NTU, NTT Site Reports

Francis Bond and Sanae Fujita[NTT]
Wang Shan, Mathieu Morey, Fan Zhenzhen,
Tan Liling, Le Tuan Anh,[NUS] Xiaocheng Yin,
Michael Goodman,[UW] Zinaida Pozen,[UW] Takaaki Tanaka[NTT]
Nanyang Technological University
[NTT]**Nippon Telegraph and Telephone Corporation**

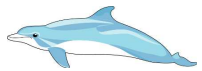
<bond@ieee.org>

2013

DELPH-IN

NANYANG TECHNOLOGICAL UNIVERSITY

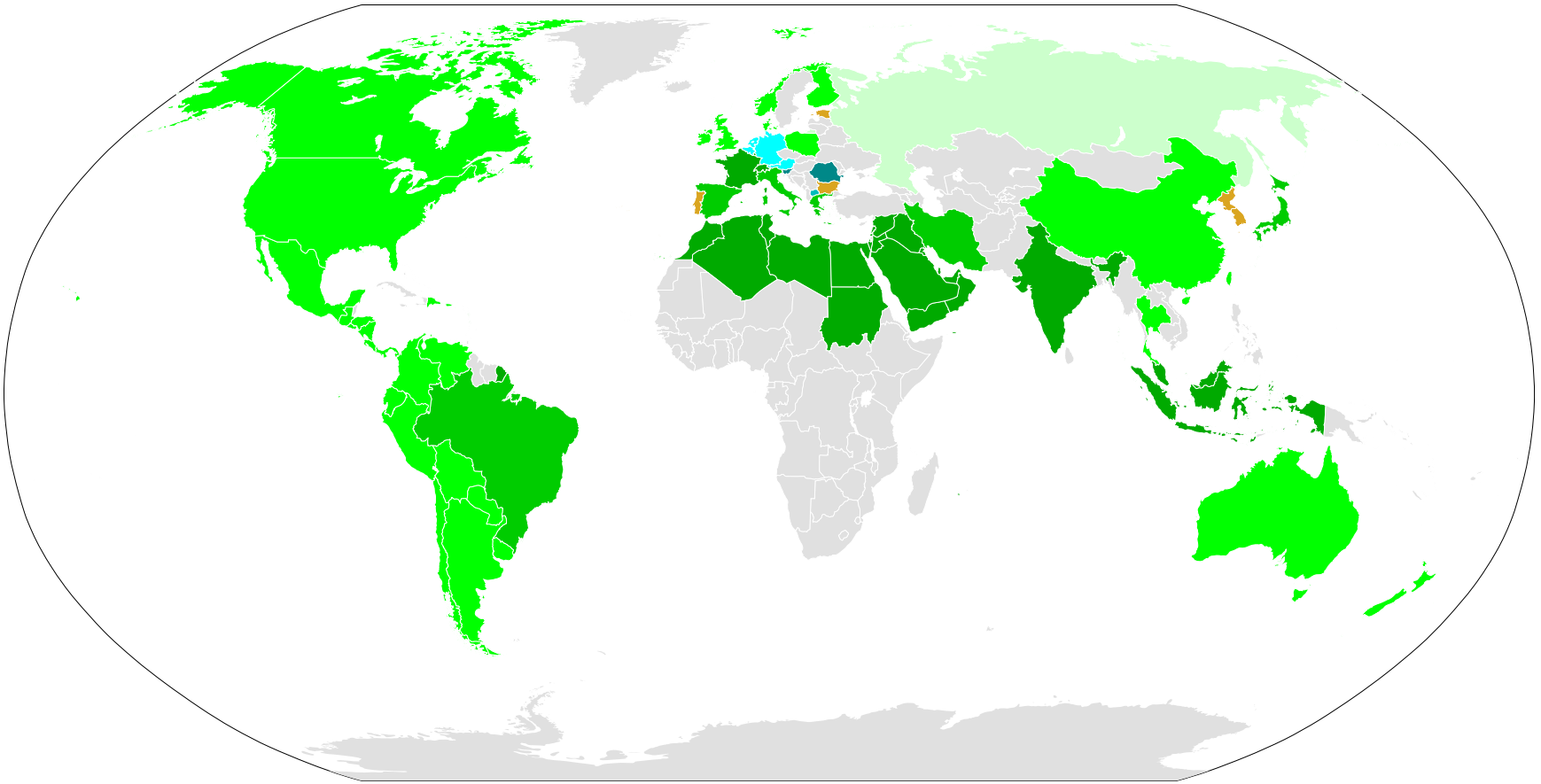
Overview



- NTU
  - Wordnets: Multi, Japanese, English, Chinese, Malay
  - Cross-lingual concept alignment
  - Grammars: MCG:Lexicon, ERG:Possession, Jacy
  - Parse Selection with Wordnets
  - DMRS visualization
  - Machine Translation (Jaen, SMT)
- NTT Report (Sanae Fujita & Takaaki Tanaka)
  Joint Research with NTU on Ontology linking, Subcat acquisition

NANYANG TECHNOLOGICAL UNIVERSITY

Wordnets in the world 2013-06



Green is free; Blue is research only; Brown costs money: Afrikaans, Dutch, Greek, Hindi to come

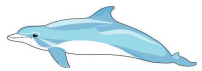
NANYANG TECHNOLOGICAL UNIVERSITY

But there are many languages

- Linked wordnet to wiktionary (accuracy 93.7%)
  - 26 wordnets with more than 10,000 synsets
  - 57 with 1,000–10,000 synsets
- Good coverage of frequently used concepts
- Automatic wordnets not as good as hand built
  but better then nothing
- Seeding new projects: Vietnamese, Singhalese, Russian?

NANYANG TECHNOLOGICAL UNIVERSITY

Wordnets ($\geq$ 10,000 entries)



	Projects			Wiktionary			Merged (+CLDR)		
Language	**Synsets**	**Senses**	**Core**	**Synsets**	**Senses**	**Core**	**Synsets**	**Senses**	**Core**
English	117,659	206,978	100	35,400	49,951	75	117,661	213,538	100
Finnish	116,763	189,227	100	21,516	31,154	65	116,830	199,435	100
Thai	73,350	95,517	81	2,560	3,193	17	73,595	97,390	81
French	59,091	102,671	92	20,449	27,150	63	61,258	109,643	95
Japanese	57,179	158,064	95	12,685	19,479	52	59,112	166,617	96
Indonesian	52,006	142,488	99	2,390	2,810	17	52,154	143,755	99
Catalan	45,826	70,622	81	8,626	10,251	36	48,007	74,806	84
Spanish	38,512	57,764	76	18,281	25,310	60	47,737	74,848	86
Portuguese	41,810	68,285	79	12,331	16,178	53	43,870	74,151	84
Malay	42,766	119,152	99	2,833	3,744	19	43,079	120,686	99
Italian	34,728	60,561	83	14,605	18,710	53	38,938	68,827	87
Basque	29,413	48,934	71	1,693	1,943	11	29,965	49,945	72
Polish	14,008	21,001	30	10,888	13,431	46	20,975	30,943	55
Galician	19,312	27,138	36	2,492	2,871	15	20,772	29,136	42
Persian	17,759	30,461	41	4,229	5,443	26	20,766	35,318	55

	Projects			Wiktionary			Merged (+CLDR)		
Language	**Synsets**	**Senses**	**Core**	**Synsets**	**Senses**	**Core**	**Synsets**	**Senses**	**Core**
Russian	0	0	0	19,983	33,716	64	20,138	34,009	64
German	0	0	0	19,675	29,616	64	19,857	29,884	64
Chinese	4,913	8,069	28	12,130	19,079	49	15,490	27,113	60
Arabic	10,165	21,751	48	6,892	9,337	38	14,861	31,337	63
Dutch	0	0	0	13,741	19,709	56	13,950	20,003	56
Czech	0	0	0	12,802	15,493	54	13,030	15,813	54
Swedish	0	0	0	12,000	16,226	51	12,221	16,512	51
Greek	0	0	0	10,308	13,071	44	10,549	13,472	44
Danish	4,476	5,859	81	7,290	8,931	35	10,328	13,551	85
Bokmål	4,455	5,586	79	7,262	9,170	35	10,322	13,612	83
Hungarian	0	0	0	9,964	12,699	45	10,213	13,029	45

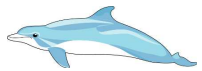
NANYANG TECHNOLOGICAL UNIVERSITY

Shared Interlingual-Index

- Interlingual-Index is a shared list of all concepts
- English definition and link in one open-source wordnet
- Allow trusted users/teams to modify/extend
  - Social issues inseparable from technical issues
- Keep persistent identifiers
- Possibly allow compositional semantics (ontological links)
  $next\ year$ = compound($next$, $year$)
- For long term maintenance, it makes more sense to link wordnets, than combine into one multilingual resource languages are interestingly different

NANYANG TECHNOLOGICAL UNIVERSITY

NTU multilingual corpus



- Small, deeply analysed corpus
  - 6,000 sentences x 3 languages (cmn, eng, jpn); 2,000 in Indonesian
    - Mainichi Newspaper (NICT translations)
    - Sherlock Holmes (many languages)
    - Cathedral and the Bazaar (many languages)
    - Singapore Tourist data (plus Korean, Viet, Indo)
  - Hand alignment, WordNet tagging, Treebanking, Information Structure, . . .
- Plus a lot more Japanese-English (and some Chinese)

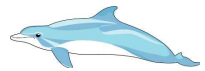
Example (from News Corpus)

(1) Jpn: 大臣$_1$ が 離党$_2$ した
daijin ga ritou shita
minister SBJ leave-party did

(2) Eng: *The minister*$_4$ *left*$_8$ *the party*$_1$

(3) Cmn: 官员$_1$ 离开$_3$ 了 政党$_1$
guanyuan likai le zhengdang
minister leave already political-party

Not linkable with our current model



(4) I am sure that I shall say nothing of the kind.

a. いやいや 、 そんな こと は
iyaiya , sonna koto wa
by+no+means , that+kind+of thing TOP
言わ-ん よ
iwa-n yo
say-NEG yo
"no no, I will not say that kind of things"

➢ ***sonna*** not in wordnet & negation makes it hard to link

➢ *iyaiya* $\leftrightarrow$ *I am sure that I shall* ???

Grammars

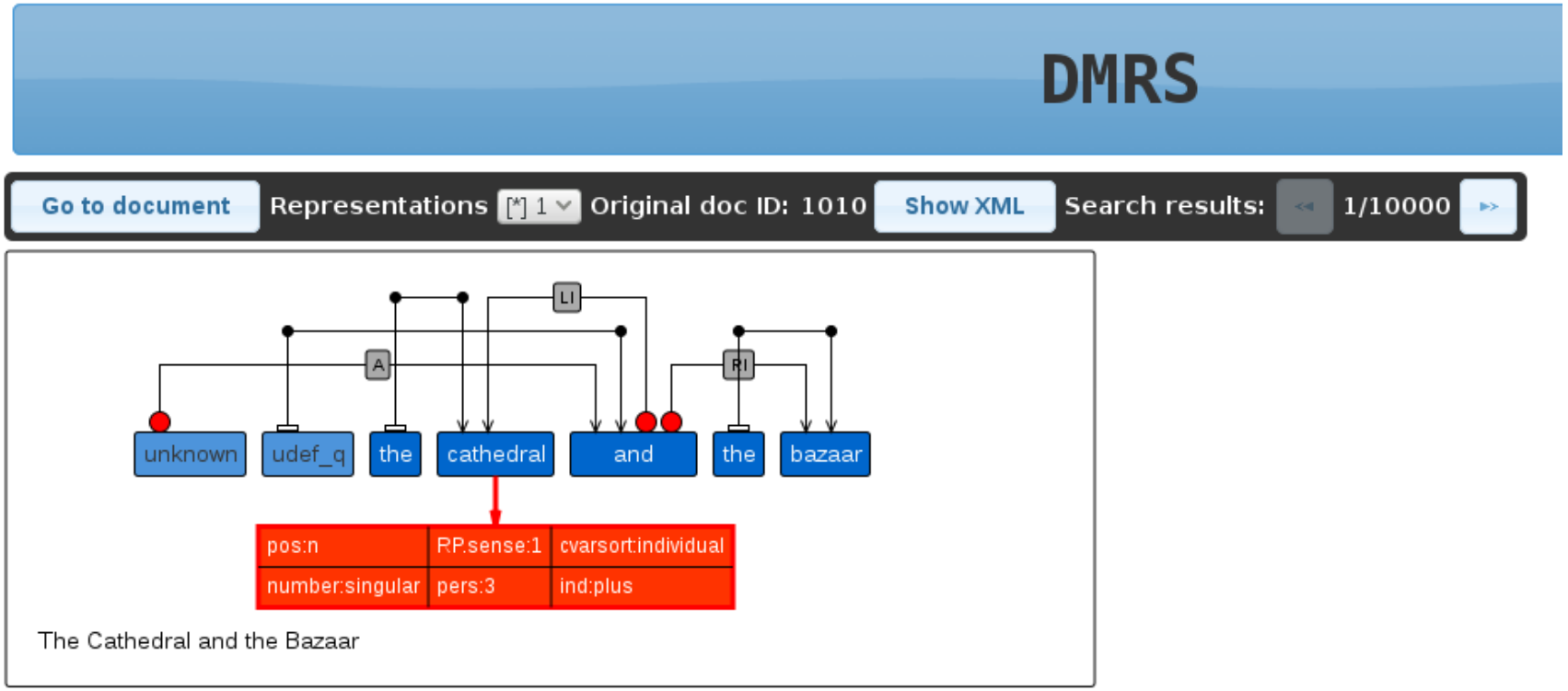
- Mandarin Chinese Grammar (Fan Zhenzhen)
  - Learning verbs from treebanks
  - Predicate name normalization
- English Resource Grammar
  - *X look as though butter wouldn't melt in X's mouth*
  - Around 300 possessive idioms
- Jacy
  - Some bug fixes

Parse Ranking with Lexical Semantics

- Backing Off to supersenses
- Some Feature Engineering
  - Xiaocheng Yin
  - Zinaida Pozen

NANYANG TECHNOLOGICAL UNIVERSITY

DMRS visualization



Machine Translation (Jaen, SMT)

- New visitor from September: cross-lingual parse ranking

NTT

- Subcat acquisition for Japanese
- Images for lexeed/wordnet