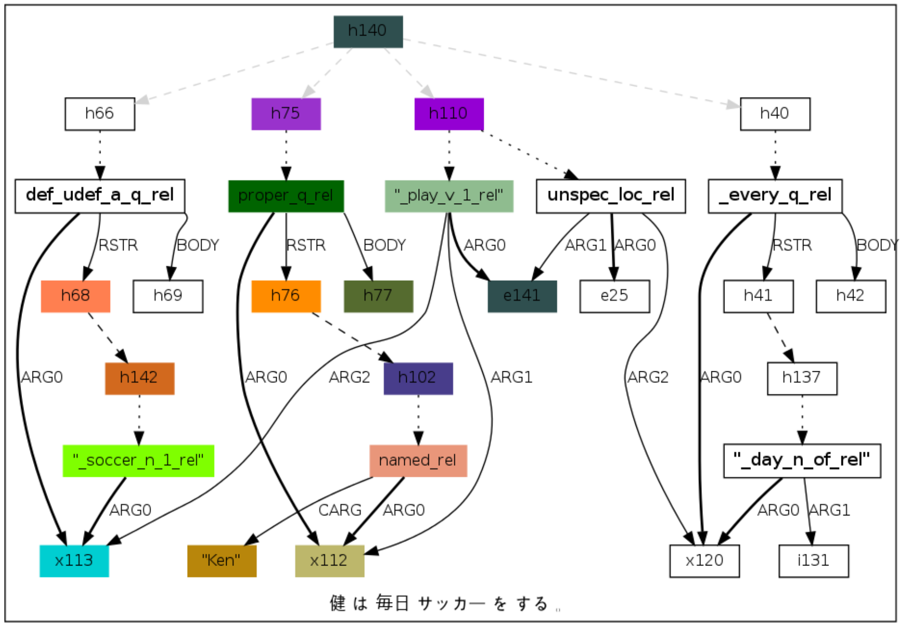
Current state

- Implementation in python
  - matching; visualisation; graph persistency
- Todo:
  - experiment with edit costs; share code; integrate

Buggy but colorful



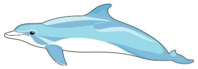
Ken plays soccer everyday.



健 は 毎日 サッカー を する 。

We have already found several bugs in Jaen.

Norsyg – Norwegian Grammar



1. Produces conventional MRS representations

2. Uses the NorKompLeks lexicon (73,000 lexical entries)

   ➢ Freely distributable (MIT licence)

   ➢ The grammar uses the REPP preprocessor

   ➢ Coverage of $\approx 30\%$ on the LOGON Jotunheimen data

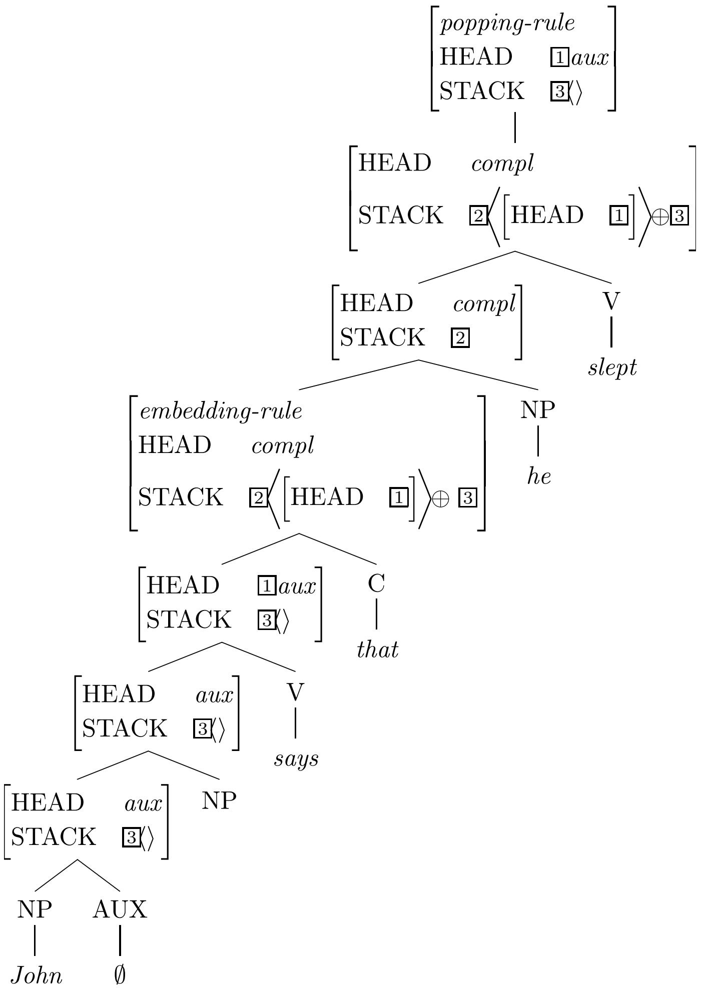
3. The grammar has been made strictly left-branching

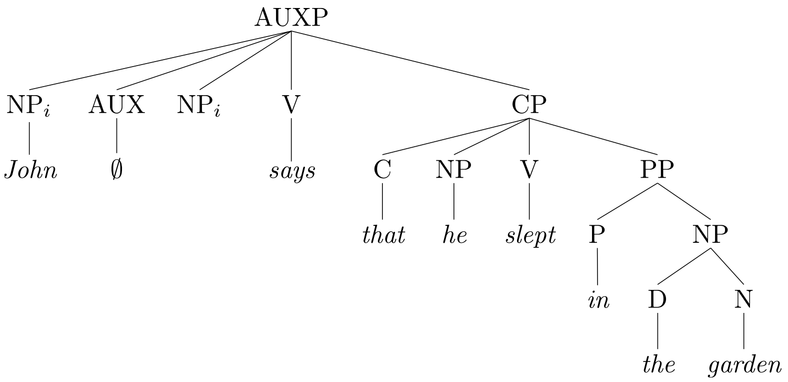
   $\Rightarrow$ all rules are of the form Phrase $\Rightarrow$ Word/Phrase, (Word)

   $\Rightarrow$ compatible with incremental parsing

   ➢ makes use of a STACK feature to account for constituent structure

Stack and Constituent Structure


popping-rule
HEAD $\boxed{1}$aux
STACK $\boxed{3}\langle\rangle$
HEAD compl
STACK $\boxed{2}\langle$[HEAD $\boxed{1}$]$\rangle \oplus \boxed{3}$
HEAD compl
STACK $\boxed{2}$
V
slept
embedding-rule
HEAD compl
STACK $\boxed{2}\langle$[HEAD $\boxed{1}$]$\rangle \oplus \boxed{3}$
NP
he
HEAD $\boxed{1}$aux
STACK $\boxed{3}\langle\rangle$
C
that
HEAD aux
STACK $\boxed{3}\langle\rangle$
V
says
HEAD aux
STACK $\boxed{3}\langle\rangle$
NP
NP
AUX
John
∅



WordNets

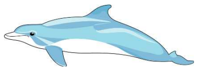
- **Japanese Wordnet**: variants, corpus, taboo words
- **Chinese Wordnet**: many new words, corpus
- **English Wordnet**: new entries, corpus
- **Wordnet Bahasa**: 50k synsets, 120k senses, corpus ☺
  - In cooperation with Malay and Indonesian projects
- **Open Multilingual Wordnet**: combining open resources
  arb, eng, fas, fin, fre, heb, ind, jpn, tha, zsm

Mainly me and my undergraduates now

NTU multilingual corpus

- Small, deeply analysed corpus
  - 6,000 sentences x 3 languages (cmn, eng, jpn)
    - Mainichi Newspaper (NICT translations)
    - Sherlock Holmes
    - Cathedral and the Bazaar (plus many languages)
    - Singapore Tourist data (plus Korean, Viet, Indo)
  - Hand alignment, WordNet tagging, Treebanking
- Plus a lot more Japanese-English (and some Chinese)

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

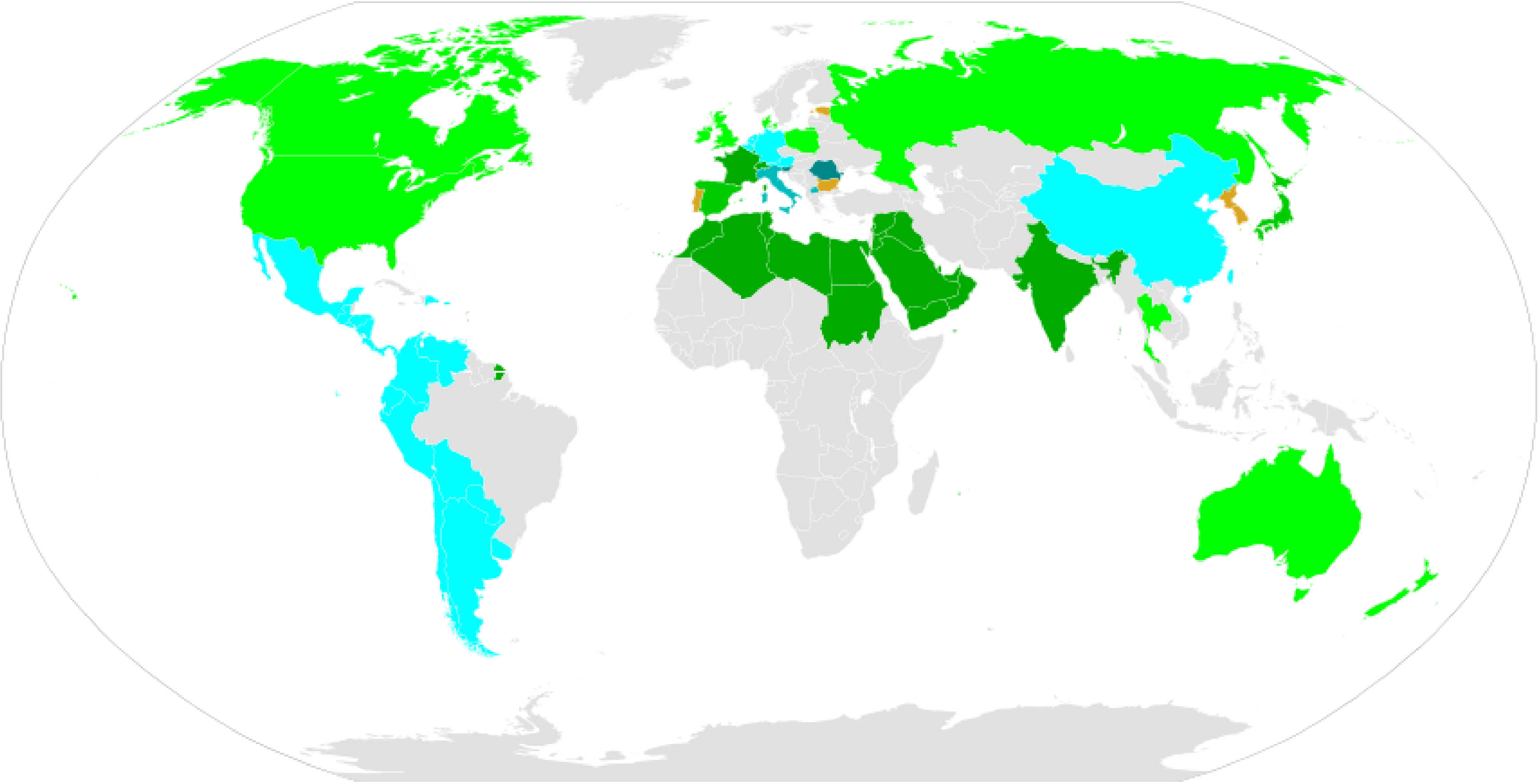
Current State (last week)

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
	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) (10 →20 this year)

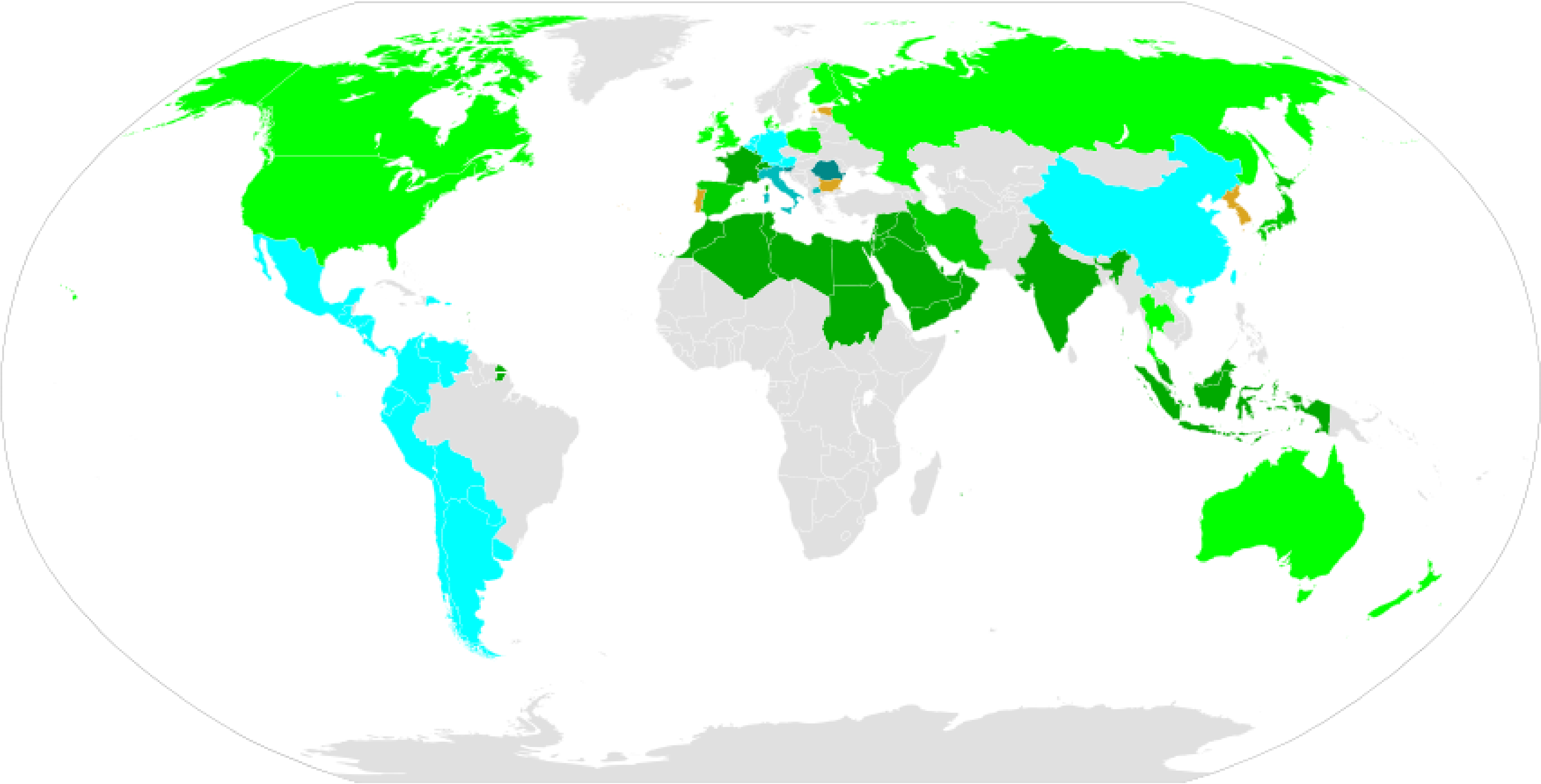
Ask me about these

Wordnets in the world 2011-06



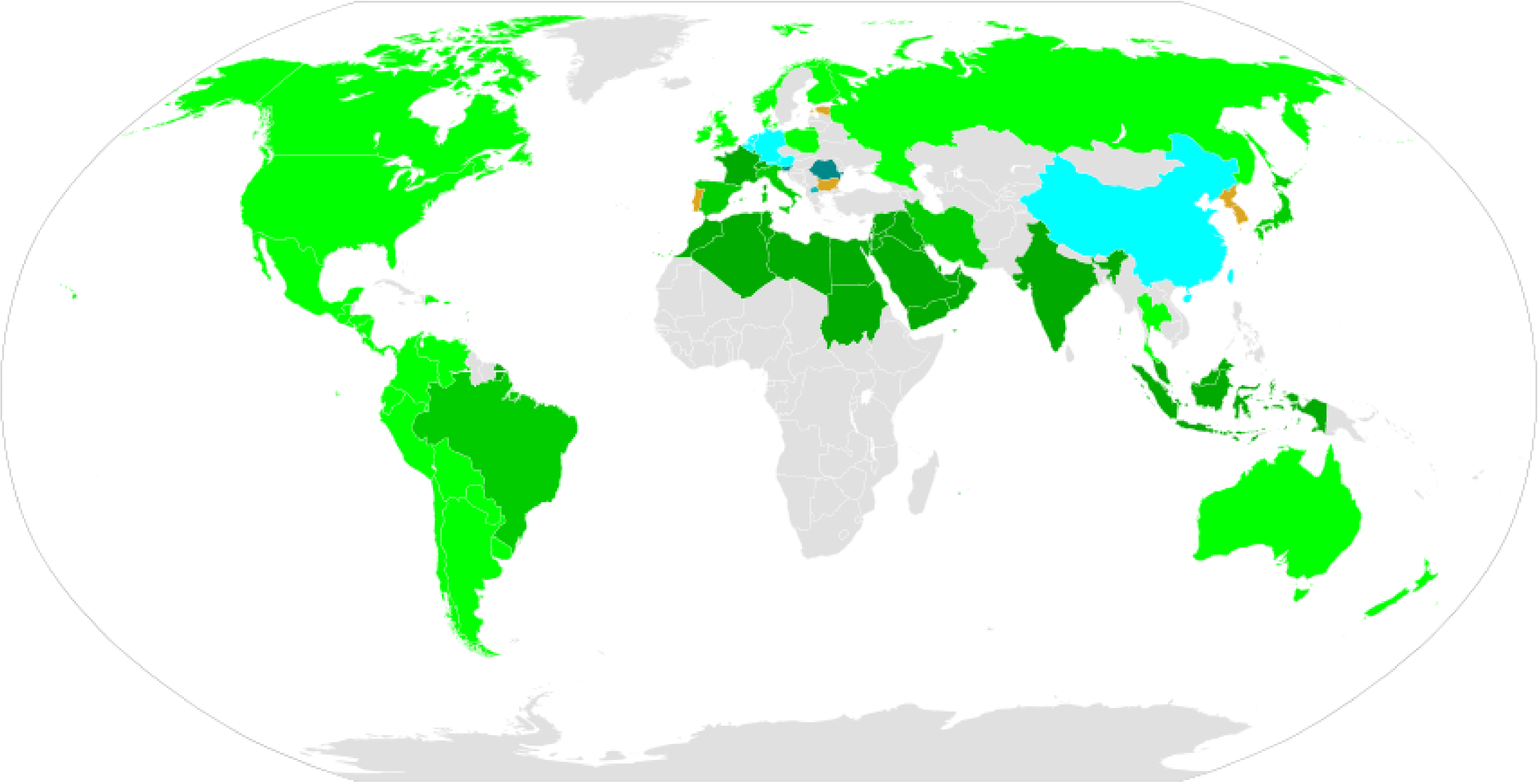
Green is free; Blue is research only; Brown costs money

Wordnets in the world 2012-01



Added: Finnish, Persian, Bahasa

Wordnets in the world 2012-06

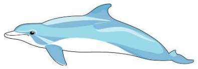


Added: Norwegian; Freed: Italian, Portuguese, Spanish

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

NANYANG TECHNOLOGICAL UNIVERSITY

What is lacking?



- German, Chinese, Bulgarian, . . . ☹
- 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)
- ToDo: Setting up shared multilingual index

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.

Synset: 07229245-n

thank you

kiitos

merci

je vous remercie

ありがとう

サンキュー

terima kasih

terima kasih

agradecimento

I can't print the Thai