The Japanese Wordnet and the Open Multilingual Wordnet

Francis **Bond**

Department of Asian Studies,
Palacký University, Olomouc, Czechia

<bond@ieee.org>

University of Ljubljana, April 2026

Faculty of Arts

Outline

Faculty of Arts

Roadmap

Faculty of Arts

Self Introduction I

- BA in Japanese and Mathematics
- BEng in Power and Control
- PhD in English on *Determiners and Number in English contrasted with Japanese, as exemplified in Machine Translation*
- 1991-2006 NTT (Nippon Telegraph and Telephone)
  - Japanese - English/Malay Machine Translation
  - Japanese corpus, HPSG grammar and ontology (Hinoki)
- 2006-2009 NICT (National Inst. for Info. and Comm. Technology)
  - Japanese - English/Chinese Machine Translation
  - Japanese WordNet
  - Open Multilingual Wordnet

Faculty of Arts

Self Introduction II

- 2009-2022 NTU (Nanyang Technological University)
  - Abui, Chinese, Malay, Multilingual Wordnets (OMW)
  - HPSGs for Chinese, Indonesian, …
  - Multilingual Meaning Banks (Treebank + Sensebank)
- 2022- UPOL (Palacký University, Olomouc)
  - Czech, Cantonese and Multilingual Wordnets
  - ChainNet for metaphor and metonymy
  - WSD using LLMs
  - Codex of HPSG implemented grammars
  - Using HPSG to measure LLM syntactic diversity

Faculty of Arts

Roadmap

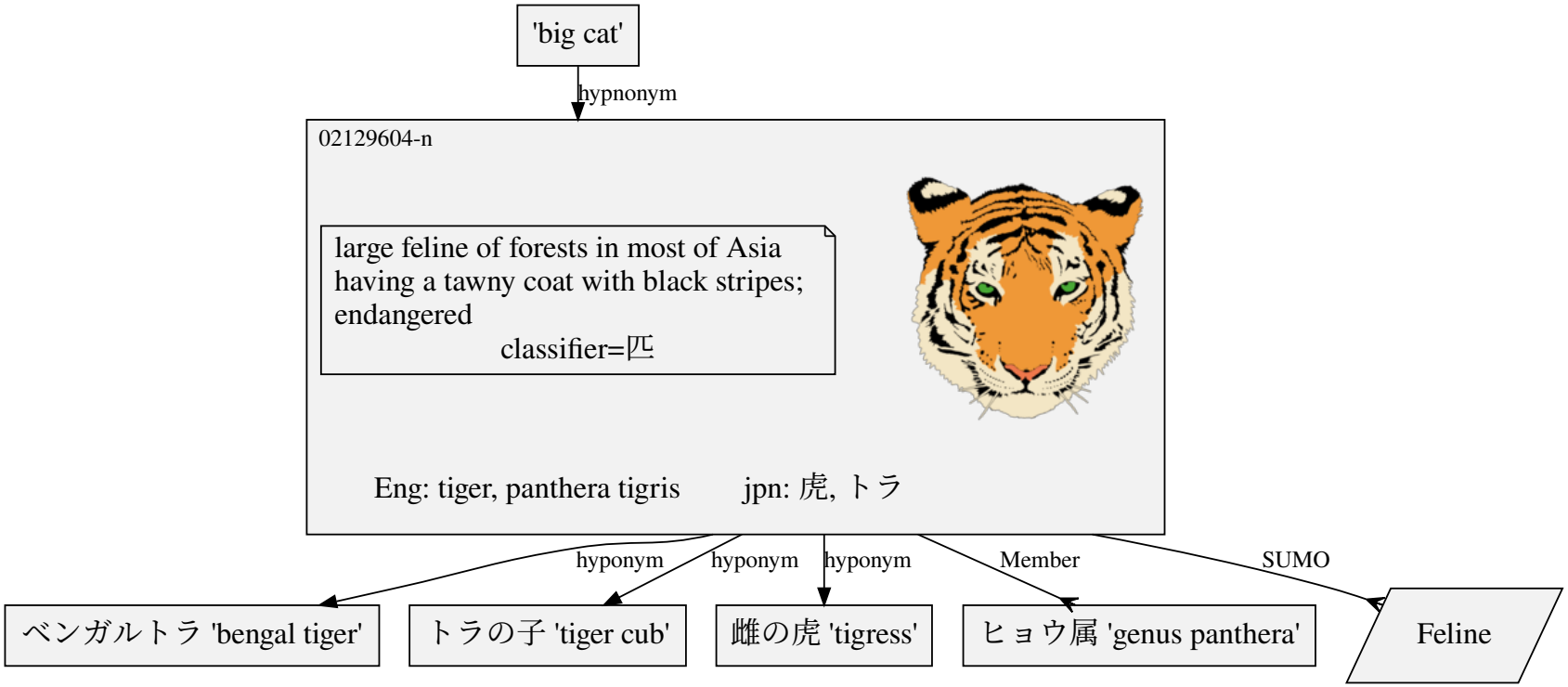
Faculty of Arts

What is a Wordnet?

- ***WordNet*** is an open-source electronic lexical database of English, originally developed at Princeton University
  http://wordnet.princeton.edu/
- Made up of four separate (but interlinked) semantic nets, for nouns, verbs, adjectives and adverbs
- A ***wordnet*** is a lexicon built with a similar structure to English

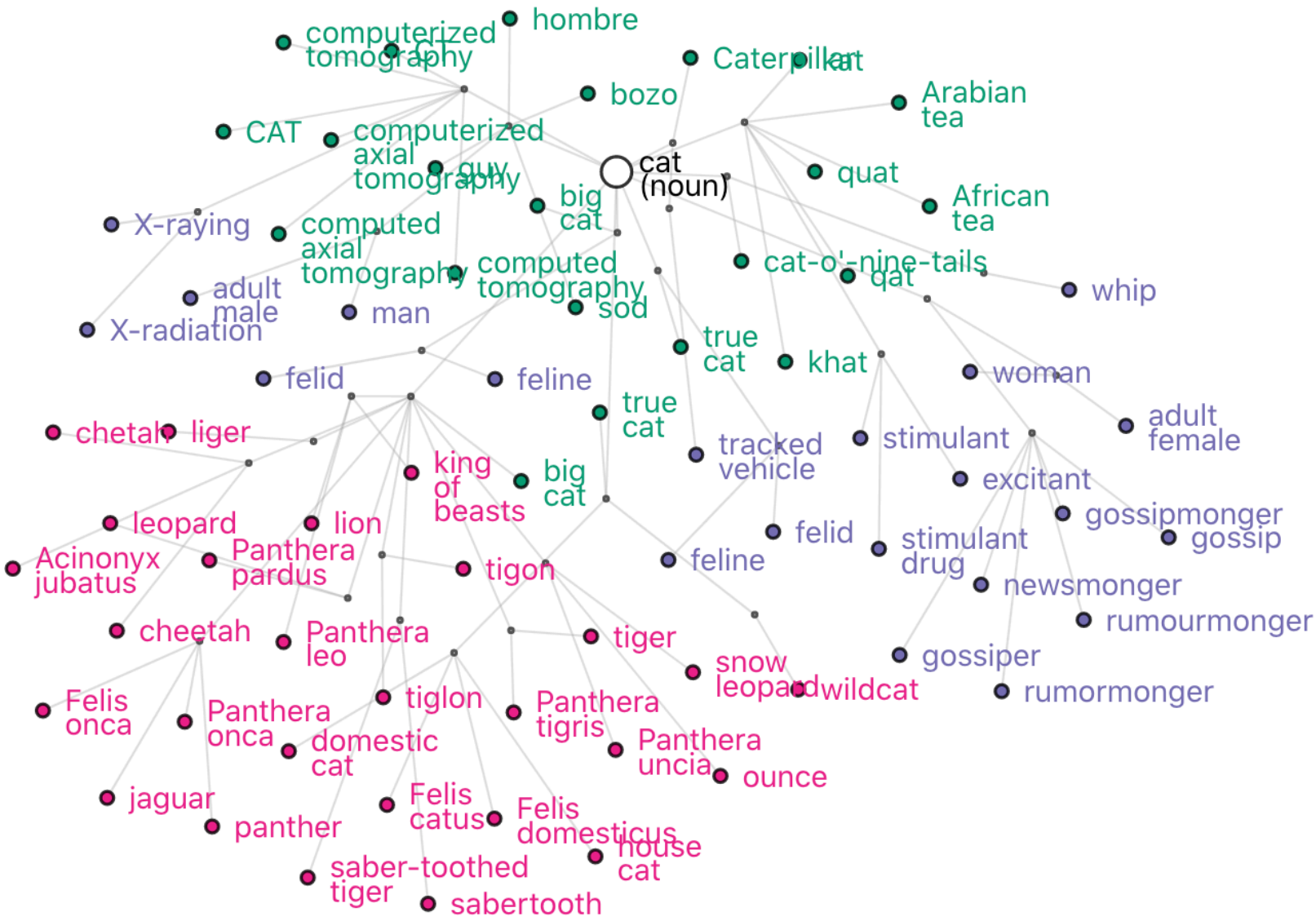
Faculty of Arts

The synset for 虎 “tiger”



Here we show English and Japanese.

Another view of *cat* as a graph



From https://github.com/aliiae/lexical-graph

Many people wanted to have wordnets

- **EuroWordNet** (Vossen, 1998): Dutch, English, German, French, Spanish, Italian
- **BalkaNet**: Bulgarian, Greek, Romanian, Serbian, Turkish
- **Asian WordNet**: Japanese, Korean, Thai, Indonesian, Vietnamese, Mongolian, Burmese
- **IndoWordNet**: Hindi, Bengali, Marathi, Gujarati, Punjabi, Urdu, Tamil, Telugu, Kannada, Malayalam, Odia, Assamese, Nepali, Konkani, Manipuri, Kashmiri, Sanskrit

And many individual projects not listed here.

NICT decided to build a wordnet for Japanese

- Stage 0
  - Semi-automatically translate English WordNet (3.0)
- Stage 1
  - Manually correct the top 10,000 entries
  - This includes the 5,000 **core** synsets
- Stage 2 (in progress: poster yesterday)
  - Correct the next most frequent 15,000 entries
  - Create a Japanese version of SemCor
  - Release WordNet-ja v1.0

Faculty of Arts

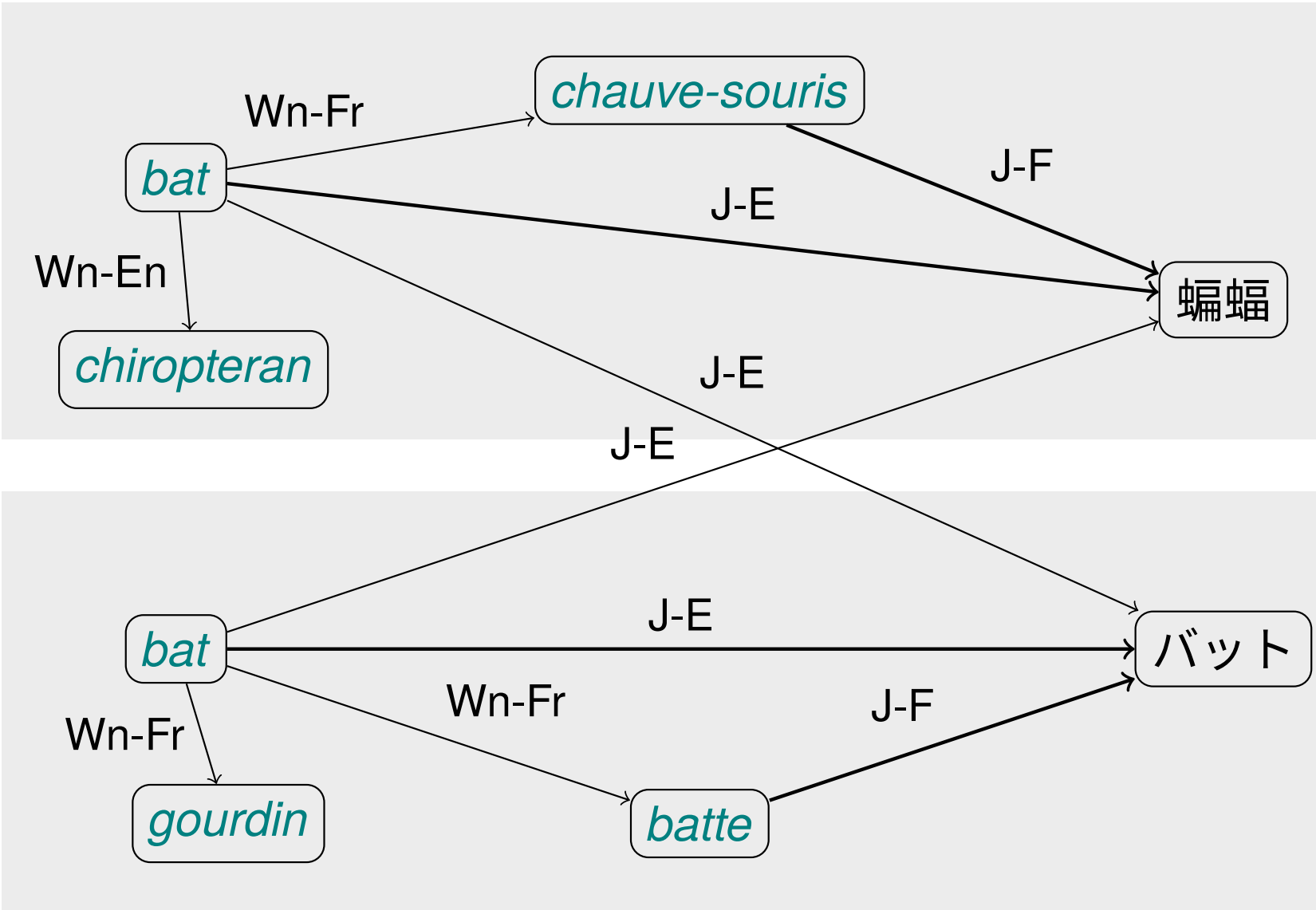
Exploit other wordnets

- For each synset in WordNet 3.0
  - Find its equivalents in WN-Fr, WN-Es, Wn-De
  - Look up translations for all equivalents $\{J_e\}$, $\{J_f\}$, $\{J_s\}$, $\{J_d\}$
  - Rank Japanese equivalents
    score $s = |\text{links}| + 10$ for links in two languages

The result is a WordNet with multiple Japanese candidates for each synset ranked by score.

Faculty of Arts

Linking with Multiple WordNets



The Japanese Wordnet

- Initially assume that the semantic structure is the same
  - ***dog*** ⊂ *animal*
  - ⇒ **犬** ⊂ **動物**
- Added Japanese words to Princeton wordnet synsets

Date	Ver	Concepts	Words	Senses	Misc
2009-02-28	0.90	49,190	75,966	156,684	initial release
2009-08-31	0.91	50,739	88,146	151,831	linked to SUMO
2009-11-16	0.91	49,655	87,133	146,811	
2010-03-05	1.00	56,741	92,241	157,398	+ definitions, exam
2010-10-22	1.10	57,238	93,834	158,058	
2012-01-06					Japanese Semcor
2014-02-06					NLTK module
2022-05-22	2.0	58,039	89,902	147,207	386,222 variants

Faculty of Arts

Japanese Wordnet 2.0

- We kept expanding the wordnet and adding more information, but it still did not describe the conceptual structure of Japanese.
- So we started to make it describe Japanese better
  - Taking advantage of a new format (GWA XML 1.0)
    - Allows for variant forms
    - Structure can be different from PWN (788 new concepts (synsets))
    - Includes frequencies from JSemCor and NTU-MC
  - A long slog by Takayuki Kuribayashi to add the forms

Orthographic Variants

- synsetID=14728724-n (Eng: protein)
  プロテイン, 蛋白質, タンパク, たんぱく質, 蛋白, タンパク 質
  ↓
  蛋白質 (タンパクシツ, たんぱ 質, タンパク 質, たんぱくしつ)
  蛋白 (タンパク, たんぱく)
  プロティン (プロテイン, ぷろてぃん)
- synsetID=02765464-v (Eng: absorb, take in)
  呑みこむ, 呑 込む, 吸引, 吸い 込む, 吸収
  ↓
  吸い込む (スイコム, 吸 込む, 吸いこむ, すいこむ)
  吸収 (キュウシュウ, きゅうしゅう) +vs
  吸引 キュウイン, きゅういん) +vs
  飲み込む (ノミコム, 飲 込む, 呑み 込む, 呑 込む, 呑みこむ,
  のみ 込む, のみこむ

Other extensions

- Added pronouns and demonstratives
  私, この, その, あの, どの
- Added classifiers
  人, 台, 匹, 回
- Added 4-character idioms
  一期一会, 海千山千, 五十歩百歩
- Added corpus-based examples and sense frequency
  対策$_3$, **策**$_3$, **措置**$_2$, **方略, 方策, 術, 打つ手** “step, measure”
- Added exclamatives
  王手, ヨイショ, お早うございます

Faculty of Arts

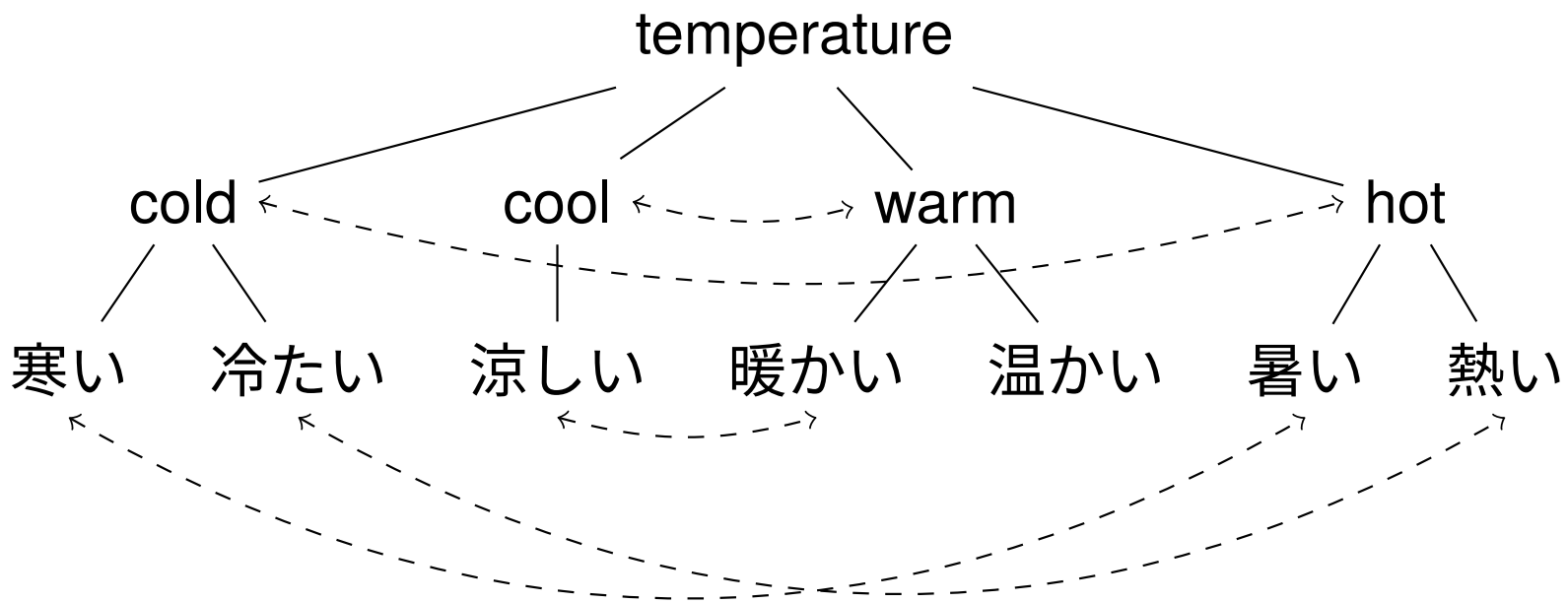
Ontological Differences: Temperature

- English uses the same words for temperature experienced by touching or as a general feeling:
  ⟨***cold, cool, warm, hot***⟩
- Japanese distinguishes
  - the feeling: ⟨**寒い, 涼しい, 温かい, 暑い**⟩
  - to-touch: ⟨**冷い, 暖か, 熱い**⟩
- English uses a single word for water of any temperature: ***water***
- Japanese uses different words for cold (non-hot) water and hot water: **水** vs **湯**
- We are doing our best to add native Japanese concepts, even if not lexicalized in English

It raised an interesting question about interlingal equivalence — should ***water*** link to **水** or do we need a non lexicalized entry **水**-or-**湯**?

Faculty of Arts

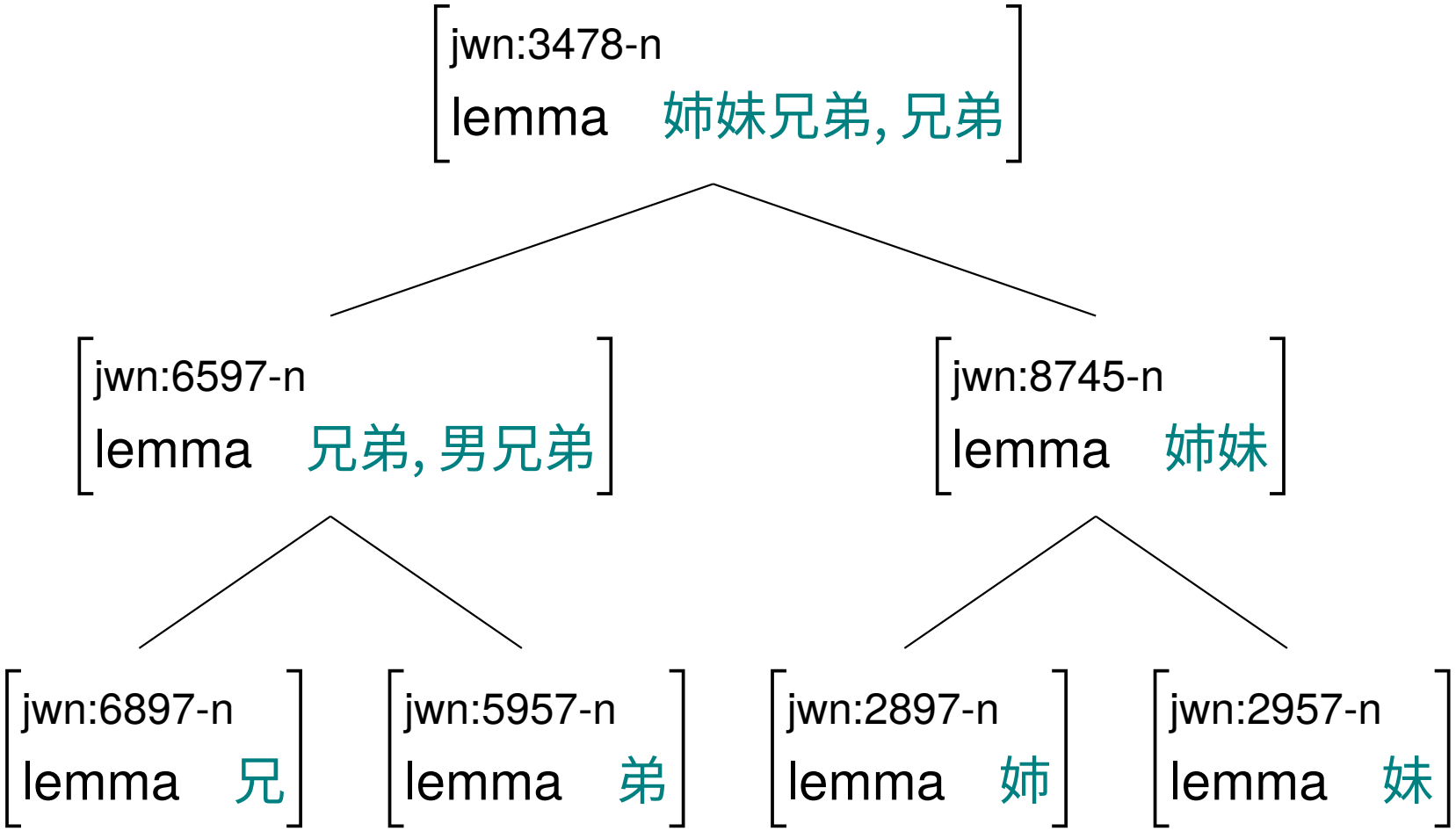
Hierarchy



- Some nodes are not lexicalized in Japanese, but still useful for the structure
- **temperature** is linked by ATTRIBUTE (属性); tree is HYPONYM dashed arrows are ANTONYM

Faculty of Arts

Ontological Differences: Kin



- More accurately models Japanese
- We aim to cover at least the TUFS basic vocabulary

Faculty of Arts

Other Examples — ロングテール “long tail” I

(1)

80001626-n (soba_noodle)	
lemmas:jpn	蕎麦
lemmas:eng	soba
def:jpn	そば 粉で 作られた 細い 麺
def:eng	narrow noodle made from buckwheat
hypernym	(noodle)

(2)

80002377-n (castle construction)	
lemmas:jpn	築城 *chikujou*
def:jpn	城の 建設
def:eng	the construction of castles
hypernym	(construction)

Faculty of Arts

Other Examples — ロングテール “long tail” II

(3)

90000315-n (hajjah)	
lemmas:jpn	ハジャ
lemmas:eng	hajjah
def:jpn	メッカへの 巡礼を 行った 女性
def:eng	a woman who has made the pilgrimage to Mecca
hypernym	(haji)
category	(muslim)

(4)

80001731-n (exchange student)	
lemmas:jpn	留学生
lemmas:eng	exchange student
def:jpn	海外で 勉強する 学生
def:eng	a student who studies abroad

Other Examples — ロングテール “long tail” III

(5)

80000338-n (Shunto)	
lemmas:jpn	春闘
lemmas:eng	spring wage negotiation, spring wage offensive, Shunto
def:jpn	毎年労働組合が、賃金引き 上げなどの 要求を 掲げて 行う 全国的な 闘争
def:eng	annual event by Japanese workers unions when wages are renegotiated
hypernym	(protest)

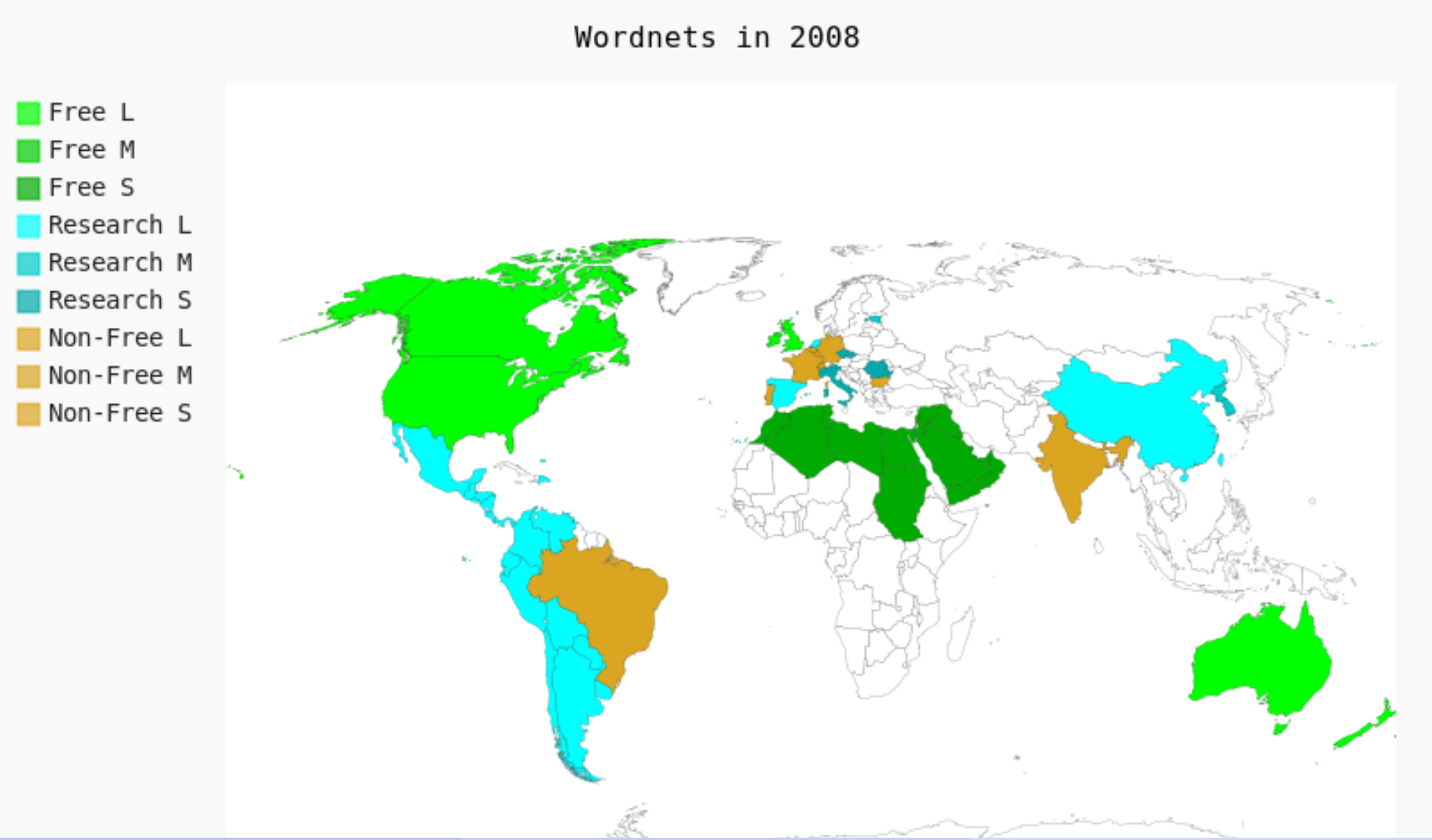
Faculty of Arts

Roadmap

Faculty of Arts

Wordnets in the world 2008-06

Many wordnets, but few free: we wanted to build an open wordnet for Japanese and needed other wordnets for cross-lingual disambiguation.



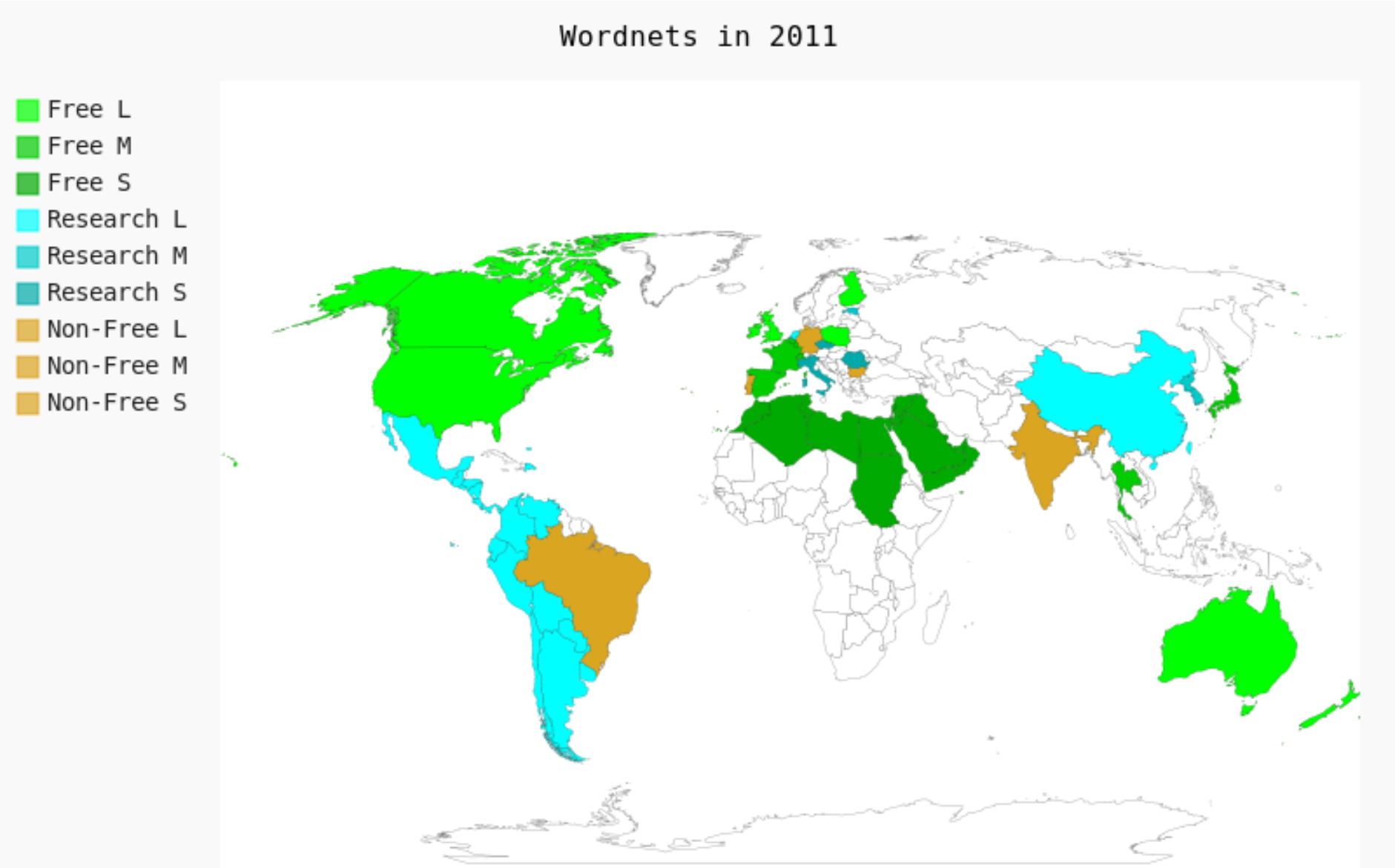
Wordnets in the world 2008-06

- When we decided to build the Japanese wordnet, we wanted to leverage existing work and use multiple languages to disambiguate
- Many wordnets existed, but few were free and almost all came in slightly (or radically) different formats.
  - PWN had a loop!
  - GWG grid had bad encodings
  - There was **no** wordnet that worked out of the box
  - Different projects even used different names for pos
    e.g., Adjective (*a* vs *j*); Adverb (*r* vs *b*)
- There was no easy shared access
- We had to massage the data for our project
- We wanted to try to save other people the trouble we experienced, and make the normalized data available (where legally possible)

Faculty of Arts

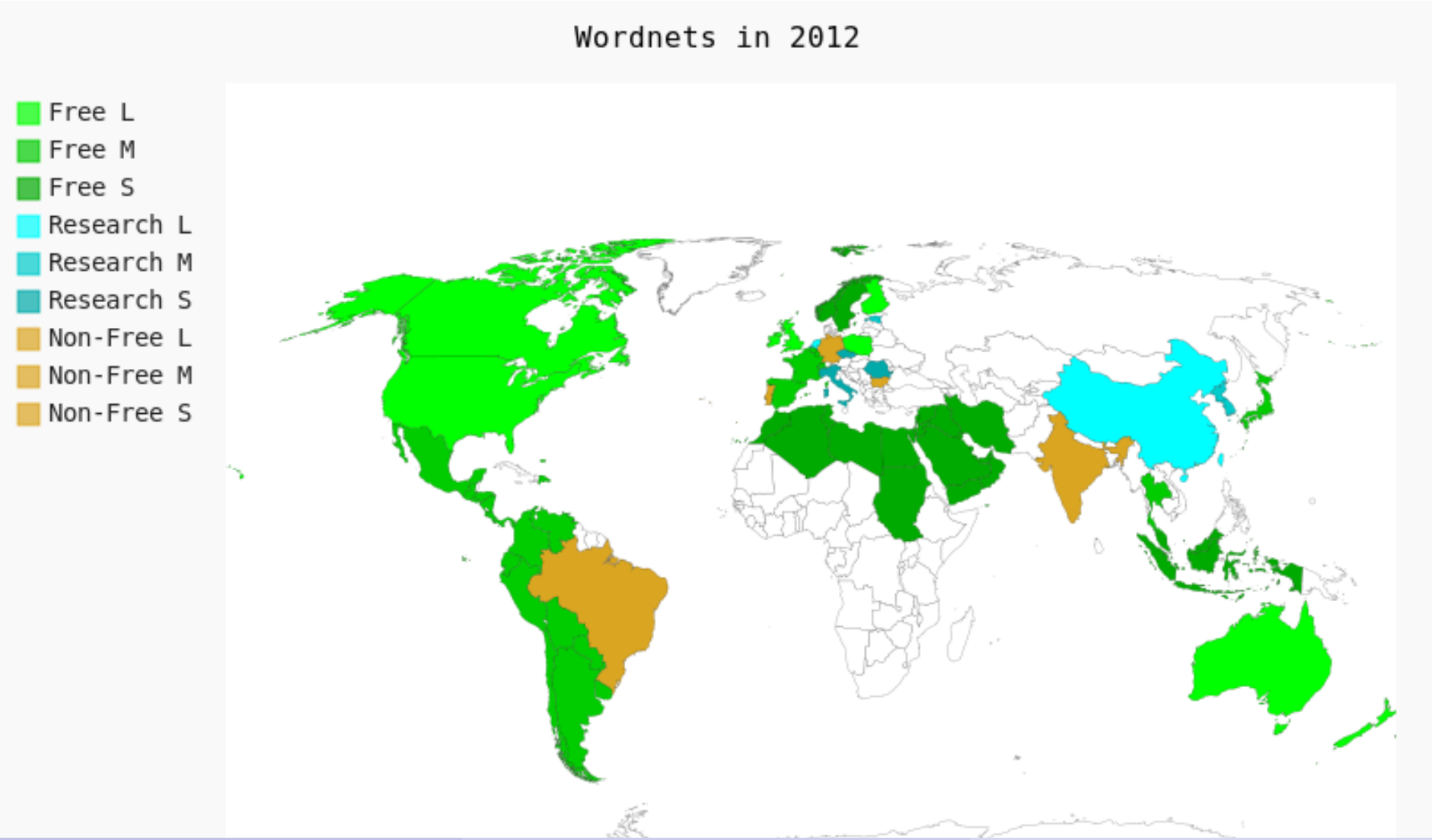
Wordnets in the world 2011-06

Free wordnets for French, Catalan, Polish, Thai and **Japanese**



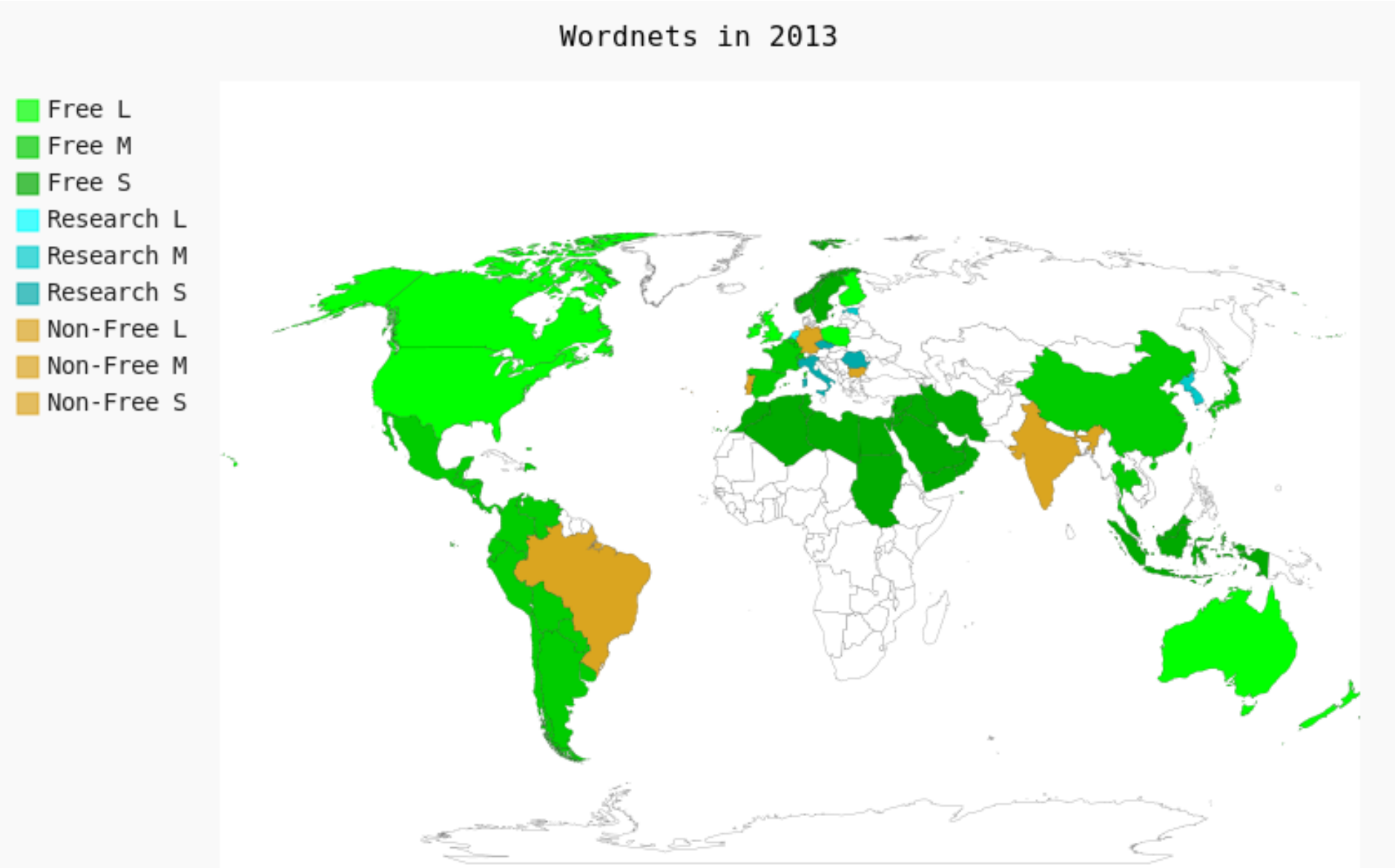
Wordnets in the world 2012-06

Spanish freed (and Galician and Basque), Farsi, Norwegian, Swedish, **Bahasa** (Malay and Indonesian)



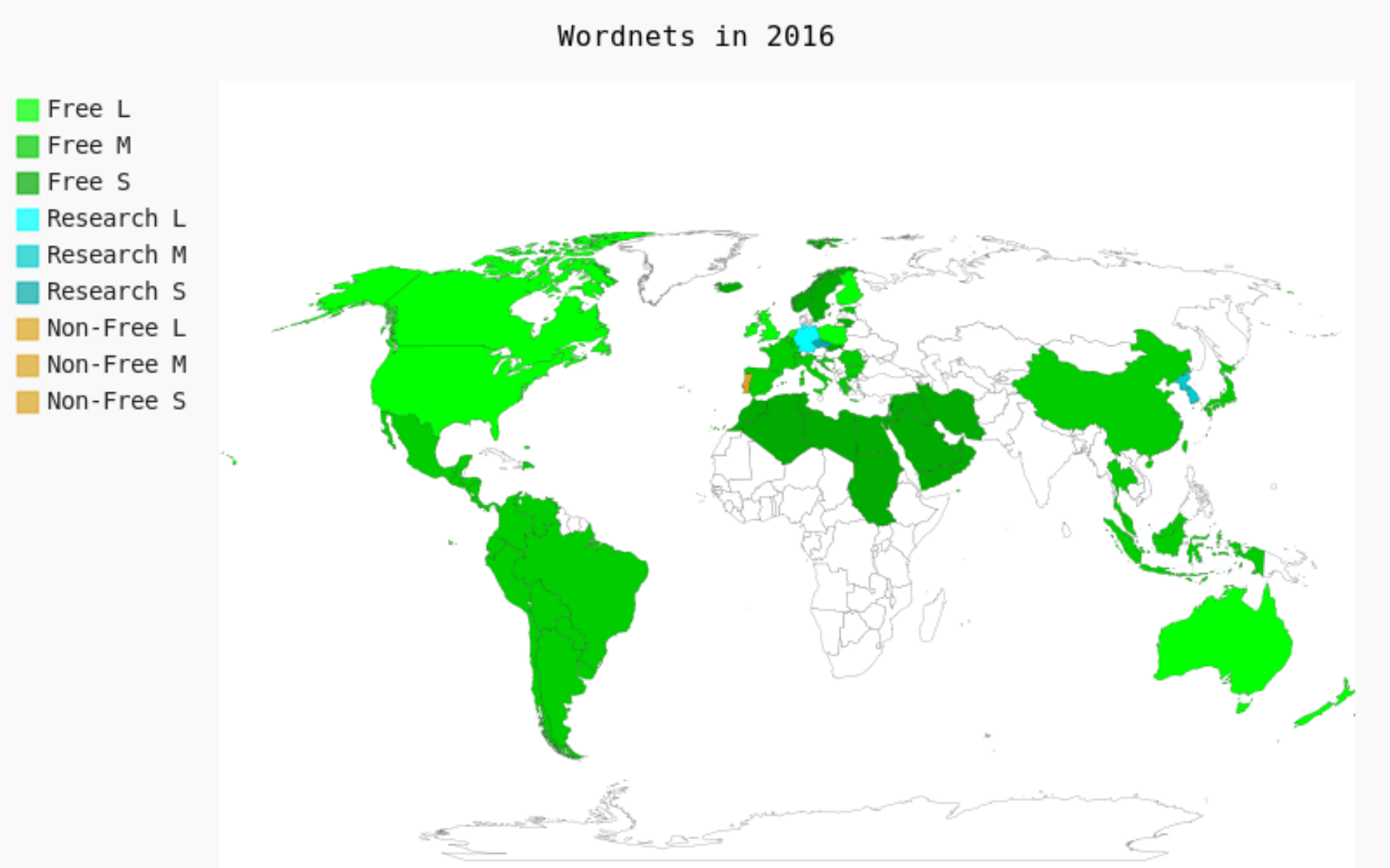
Wordnets in the world 2013-06

Chinese Open Wordnet (and Chinese wordnet)



Wordnets in the world 2016-01

Swedish, Dutch, Icelandic, Lithuanian, Romanian, …



Why do we want so many?

- Every new wordnet makes the network much richer $n(n-1)$ lexicons!
- Multilingual disambiguation makes it easier to add new languages
- New languages add new phenomena
  - and new synsets/concepts
  - and new relations
- More users mean more bugs found
- New approaches can be shared
  - UKB (does graph-based WSD) Basque
  - Wordnet glosses (disambiguates wordnet definitions) Princeton
  - Logical forms (gives LF for wordnet definitions) USC/ISI
  - UKB + Wordnet glosses + Logical forms (better WSD) Bulgaria

  ...

Faculty of Arts

How did we open the data?

- Leading by example (we made open wordnets for Japanese, Bahasa, Chinese)
- Appeal to self-interest: we showed that open resources are cited more (Bond and Paik, 2012)
- Simple format for sharing (tsv) — I wrote many converters
  Many checks for ill-formed lexicons
- Public praise for freed resources
- Private persuasion for non-open resources
- Open website
  - online interface (with statistics on coverage)
  - downloadable in multiple formats
  - linked to other resources (sentiment, time, SUMO, …)
- Used by other projects: Google Translate; Natural Language Toolkit (NLTK); Babelnet

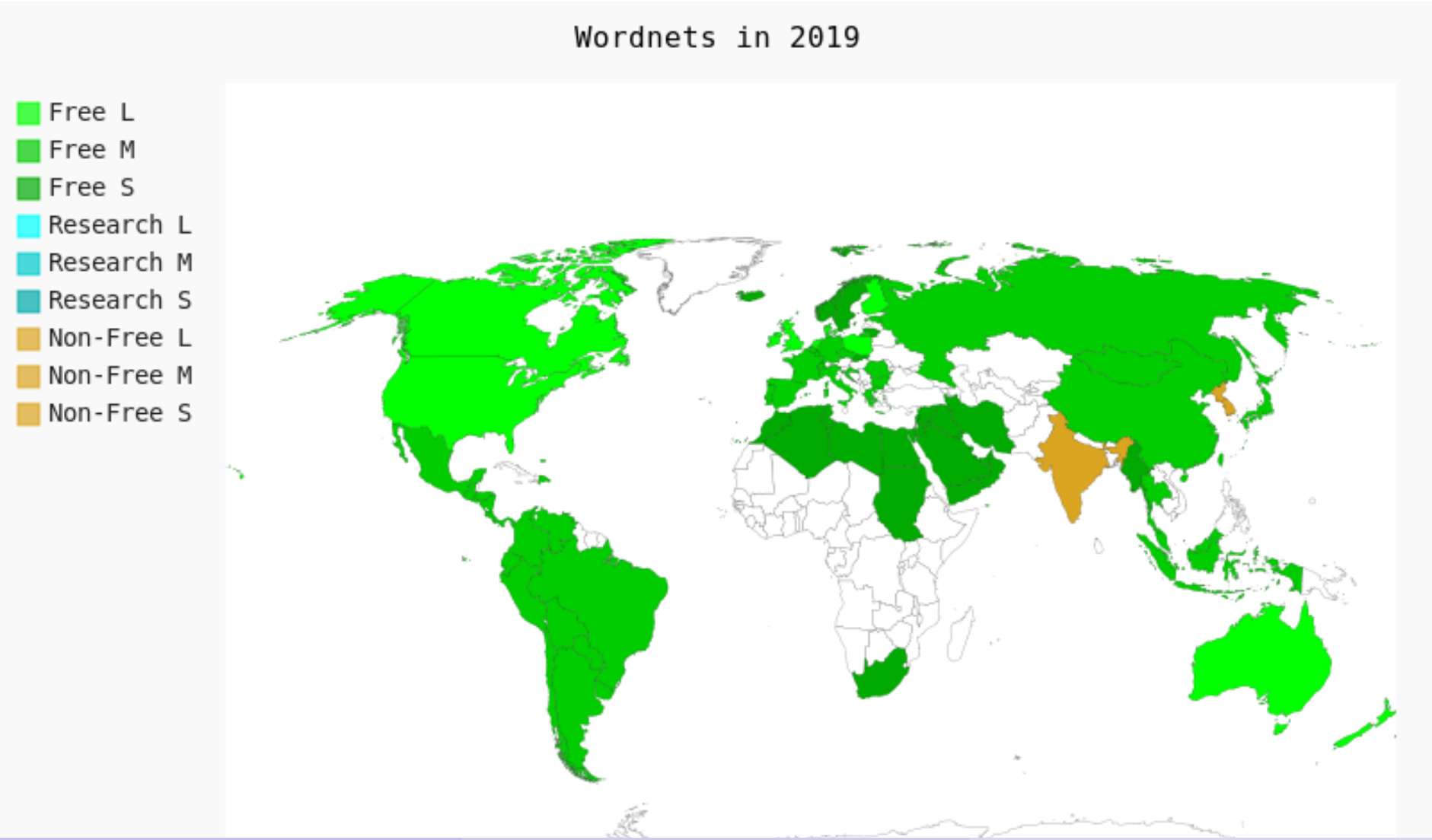
Faculty of Arts

Wordnets in the world 2017-12

Added: Moroccan Arabic, Abui, Myanmar, Sesotho, Tswana, Venda, Xhosa, Zulu;
Ancient: Greek, Latin, Sanskrit;
Freed: German, Estonian, Romanian, Czech

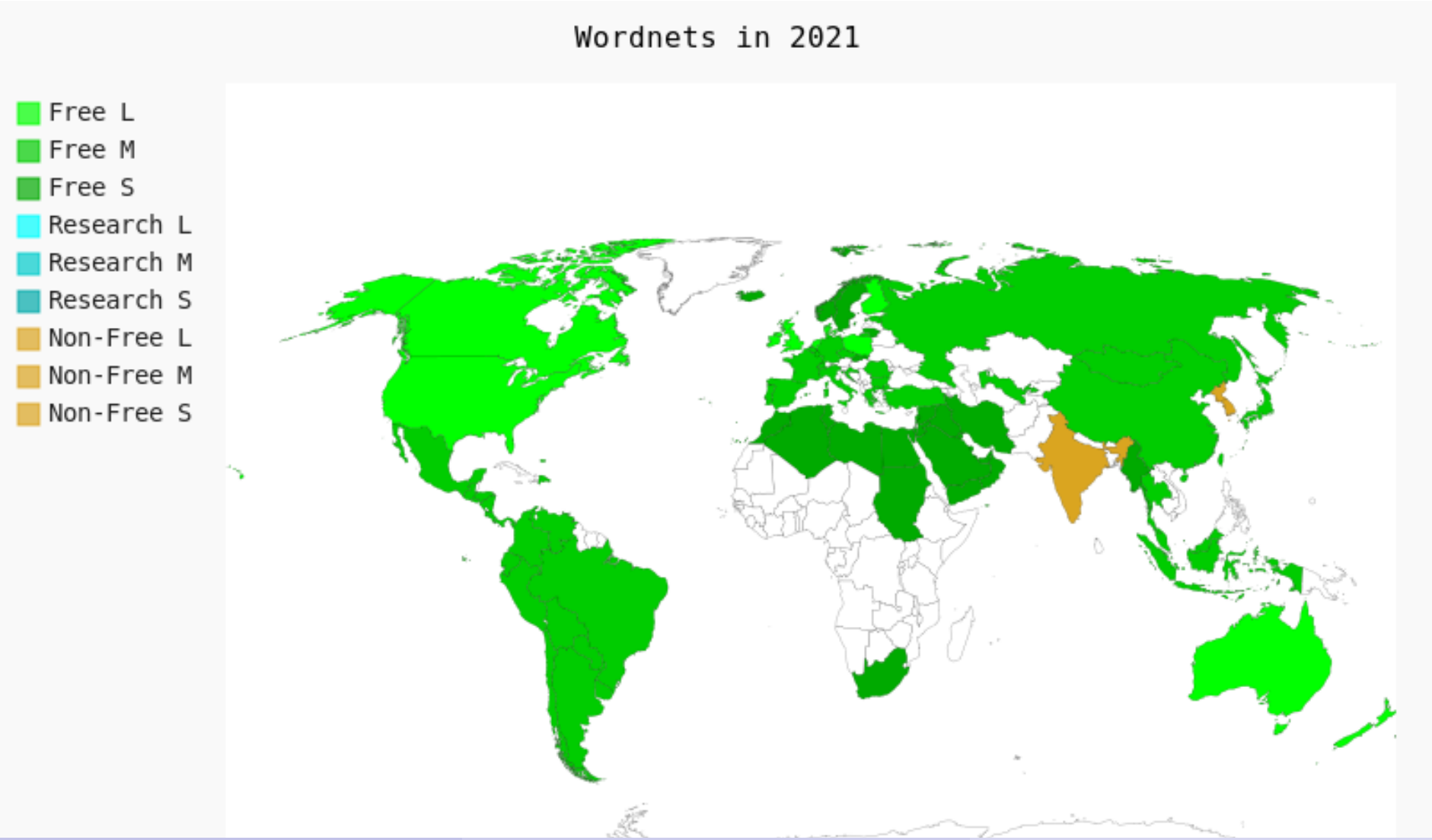
Wordnets in the world 2019-06

Added: Cantonese, Coptic, Russian, Mongolian, Danish
Expanded: English



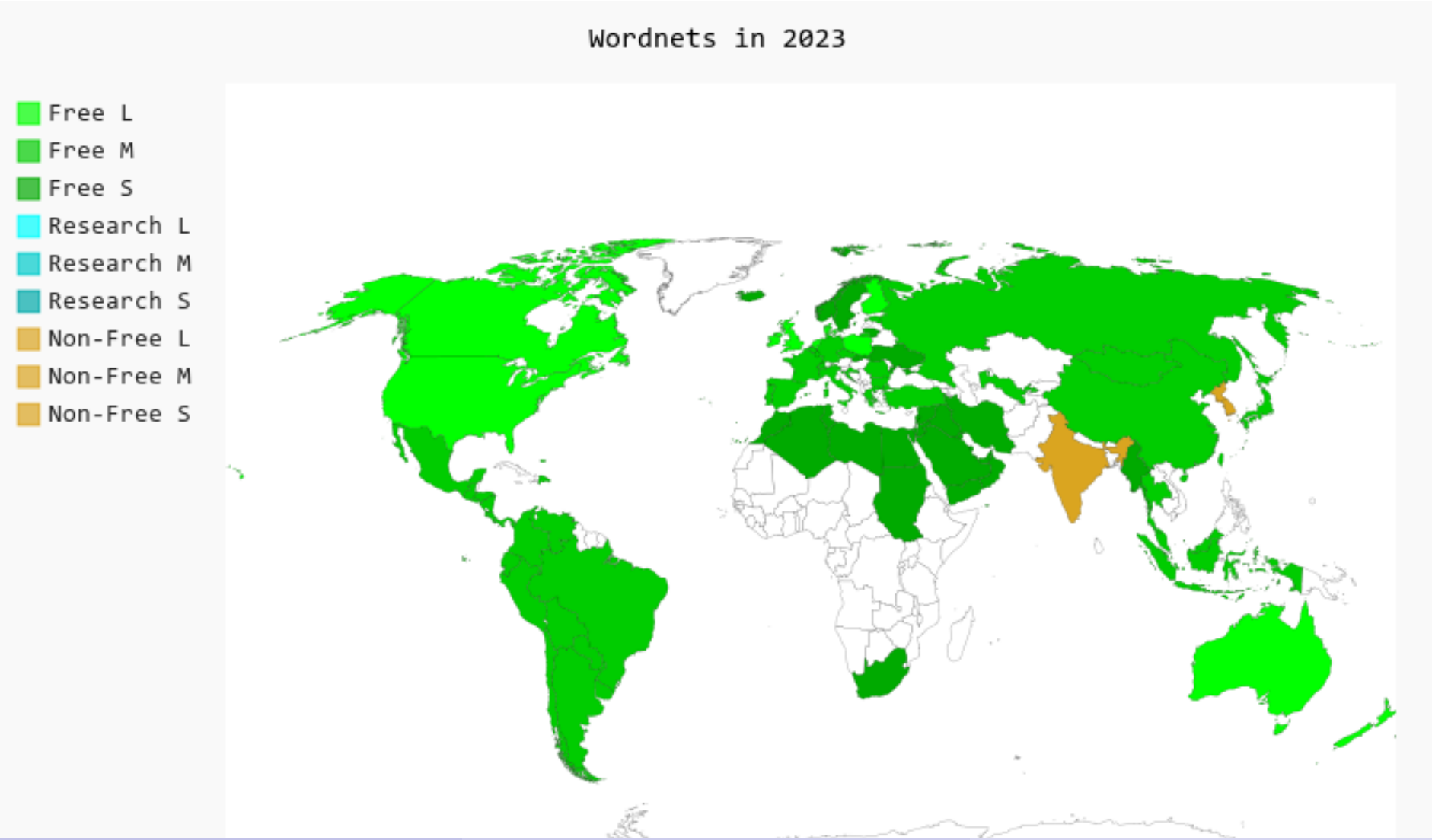
Wordnets in the world 2021-01

Added: Uzbek, Turkish;
Richer inventory of relations, better documentation



Wordnets in the world 2023-01

Added: Ukrainian;
Richer inventory of relations, better documentation



Current state

- 120,000+ synsets from the Open English Wordnet
  - 2,000 added locally
- 35 curated wordnets with 2,000,000 senses
- 250 languages with automatically created senses from wiktionary (at least 100 senses per language)
- 1,200 languages with senses mapped from various Swadesh lists (around 100-200 senses/language)

Faculty of Arts

The Open Multilingual Wordnet 1.0

- Defined a minimal shared format (incrementally useful) based on the hierarchy of PWN 3.0 (most wordnets are translations)
  - **name, license, URL** (metadata)
  - **synset & lemma** pairs (linked to PWN)
  - later also **definitions** and **examples**
- Wrote conversion scripts to extract this information
  - Provided feedback when there were errors
  - Occasionally suggested improvements
- Encouraged people to choose an open license
  we are deriving a new resource and redistributing
  — this needs permission
- Linked each resource to its canonical citation
  Encouraged people to cite the papers when they used the wordnets

Mostly done by me on weekends, some help from Lars Nygaard, John McCrae and of course the individual projects.

What else did we do?

- Minimal evaluation
  - Coverage of core concepts
    made the core concept list more easily available
- Created a common search interface
  - with links to other external resources: SUMO, sentiment, temporal, …
- Provided downloads in standard formats (LMF, LEMON)
- Automatically created wordnets from Wiktionary and CLDR
  - shared this data freely
- Created a Python API to access the data in NLTK (Bird et al., 2009)
  - Extended the NLTK corpus class to add information about recommended citation and licenses (for all corpora, not just the wordnets)
  - New wordnets and updates are still being added

Faculty of Arts

The Open English Wordnet

- The successor to the Princeton Wordnet (which is no longer developed)
- New vocabulary (thousands of new words and senses)
- New relationships (TABOO, AGENT, FEMALE-FORM)
- Many, many bug fixes
- Pronunciation (in IPA)

Faculty of Arts

Roadmap

Faculty of Arts

Many interesting things are not NVAR

From *The Adventure of the Speckled Band* (Conan Doyle, 1892), a short story currently being annotated as part of the NTUMC: some pronouns and exclamatives.

- *"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?"*
- *"Yes, but very small ones. Too narrow for anyone to pass through."*
- *"Thank you. That is quite settled" said he, rising and putting his lens in his pocket.*
- *"Hullo! Here is something interesting"*

Pronoun Motivation and Overview

- Attempting to model lexical and structural semantics
  - For multiple languages — identify cross-lingual differences
  - Exploit them to learn meaning (make the **implicit explicit**)
- Started by annotating **content words** (with **wordnets**)
- But nouns were often translated as pronouns$_i$ — so tag them$_i$
  1. **Identify pronouns** used in the corpus
  2. **Analyze in terms of components** — aids matching
     - Extended wordnet gives **full decompositional analysis**
  3. **Annotate** the pronouns **monolingually** in each language
     - **Link to** extended wordnet for **analysis**
  4. **Annotate** their correspondences across **languages**
  5. **Analyze the distribution cross-lingually**

Faculty of Arts

Identifying Pronouns

- Examined words tagged as pronouns in (Mandarin) Chinese, English, Japanese (and later Indonesian) parts of the NTU Multilingual Corpus (NTU-MC) — used the POS tags
  - Different tag-sets identified quite different collections
- We took the union, and filled in missing entries by hand
  - also referred to reference grammars
  - not complete, but getting there
- 117 different types; 249 tokens:

Chinese	57	
English	68	
Indonesian	40	(in progress)
Japanese	84	

- We include related determiners (demonstratives and quantifiers)

[numbers out of date]

Faculty of Arts

Components: Place

Head	Type/Proximity	English	Japanese	Chinese
Place	Interrogative	*where*	何処, どこ “doko”	哪里 “nǎlǐ”
	Proximal	*here*	此処, ここ “koko”	这里 “zhèlǐ”
	Distal	*there*		那里 “nàli”
	Medial		其処, そこ “soko”	
	Remote		彼処, あそこ “asoko”	
	Universal	*everywhere*	どこも “doko mo”	到处 “dàochù
	Existential		どこか “doko ka”	某处 “mǒuch
	Assertive	*somewhere*		
	Elective	*anywhere*		
	Other	*elsewhere*	よそ “yoso”	别处 “biéchù”

Not all lemmas shown

Faculty of Arts

Tagging Pronouns Mono-lingually

- Tagged one document by hand *The Adventure of the Speckled Band*
Language	English	Chinese	Japanese
Contentful	1,370	1,177	463
Other	75	19	51
Total	1,445	1,196	514
Sentences	599	620	702
Words	11,628	12,433	13,902
- Distinguished existential ***there*** (but not dummy ***it***) with POS tags
- *other* includes relative pronouns, dummy ***it***, idioms and segmentation errors

Faculty of Arts

Tagging and Analyzing Pronouns Cross-lingually

- Automatically linked by matching features
- Hand corrected:

	Linked Pronouns						Non-linked Prono	
	# Matching Features					Pronoun	English	Othe
	5	6	7	8	9	to Noun		
# Chinese	5	19	54	789	58	134	369	215
# Japanese	15	120	114	37	32	139	943	109

- Case and politeness mismatches common
- A surprising number of non-linked pronouns in Chinese and Japanese

Faculty of Arts

Interesting Cross-Linguistics Differences

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

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

b. 她 拿 枪 先 打 丈夫 ， 然后 打 自己
tā ná qiāng xiān dǎ zhàngfū, ránhòu dǎ zìjǐ
She$_i$ took the gun to first shoot husband$_j$, and then shot self$_i$.

Faculty of Arts

Interesting Cross-Linguistic Differences

(7) [many (cases) strange] …but none commonplace …

a. 但是 却 没有 一例 是 平淡无奇 的
Dan4shi4 que4 mei2you3 yi1li4 shi4 ping2dan4wu2qi2 de
'But, there is not one case that is featureless.'

b. どれも 尋常で は ない 事件 である
Dore mo jinjode wa nai jiken dearu
'Everything is a case which is not usual.'

(8) It is a swamp adder!

a. 这 是 一条 沼地 蝰蛇 !
Zhe4 shi4 yi1tiao2 zhao3di4 kui2she2 !
'This is a swamp adder !'

b. 沼蛇 だ !
numahebi da !
'ϕ is a swamp snake'

Faculty of Arts

Interjections

- words or phrases that
  - constitute a whole linguistic act
    (do not combine in integrated syntactic constructions)
  - do not refer to events, 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 …
- We introduce a new POS **x** for these non-referential expressions (also used for numeral classifiers in languages like Japanese)
  - definition with the form “an expression that is uttered ...:
  - EXEMPLIFIES: utterance (07109847-n)
  - enrich this flatter hierarchy with links to other existing concepts
    (when possible)

Faculty of Arts

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 tail of onomatopoeia
  (e.g. *ding-dong, woof-woof, miao, ...*)

Faculty of Arts

New Interjective Senses

Concept	Senses	Concept	Senses
Surprise, Wonder	58	Pity, Sorrow	19
Joy, Pleasure	17	Anger, Annoyance, Irritation	41
Approval, 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		
Total number of senses	336		

Faculty of Arts

New Interjective Senses

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
exemplifies	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
exemplifies	15167474-n (utterance)
see also	00167764-n (checkmate)

Roadmap

Faculty of Arts

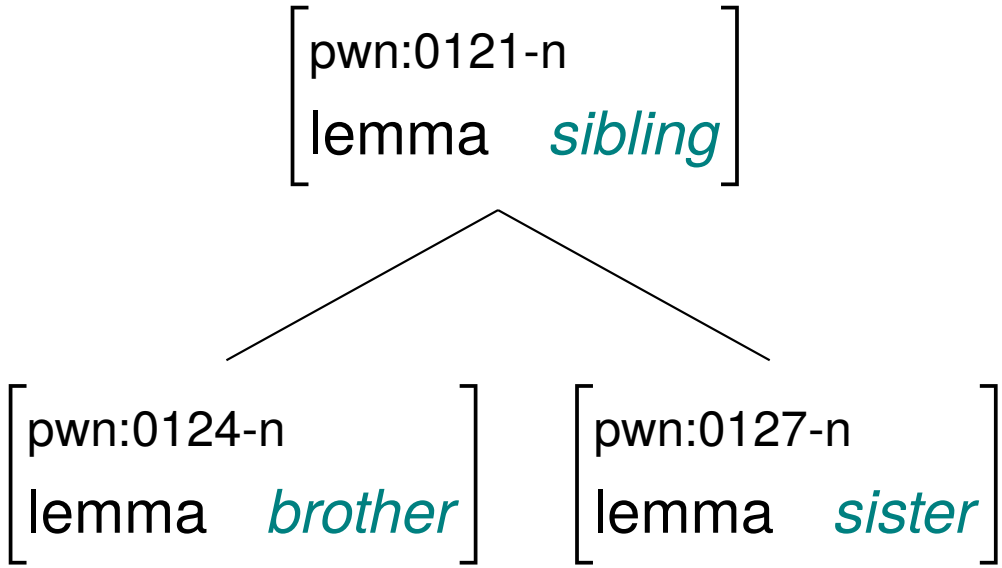
Why the Collaborative InterLingual Index?

We want to make it easier to link things together.

- There are wordnets for many languages
  - Currently they link through PWN (3.0)
- Many projects are adding new synsets
  - And not just synsets: lemmas, relations, POS, meta-data (domains, sentiment …)
- We want to be able to link them even if they are not in PWN
- We want to minimize wasted effort
  - Adding the same thing in different projects
  - Fixing the same errors in different projects
- We want to spread the burden of development

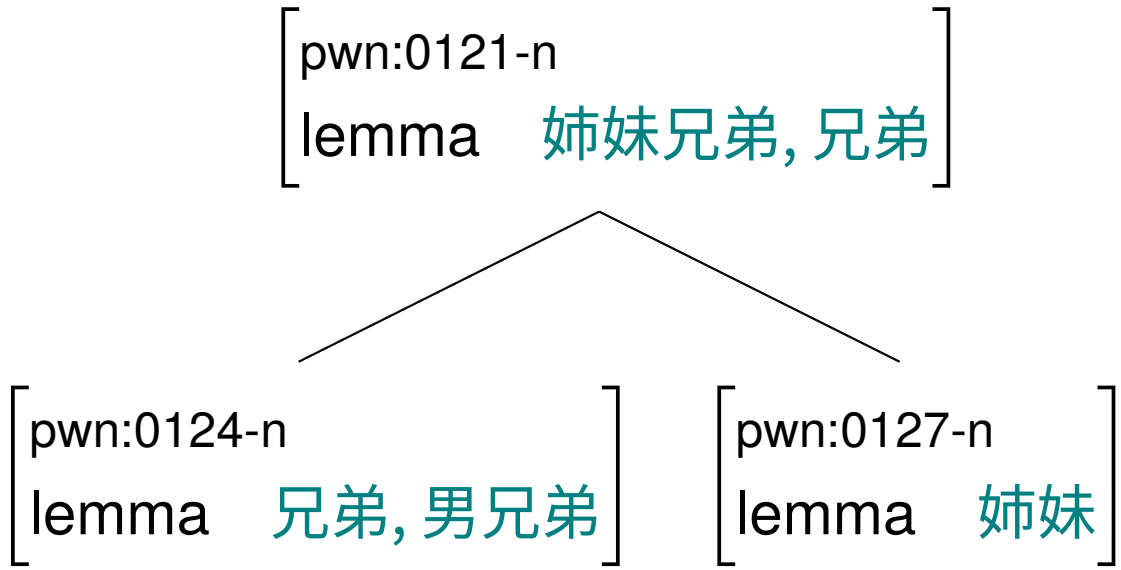
Faculty of Arts

The Princeton Wordnet of English



- Models the lexical structure of English

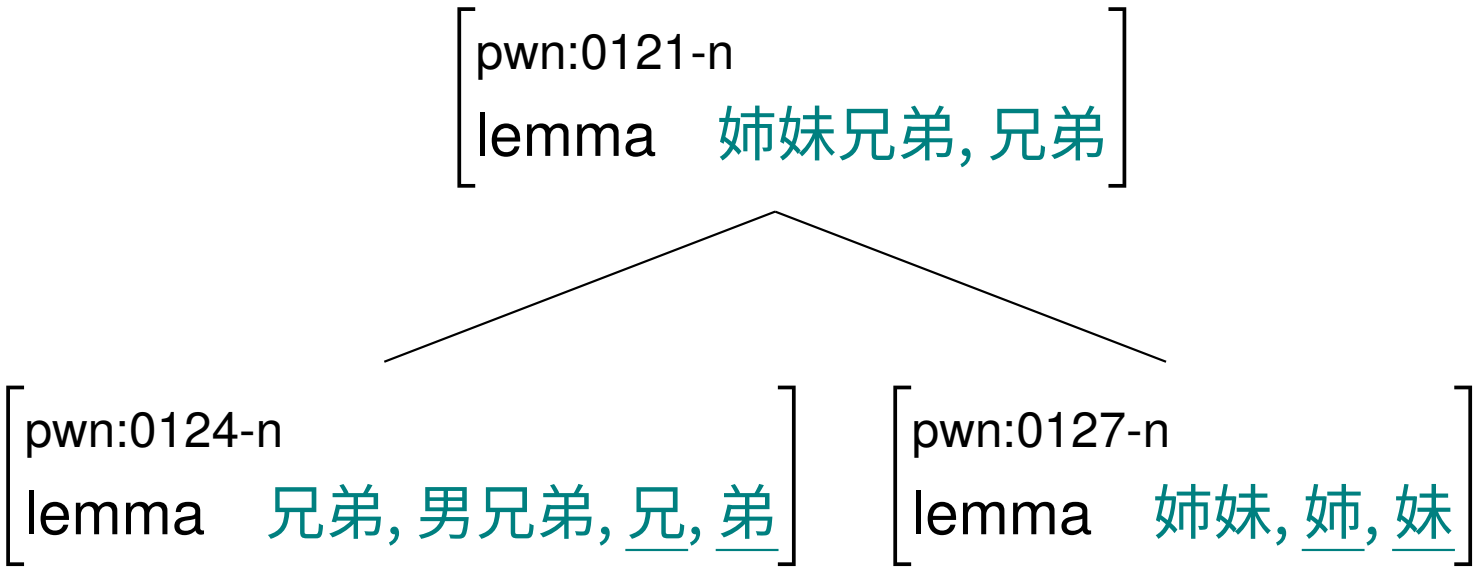
Japanese Wordnet (transfer structure) (–)



- Saves time by copying the English structure
- But misses some concepts (younger and older siblings)

Faculty of Arts

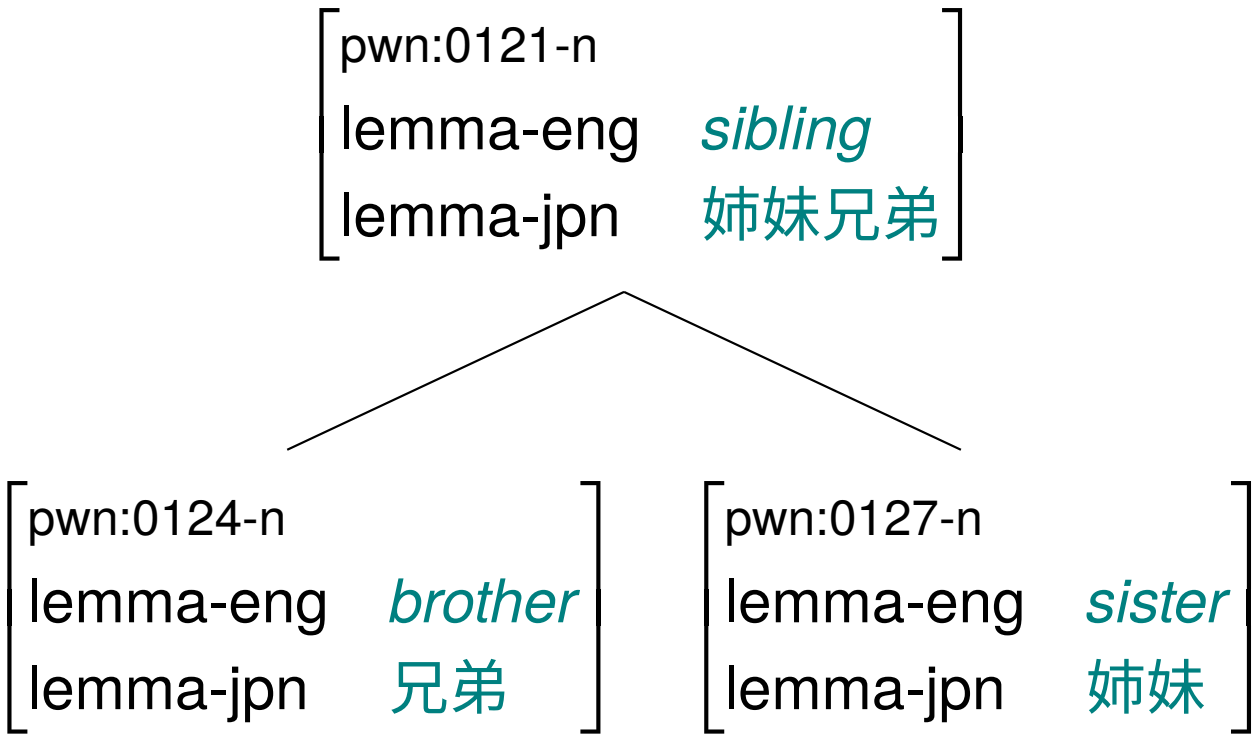
Japanese Wordnet (transfer structure) (+)



- Saves time by copying the English structure
- But merges dis-similar concepts (younger and older brother)

Faculty of Arts

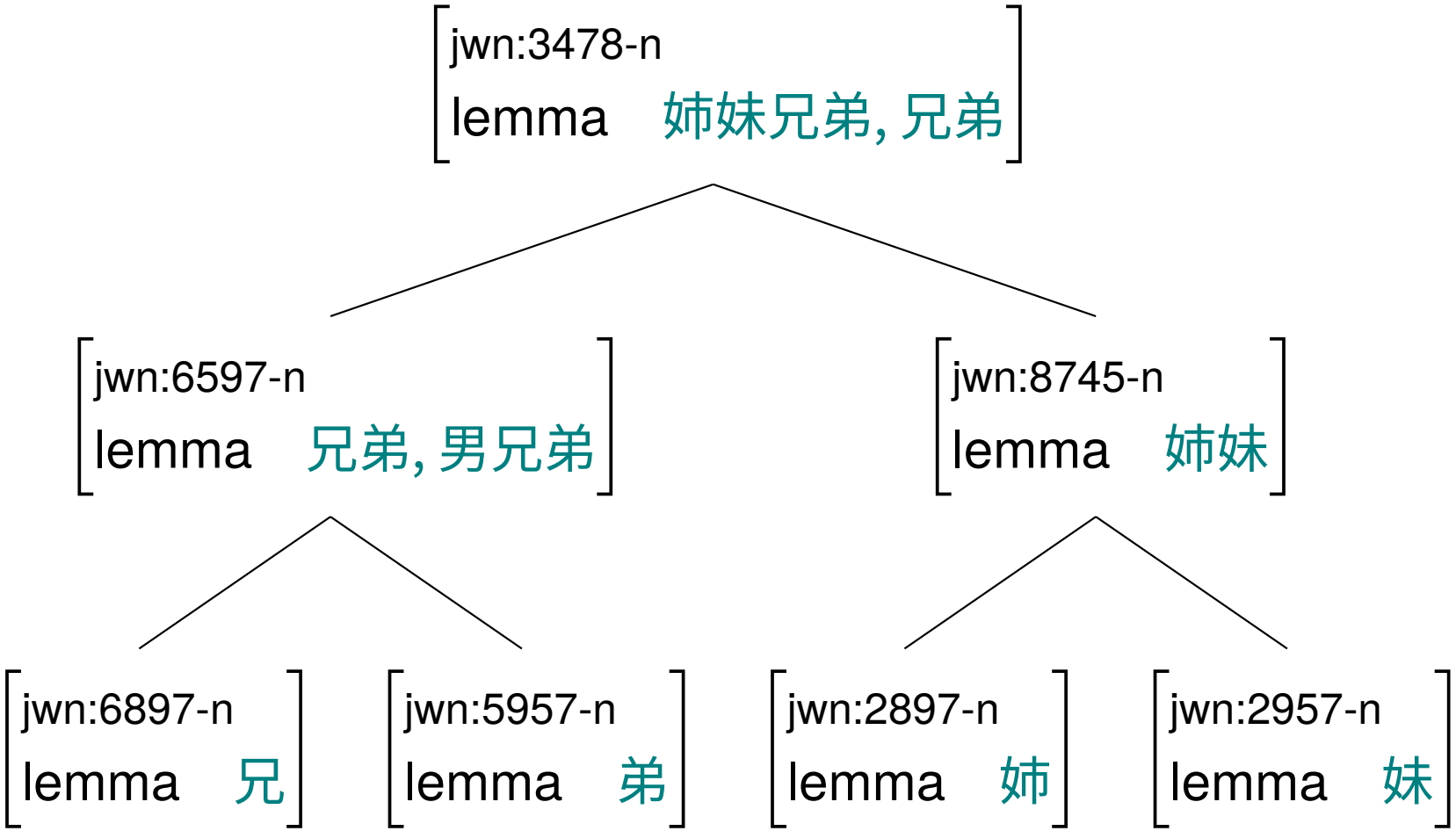
OMW 1.0



- Uses the English structure
- Loses any nodes not linked to English

Faculty of Arts

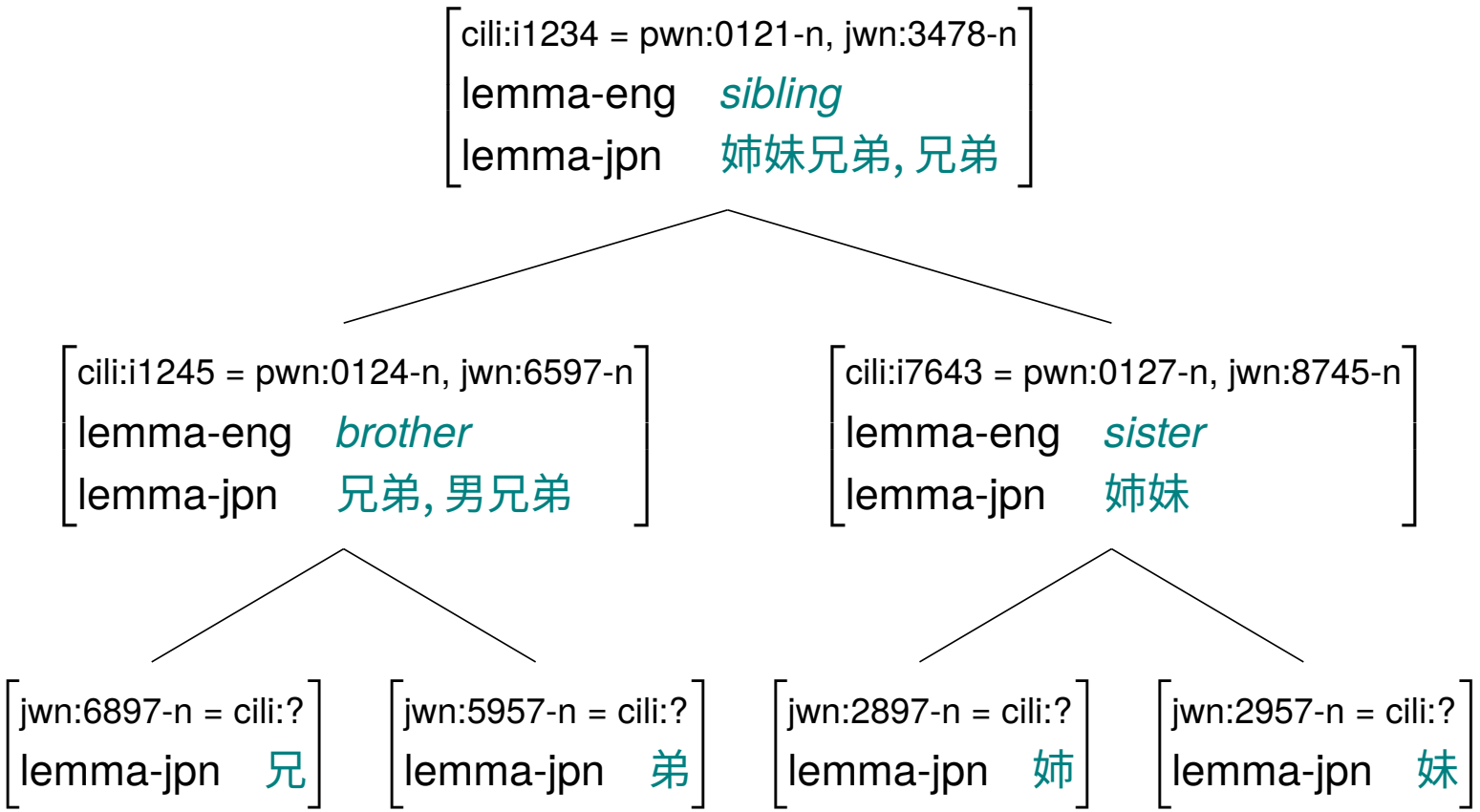
Japanese Wordnet (native structure)



- More accurately models Japanese

Faculty of Arts

OMW 2.0



- Unifies the structures of all languages
- Linked by the Collaborative Interlingual Index (CILI)
- Requires some coordination, …

Faculty of Arts

The solution: an InterLingual Index (ILI)

1. The Interlinear Index (ILI) should be a flat list of concepts (and instances).
2. The semantic and lexical relations should mean the same things for all languages.
3. Concepts should be constructed for salient and frequent lexicalized concepts in all languages.
4. Concepts linked to Multiword units (MWUs) in wordnets should be included.
5. A formal ontology could be linked to but separate from the wordnets.

Basic idea well known (Fellbaum and Vossen, 2008).

Faculty of Arts

Collaboratively Developed (CILI) …

1. The license must allow redistribution of the index
2. ILI IDs should be persistent: we never delete, only **deprecate** or **supercede**; we should not change the meaning of the concept
3. Each new ILI concept should have a definition in English, as this is the only way we can coordinate across languages.
   The definition should be unique (which is not currently true).
   Definition changes will be moderated.

…Collaboratively Developed (CILI)

1. Each new ILI concept should link to a synset in an existing project that is part of the OMW with one of a set of known relations (HYPERNYMY, MERONOMY, ANTONYMY, …)
2. This synset should link to another synset in an existing project that is part of the OMW and links to an ILI concept.
   - ⇒ each concept is linked to another concept through at least one wordnet in the grid
3. Any project adding new synsets should first check that they do not already exist in the CILI
   - New concepts are added through their existing in a wordnet
   - If something fulfills the criteria is proposed
   - If no objections after three months then it is added

Faculty of Arts

Roadmap

Faculty of Arts

Open Multilingual Wordnet 2.0

- OMW and merging hosted at UP, proxied through omwn
  http://omwn/omw/
  show provenance and confidence
- Individual wordnets can be uploaded
- ILI as RDF — shared on github
- Each new wordnet release can suggest ILI candidates
  - Mark as `ili='in'` "ILI new"
  - Only projects can add new ILI entries
- Review period (1–3 months)? — accept if no comments
- Allow orthographic variants
- Documentation (and all code) hosted on github

Faculty of Arts

Roadmap

Faculty of Arts

Where do we go from here?

- Convert existing wordnets to LMF
  - **Validate them**

  Possibly revise the LMF as needed
- Upload it to the Grid
  - We will validate it again
  - We add the wordnet to OMW: linking through ILI
  - Look for `ILI='in'` "ILI new"
  - Check the definition is good (and not too close); parse it?
  - Check it is linked to something existing
- Finish New OMW interface
- Add an interface for changing definitions, deprecating and superseding

Faculty of Arts

Where do we go from here?

- New POS: 'x', 'q', 'c', 'p', …
- Orthographic variants explicitly modeled
- More languages, links, words
- Sense tagged corpora and live examples
- More sense groupings (iliXXX ≈ iliYYY)
- Decompositional semantics: ***un-clasp***, ***today*** "this day"
- External DBS: `wikidata, species, geo-wordnet, images, National Cancer Institute Thesaurus, …`
- Integration with vector-space representations
  - modeling fuzziness
  - sense-embeddings
  - typing relations

Faculty of Arts

What will this enable?

- Better WSD
  - Fewer indistinguishable senses
  - More training data
  - More relations
- New synsets and senses:
  - ***smart phone***$_{n:1}$ “a mobile phone with more advanced computing capability and connectivity than basic feature phones”
  - ***klunen***$_{v:1}$ “to walk on skates across non-ice” (nl)
  - ***satay***$_{n:1}$ “A popular dish made from small pieces of meat or fish grilled on a skewer and served with a spicy peanut sauce (Nusuntara)”
  - around 20-30,000 coming soon
  - … and now it is easier to add them!

Faculty of Arts

References I

Stephen Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly. (www.nltk.org/book).

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

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. Saldo: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211. URL dx.doi.org/10.1007/s10579-013-9233-4.

Fišer Darja, Jernej Novak, and Tomaž. 2012. sloWNet 3.0: development, extension and cleaning. In *Proceedings of 6th International Global Wordnet Conference (GWC 2012)*, pages 113–117. The Global WordNet Association.

Faculty of Arts

References II

Valéria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: an open Brazilian Wordnet for reasoning. EMAp technical report, Escola de Matemática Aplicada, FGV, Brazil.

Sabri Elkateb, William Black, Horacio Rodríguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Building a wordnet for Arabic. In *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006)*.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Radovan Garabík and Indrė Pileckytė. 2013. From multilingual dictionary to lithuanian wordnet. In Katarína Gajdošová and Adriána Žáková, editors, *Natural Language Processing, Corpus Linguistics, E-Learning*, pages 74–80. Lüdenscheid: RAM-Verlag. http://korpus.juls.savba.sk/attachments/publications/lithuanian_wordnet_2013.pdf.

Faculty of Arts

References III

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese wordnet: Design and implementation of a cross-lingual knowledge processing infrastructure. *Journal of Chinese Information Processing*, 24(2):14–23. (in Chinese).

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*. Marrakech.

Faculty of Arts

References IV

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

Nurril Hirfana Mohamed Noor, Suerya Sapuan, and Francis Bond. 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.

Mortaza Montazery and Heshaam Faili. 2010. Automatic Persian wordnet construction. In *23rd International conference on computational linguistics*, pages 846–850.

Antoni Oliver, K. Šojat, and M. Srebačić. 2015. Automatic expansion of Croatian wordnet. In *Proceedings of the 29th CALS international conference "Applied Linguistic Research and Methodology"*. Zadar (Croatia).

Faculty of Arts

References V

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

Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. DanNet — the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3):269–299.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: Developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302. Mysore, India.

Faculty of Arts

References VI

Maciej Piasecki, Stan Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Wroclaw University of Technology Press. URL http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/A_Wordnet_from_the_Ground_Up.pdf, (ISBN 978-83-7493-476-3).

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

Ida Raffaelli, Božo Bekavac, Željko; Agić, and Marko Tadić. 2008. Building Croatian wordnet. In Attila Tanács, Dóra Csendes, Veronika Vincze, Christianne Fellbaum, and Piek Vossen, editors, *Proceedings of the Fourth Global WordNet Conference 2008*, pages 349–359. Szeged.

Faculty of Arts

References VII

Ervin Ruci. 2008. On the current state of Albanet and related applications. Technical report, University of Vlora. (http://fjalnet.com/technicalreportalbanet.pdf).

Benoît Sagot and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco.

Kiril Simov and Petya Osenova. 2010. Constructing of an ontology-based lexicon for Bulgarian. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta.

References VIII

URL http://www.lrec-conf.org/proceedings/lrec2010/summaries/848.html.

Sofia Stamou, Goran Nenadic, and Dimitris Christodoulakis. 2004. Exploring Balkanet shared ontology for multilingual conceptual indexing. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, page 781–784. Lisbon.

Sareewan Thoongsup, Thatsanee Charoenporn, Kergrit Robkop, Tan Sinthurahat, Chumpol Mokarat, Virach Sornlertlamvanich, and Hitoshi Isahara. 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.

Faculty of Arts

References IX

Antonio Toral, Stefania Bracal, Monica Monachini, and Claudia Soria. 2010. Rejuvenating the Italian wordnet: upgrading, standardising, extending. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*.

Dan Tufiş, Radu Ion, Luigi Bozianu, Alexandru Ceauşu, and Dan Ştefănescu. 2008. Romanian wordnet: Current state, new applications and prospects. In *Proceedings of the 4th Global WordNet Association Conference*, pages 441–452. Szeged.

Piek Vossen, editor. 1998. *Euro WordNet*. Kluwer.

Piek Vossen and Marten Postma. 2014. Open Dutch wordnet. In *Proceedings of the 7th Global WordNet Conference (GWC 2014)*. Tartu. (presentation only).

Faculty of Arts

References X

Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP-2013*, pages 10–18. Nagoya.