

日本語の意味を世界につなぐ 日本語 Wordnet / The Japanese Wordnet linking Japanese meanings to the world

Francis Bond
and many more

Linguistics and Multilingual Studies,
Nanyang Technological University (南洋理工大学)

[<bond@ieee.org>](mailto:bond@ieee.org)

TUFS 2016-12-13

① 日本語 Wordnet

- Intro
- The Japanese Wordnet
- Multilingual Wordnet
- NTU-MC
- CILI: the Collaborative InterLingual Index
- Future Work

② Jacy: Japanese HPSG

③ Technology Enhanced Learning: 技術強化された学習

- Vocabulary Learning
- First Language Learning
- Second Language Learning

Self Introduction: 凡士 (ボンド) フランシス

- 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
- 2009- NTU (Nanyang Technological University)
 - ▶ Abui, Chinese, Malay, Multilingual Wordnets
 - ▶ HPSGs for Chinese, Indonesian, ...
 - ▶ Multilingual Meaning Banks (Treebank + Sensebank)



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

<http://wordnet.princeton.edu/>

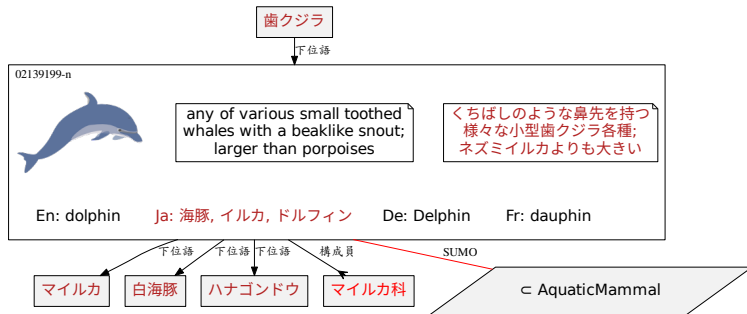
- Made up of four separate (but interlinked) semantic nets, for nouns, verbs, adjectives and adverbs
- WordNets exist for many languages, at NTU we work on:
 - ▶ Japanese
 - ▶ Bahasa Malay/Indonesian
 - ▶ Chinese
 - ▶ Abui (a Papuan language of Alor)
 - ▶ The shared **Open Multi-lingual Wordnet** (150+ languages)

<http://compling.hss.ntu.edu.sg/omw/>

Wordnet Structure

- Lexical items are categorised into $\sim 115\text{K}$ (v 3.0) glossed **synsets** (= synonym sets \approx concepts)
wordnet has grown over time, and continues to do so
- Lexical relations are at either the synset level or sense (= combination of lexical item and synset) level
- Strongly lexicalist (originally):
 - ▶ synsets only where words exist
 - ▶ but many multiword expressions ($\approx 50\%$)
 - ▶ (near) absence of frame semantics

The synset for *dolphin*



Here we show multiple languages: Japanese, English, French and German

Semantic Relations

- Constitutive

*driver*_{n:1} HYPONYM *operator*_{n:1}

*driver*_{n:1} ANTONYM *nondriver*_{n:1}

*driver*_{n:3} SYNONYM *number one wood*_{n:1}

*golf club*_{n:1} MERONYM *clubhead*_{n:1}

- Derivational

*driver*_{n:1} DERIVED *drive*_{v:1}

- Descriptive

*driver*_{n:2} DOMAIN *computer science*_{n:1}

- Implied (from e.g. definitions)

*driver*_{n:1} “the *operator*_{n:1} of a *motor vehicle*_{n:1}”

*driver*_{n:1} “a *program*_{n:3} that *determines*_{v:2} how a *computer*_{n:2} will
*communicate*_{n:1} with a *peripheral device*_{n:1}”

- Wordnets are used for
 - ▶ studying word similarity
 - ▶ describing the meaning of texts
 - ▶ testing WSD algorithms
 - ▶ improving parse ranking
 - ▶ machine translation
 - ▶ puzzle generation and solving
 - ▶ joke generation
 - ▶ distinguishing senses for word embeddings
 - ▶ language learning

The Japanese Wordnet

- Initially assume that the semantic structure is the same

▶ *dog* \subset *animal*

⇒ 犬 \subset 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, examples
2010-10-22	1.10	57,238	93,834	158,058	
2012-01-06					Japanese Semcor
2014-02-06					NLTK module
2017-??-??	2.0	60,000	80,000	160,000	+220,000 variants

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, たんぱく質, 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, ぷろていん)

- 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)

吸引 (キュウイン, きゅういん)

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

五十歩百歩

- Added corpus-based examples and sense frequency

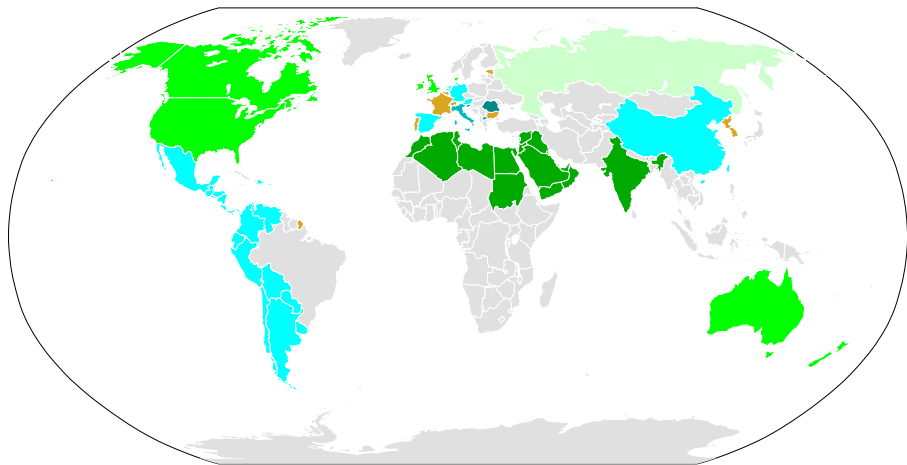
対策₃, 策₃, 措置₂, 方略, 方策, 術, 打つ手 “step, measure”

- Added exclamatives

大手, ヨイショ, お早うございます

Wordnets in the world 2008-06

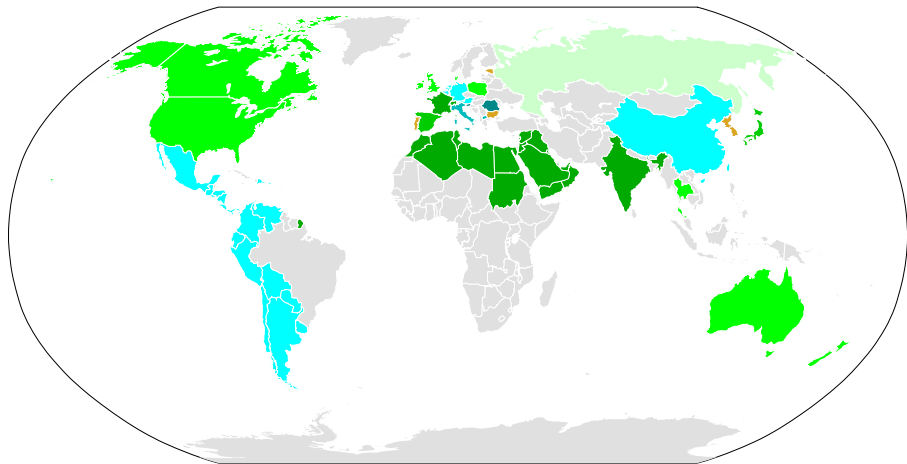
Many wordnets, but few free: we decided to build an open wordnet for Japanese and needed other wordnets for cross-lingual disambiguation.



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

Wordnets in the world 2011-06

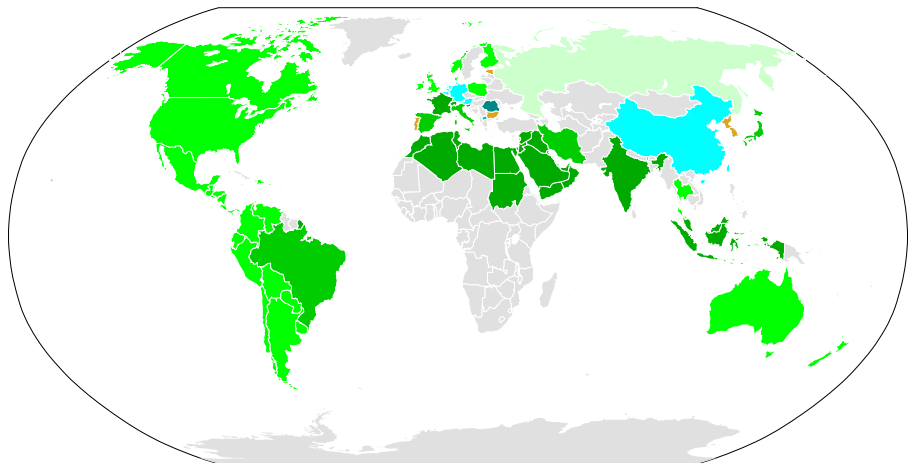
Free wordnets for French, Catalan, Polish, Thai and **Japanese**



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

Wordnets in the world 2012-06

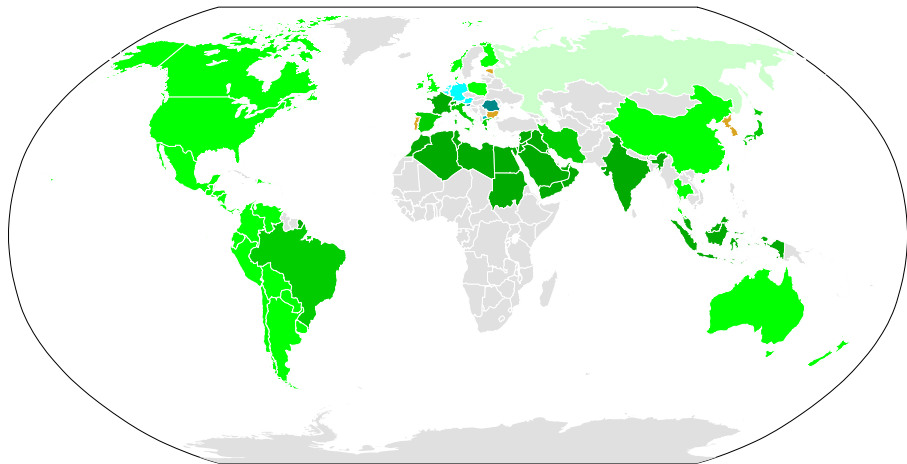
Spanish freed (and Galician and Basque), Farsi, Norwegian, Swedish, **Bahasa** (Malay and Indonesian)



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

Wordnets in the world 2013-06

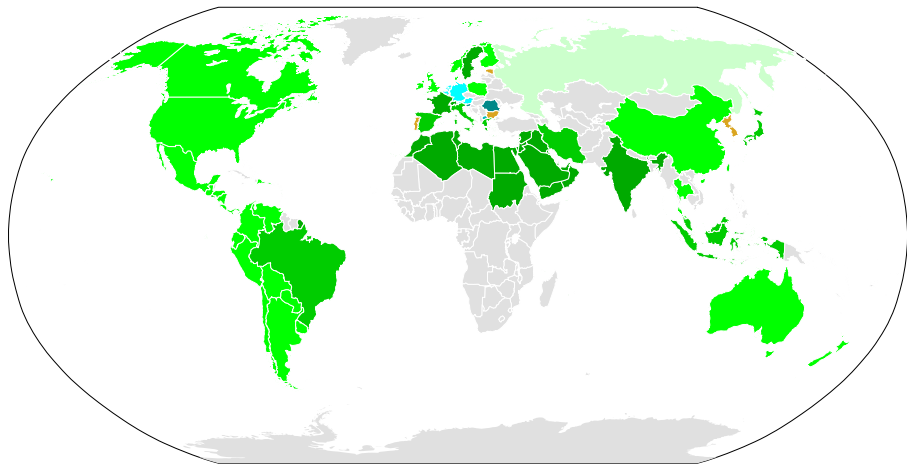
Chinese Open Wordnet (and Chinese wordnet)



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

Wordnets in the world 2014-06

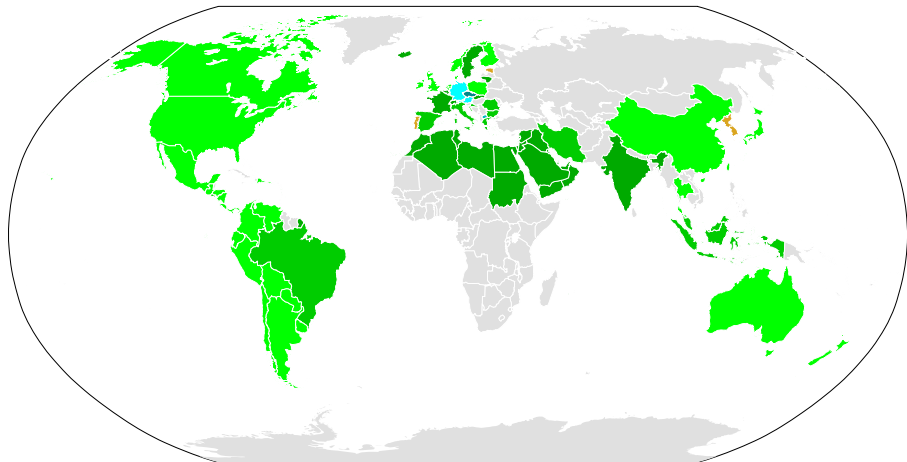
Swedish



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

Wordnets in the world 2016-01

Dutch, Icelandic, Lithuanian, Romanian, ...



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

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)
 - ▶ Wordnet glosses (disambiguates wordnet definitions)
 - ▶ Logical forms (gives LF for wordnet definitions)
 - ▶ UKB + Wordnet glosses + Logical forms (better WSD)
- ...

Basque
Princeton
USC/ISI
Bulgaria

How did we do it?

- Leading by example (all our wordnets are open)
- Appeal to self-interest: 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

Open Multilingual Wordnet (1.3)

- 117,000 synsets from Princeton wordnet
 - ▶ 2,000 added locally
- 34 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)
- all linked to Japanese through the Japanese wordnet

NTU multilingual corpus

We use it to describe the meaning of texts

- Small, deeply analysed corpus
 - ▶ 6,000+ sentences x 3 languages (cmn, eng, jpn); 2,000 in Indonesian
 - ★ Mainichi Newspaper (NICT translations)
 - ★ Sherlock Holmes ([DANC](#), [SPEC](#), [REDH](#)[eng])
 - ★ Cathedral and the Bazaar (many languages)
 - ★ Spider's thread 「蜘蛛の糸」
 - ★ Singapore Tourist data (plus Korean, Viet, Indo)
 - ▶ [Hand alignment](#), [WordNet tagging](#), Treebanking, Sentiment, ...
- Revised the DB structure; second round of wordnet tagging
 - ▶ Slightly behind schedule
 - ▶ Wordnet extension is being done in parallel
- Tagged Italian ([SPEC](#))
ready to start Bulgarian, Dutch, German, Polish

Interesting Cross-Linguistics Differences

(1) Said he suddenly

- a. ホームズが 突然 口 を 開く
ho-muzu ga totsuzen kuchi wo hiraku
Holmes NOM suddenly mouth ACC open
Holmes opens his mouth suddenly

pronoun ↔ noun

adv ↔ adj

verb ↔ noun+verb

Interesting Cross-Linguistics Differences

(2) 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

Interesting Cross-Linguistic Differences

(3) [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.'

(4) It is a swamp adder!

a. 这 是 一 条 沼地 蝮蛇!
Zhe4 shi4 yiltiao2 zhao3di4 kui2she2 !
'This is a swamp adder!'

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

Why the Collaborative InterLingual Index?

We want to make it easier to link things together.

- There are wordnets for many languages
 - ▶ Currently they mainly 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
御飯, 米, 玄米 (jpn); *nasi, beras, gabah* (ind)
来年 (jpn); 明年 (mcn); *next year* (eng)
 - ▶ Fixing the same errors in different projects
- We want to spread the burden of development

Where do we go from here?

- A Collaborative Interlingual Index (CILI)
 - ▶ Shared by all projects
 - ▶ New concepts and relations can be added
 - ▶ Coordinated on git-hub
- Convert existing wordnets to LMF linked to CILI
- 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

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!

Lexical and Structural Meaning

- Wordnet gives us the link from words to the world (lexical semantics)
- But we also want to know who does what to whom
 - ▶ We need structural semantics too
- So we develop a grammar of Japanese: JACY
 - ▶ HPSG-based Grammar of Japanese:
 - ▶ Morphological analyser: **Mecab** (NAIST-IPA)
 - ▶ 37,000 word vocabulary (contains all core vocabulary)
gives lexical-type for unknown words using pos
 - ▶ Semantic representation: Minimal Recursion Semantics
 - ▶ Development with CSLI, DFKI, Kobe Shoin, Osaka Graduate Uni., NTT, NAIST, Saarland, Washington, NTU
- New book out now!
Melanie Siegel, Emily Bender and Francis Bond (2016) *JACY an implemented grammar of Japanese*, CSLI Publications

Why HPSG?

- Head-driven Phrase Structure Grammar is a theory of grammar with several desirable properties
 - ▶ **Mono-stratal**: Orthography, Syntax, Semantics, Pragmatics are all handled in a single structure (the **sign**)
 - ▶ **Constraint-based**: parses are built up compositionally, with new information constraining the range of interpretations
 - ▶ **Lexicalist**: Word structure and phrase structure are governed by partly independent principles. Words and phrases are two kinds (subtypes) of sign. Lexical information is organized in terms of multiple inheritance hierarchies and lexical rules that allow complex properties of words to be derived from the structure of the lexicon.
 - ▶ **Constructionist**: Constructions (phrase-rules) are also modeled as feature structures. This allows constructions to be analyzed via multiple inheritance hierarchies modeling the fact that constructions cluster into groups with a family resemblance that corresponds to a constraint on a common supertype.

Deep Linguistic Processing with HPSG (DELPH-IN)

- Informal collaboration adopting Head-Driven Phrase Structure Grammar (HPSG) and Minimal Recursion Semantics (MRS).
- A shared format for grammatical representation
 - ▶ Typed Feature Structures (written using TDL)
- A repository of *open-source* tools and test sets
 - ▶ Grammar development environment (LKB)
 - ▶ Efficient parsers/generators for NLP (PET, ACE)
 - ▶ Dynamic treebanking (ITSDB, ACE)
 - ▶ Machine Translation engine (LOGON, ACE)

With stochastic models to select the most plausible interpretation

- A yearly summit
- Collaboration between linguists and computer scientists

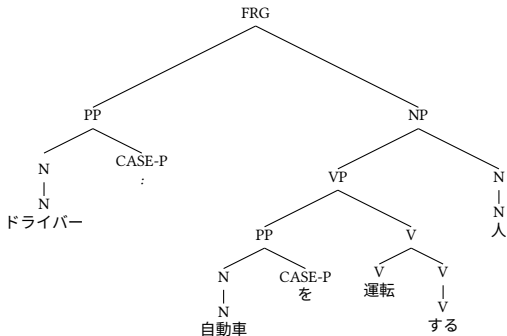
- English Resource Grammar (ERG)
 - ▶ Robust wide-scale grammar of English
 - ▶ Used for MT, CALL, finding negation, grammar correction, ...
 - ▶ Large treebank (Redwoods)
- **Jacy** (Japanese grammar)
 - ▶ Medium sized grammar of Japanese
 - ▶ Smallish treebank (Hinoki 檜)
 - ▶ Used to build accurate but brittle MT (JaEn)
- GG (German)
- SRG (Spanish)
- **Zhong** (Chinese)
- **INDRA** (Indonesian)
- ...

Jacy: ドライバー: 自動車を運転する人

(5) ドライバー: 自動車を運転する人

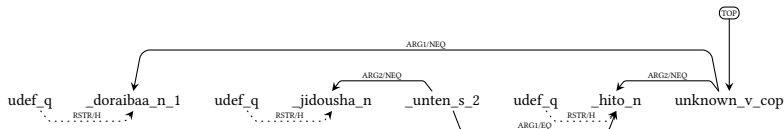
driver: a person who drives a car

(6)



Jacy: ドライバー: 自動車を運転する人

(7)

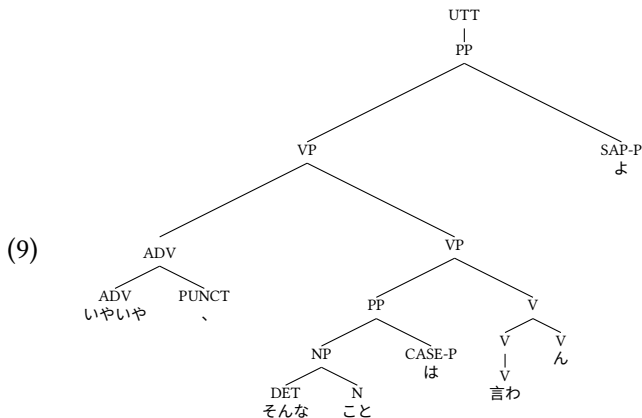


- Some words do not have predicates (e.g. が, を)
- Some constructions add predicates (e.g. **NP:NP** unknown_v_cop)
- Dependency MRS underspecifies quantifier scope
- Jacy adds quantifiers (not shown here)
- Demo here: <http://delph-in.github.io/delphin-viz/demo>

Jacy: いやいや、そんなことは言わんよ

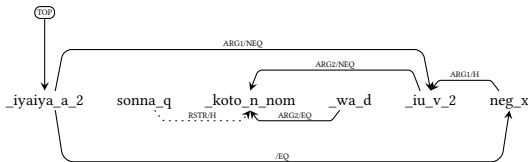
(8) いやいや、そんな_k こと_j は言わ_s ん_i よ

I am sure that I shall say_s no_i thing_j of the kindg_k



Jacy: いやいや、そんなことは言わんよ

(10)



- *iyaiya* is a scopal adverb
- The negation is also scopal
- We could decompose sonna(x) further
sono(e) kind(e,x) “that kind of”

Jacy Summary

- Jacy is a medium coverage grammar of Japanese
- It attempts to give meaningful semantics
 - ▶ compatible with other DELPH-IN grammars
- It has been used for
 - ▶ machine translation
 - ▶ knowledge acquisition from dictionaries
 - ▶ treebanking corpora
- It is still being (slowly) developed

Introduction

- Language Tutoring Goals
- Teaching Vocabulary
- Teaching Writing (feedback through mal-rules)
- Teaching a Second Languages (feedback through translation)
- Learning a user model

What about Language Learning and Technology?

Education and Technology:

Technology is an increasingly important part of higher education! (MOOCS, E-Learning Platforms, Blended-Learning Classrooms, etc.)

But the technology for Language Learning is not fully developed:

- to know and help individual student's weaknesses
- to drill student's weaknesses with exercises
- to provide precise, informed feedback on how to improve
- to help assess students' level of proficiency
- to scale to 100s or 1000s of students

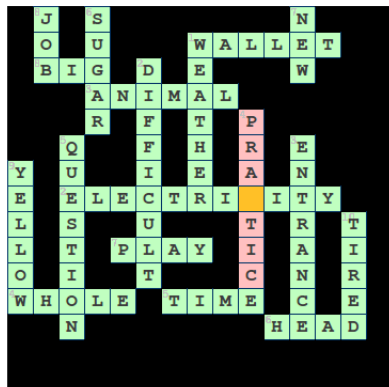
Vocabulary Learning

- You need to know about 2 ~ 3,000 words to be able to boot-strap communication in a language: that is enough to ask about words you don't understand and understand the answer
- Once you know these, you can proceed monolingually
- But first you have to learn these — it takes a kid four years to get to here.
- But we can use another language as scaffold

The cross-lingual crossword

Undergraduate Honor's Thesis (Jeanette Tan), code and demo at: <https://github.com/zenador/multi-xwords>

> Timeout waiting for server to reply!



Game Over!

You have 97 points.

Score: 97 points

Across

1. さいふ
2. でんき
3. どうぶつ
4. ぜんぶ
5. じかん
6. あたま
7. あそびます
8. おおきい

Down

1. でんき
2. むずかしい
3. いりぐち
4. れんしゅう
5. しつもん
6. さとう
7. あたらしい
8. しごと
9. きいろい
10. つかれました

Time: 181:15

The cross-lingual crossword (ii)

- **Clues** are shown all the time
- **Hints** can be asked for (and lose you points)
- We get data from the wordnets in the Open Multilingual Wordnet (OMW: Bond and Foster, 2013). This gives us:
 - ▶ L1 definition, synonyms
 - ▶ L2 definition, synonyms
 - ▶ Pictures (for the concept)
 - ▶ For some languages, pronunciation/transliteration

Not all wordnets have all the information for all synsets

See <http://compling.hss.ntu.edu.sg/omw/> for the current list.

The cross-lingual crossword (hints)

For example: we want *dog* (L2)

- L2 definitions and synonyms (Fellbaum, 1998)
 - ▶ a member of the genus *Canis* (probably descended from the common wolf) that has been domesticated by man since prehistoric times
 - ▶ *domestic dog*, *Canis familiaris*
- L1 definitions, synonyms, transliterations and examples (Bond et al., 2010)
 - ▶ 有史以前から人間に家畜化されて来た（おそらく普通のオオカミを先祖とする）イヌ属の動物
 - ▶ 犬
 - ▶ イヌ
 - ▶ いぬ
 - ▶ *inu*
 - ▶ “彼女は犬と一緒に散歩します。”

- Pictures



The cross-lingual crossword (hints)

- L1 definitions, synonyms and transliterations (Bulgarian: Simov and Osenova, 2010)
 - ▶ Вид домашно животно от семейство хищни бозайници, с различна големина, цвят на козината и различни породи, което лае и често се използва като пазач на дома и имота, за лов, може да бъде дресирано и обучавано за различни служебни цели.
 - ▶ *куче, kuche*
- L1 definitions, synonyms and transliterations (Greek: Stamou et al., 2004)
 - ▶ σκύλος του γένους Canis familiaris που συνήθως προέρχεται από τον κοινό λύκο και έχει εξημερωθεί από τους προϊστορικούς χρόνους
 - ▶ *σκύλος γένους, skýlos génou*
- L1 definitions, synonyms and transliterations (Basque: Pociello et al., 2011)
 - ▶ *zakur, txakur, or*

The cross-lingual crossword (Results)

A small experiment with Japanese tested several configurations:

- ① English Definition → Hiragana
a member of the genus *Canis* that has been domesticated by man since prehistoric times → いぬ
- ② English Lemmas → Hiragana
domestic dog, *Canis familiaris* → いぬ
- ③ Kanji → Hiragana
犬 → いぬ
- ④ Kanji → English
犬 → dog
- ⑤ Hiragana → English
いぬ → dog

All showed an improvement in vocabulary retention except (5), but the sample size was small (Tan, 2012).

- In theory the crossword is language independent
 - ▶ In practice some languages are not good for crosswords (Chinese has an average word length of 2 characters)
 - ▶ transliterations only exist for some languages and some have multiple transliterations:
hiragana/katakana/romaji; with/without vowels; ...
 - ▶ the crossword code does not work for right-to-left scripts
- Wordnet definitions are not ideal clues
 - ▶ Rewrite simple definitions for language learning
 - ▶ Use different clues (translations, pictures, examples?)
- Learners should learn words in context

Crosslingual Crossword – ToDo

- Combine with hierarchy and frequency
 - ▶ Crossword of fruit (hyponyms), with high frequency
- Combine with graded vocabulary lists; lesson plans
 - ▶ Grade n only, Grade $< n$, ...
- Gamify it more
 - ▶ (Singapore) students like to win points
 - ▶ Show highest score, fastest, ...
- Allow feedback — there are still mistakes in the wordnets
- Add more games (hangman)
- Show examples when they get a word right (from dictionary or corpus)

Using an HPSG implementation to teach writing skills Dan Flickinger (Pat et al., 2014)

Language Arts and Writing

- For monolingual English speakers grades 2–6
- Now used in classrooms in public schools
- Goal: to help students improve writing skills
- Automated exercise-based course with immediate feedback

Exercise design for sentence composition

- Present a few sentences of context
- Ask a question
- Provide a set of (fully inflected) words, listed by part-of-speech
- Ask the student to compose an answer as a complete sentence
- Evaluate the answer, and if incorrect, identify error where possible
- Ask the student to try again once

An example from Grade 5

Abigail didn't want to go hiking with her parents because she felt too tired and wanted to rest instead.

Why didn't Abigail want to go hiking?

Verb	Preposition	Noun	Adjective	Pronoun	Conjunction	Contraction	Adverb
want was were go hiking hike	to	Abigail hike	tired hungry sick	she	because	didn't	too
>							

Please select words from the lists (click to select, drag outside to deselect)

There are a few possible answers

She was tired.

She was too tired.

She was too tired to.

She was too tired to go.

She was too tired to hike.

She was too tired to go hike.

She was too tired to go hiking.

She didn't because she was too tired.

She didn't want to because she was too tired.

She didn't want to go because she was too tired.

She didn't want to hike because she was too tired.

She didn't want to go hike because she was too tired.

She didn't want to go hiking because she was too tired.

Really, there are quite a few possible answers

She didn't because she was tired.
She didn't because she was too tired to.
She didn't because she was too tired to go.
She didn't because she was too tired to go hike.
She didn't because she was too tired to go hiking.
She didn't because she was too tired to hike.
She didn't want to because she was tired.
She didn't want to because she was too tired to.
She didn't want to because she was too tired to go.
She didn't want to because she was too tired to go hike.
She didn't want to because she was too tired to go hiking.
She didn't want to because she was too tired to hike.
She didn't want to go because she was tired.
She didn't want to go because she was too tired to.
She didn't want to go because she was too tired to hike.
She didn't want to go because she was too tired to go hike.
She didn't want to go because she was too tired to go hiking.
She didn't want to go because she was too tired to hike.
She didn't want to go hike because she was tired.
She didn't want to go hike because she was too tired to.
She didn't want to go hike because she was too tired to go.
She didn't want to go hike because she was too tired to go hike.
She didn't want to go hike because she was too tired to go hiking.
She didn't want to go hike because she was too tired to hike.
She didn't want to go hiking because she was tired.
She didn't want to go hiking because she was too tired to.
She didn't want to go hiking because she was too tired to go.
She didn't want to go hiking because she was too tired to go hike.
She didn't want to go hiking because she was too tired to go hiking.
She didn't want to go hiking because she was too tired to hike.

More than you might think

She didn't want to hike because she was tired.
She didn't want to hike because she was too tired to.
She didn't want to hike because she was too tired to go.
She didn't want to hike because she was too tired to go hike.
She didn't want to hike because she was too tired to go hiking.
She didn't want to hike because she was too tired to hike.
She didn't because she was tired.
She didn't because she was too tired to.
She didn't because she was too tired to go.
She didn't because she was too tired to go hike.
She didn't because she was too tired to go hiking.
She didn't because she was too tired to hike.
She didn't want to because she was tired.
She didn't want to because she was too tired to.
She didn't want to because she was too tired to go.
She didn't want to because she was too tired to go hike.
She didn't want to because she was too tired to go hiking.
She didn't want to because she was too tired to hike.
She didn't want to go because she was tired.
She didn't want to go because she was too tired to.
She didn't want to go because she was too tired to hike.
She didn't want to go because she was too tired to go hike.
She didn't want to go because she was too tired to go hiking.
She didn't want to go because she was too tired to hike.
She didn't want to go hike because she was tired.
She didn't want to go hike because she was too tired to.
She didn't want to go hike because she was too tired to go.
She didn't want to go hike because she was too tired to go hike.
She didn't want to go hike because she was too tired to go hiking.
She didn't want to go hike because she was too tired to hike.
She didn't want to go hiking because she was tired.
She didn't want to go hiking because she was too tired to.

Probably more than you want to enumerate

Abigail didn't because she was tired.
Abigail didn't because she was too tired to.
Abigail didn't because she was too tired to go.
Abigail didn't because she was too tired to go hike.
Abigail didn't because she was too tired to go hiking.
Abigail didn't because she was too tired to hike.
Abigail didn't want to because she was tired.
Abigail didn't want to because she was too tired to.
Abigail didn't want to because she was too tired to go.
Abigail didn't want to because she was too tired to go hike.
Abigail didn't want to because she was too tired to go hiking.
Abigail didn't want to because she was too tired to hike.
Abigail didn't want to go because she was tired.
Abigail didn't want to go because she was too tired to.
Abigail didn't want to go because she was too tired to hike.
Abigail didn't want to go because she was too tired to go hike.
Abigail didn't want to go because she was too tired to go hiking.
Abigail didn't want to go hike because she was tired.
Abigail didn't want to go hike because she was too tired to.
Abigail didn't want to go hike because she was too tired to go.
Abigail didn't want to go hike because she was too tired to go hike.
Abigail didn't want to go hike because she was too tired to go hiking.
Abigail didn't want to go hike because she was too tired to hike.
Abigail didn't want to go hiking because she was tired.
Abigail didn't want to go hiking because she was too tired to.
Abigail didn't want to go hiking because she was too tired to go.
Abigail didn't want to go hiking because she was too tired to go hike.
Abigail didn't want to go hiking because she was too tired to go hiking.
Abigail didn't want to go hiking because she was too tired to hike.
Abigail didn't want to hike because she was tired.
Abigail didn't want to hike because she was too tired to.
Abigail didn't want to hike because she was too tired to go.

Many, many more

Because she was tired, she didn't.
Because she was too tired to, she didn't.
Because she was too tired to go, she didn't.
Because she was too tired to go hike, she didn't.
Because she was too tired to go hiking, she didn't.
Because she was too tired to hike, she didn't.
Because she was tired, she didn't want to.
Because she was too tired to, she didn't want to.
Because she was too tired to go, she didn't want to.
Because she was too tired to go hike, she didn't want to.
Because she was too tired to go hiking, she didn't want to.
Because she was too tired to hike, she didn't want to.
Because she was tired, she didn't want to go.
Because she was too tired to, she didn't want to go.
Because she was too tired to hike, she didn't want to go.
Because she was too tired to go hike, she didn't want to go.
Because she was too tired to go hiking, she didn't want to go.
Because she was too tired to hike, she didn't want to go.
Because she was tired, she didn't want to go hike.
Because she was too tired to, she didn't want to go hike.
Because she was too tired to go, she didn't want to go hike.
Because she was too tired to go hike, she didn't want to go hike.
Because she was too tired to go hiking, she didn't want to go hike.
Because she was too tired to hike, she didn't want to go hike.

Language is infinite, remember

Because Abigail was tired, she didn't want to go hike.
Because Abigail was tired, she didn't want to go hiking.
Because Abigail was tired, she didn't want to go.
Because Abigail was tired, she didn't want to hike.
Because Abigail was tired, she didn't want to.
Because Abigail was tired, she didn't.
Because Abigail was too tired to go hike, she didn't want to go hike.
Because Abigail was too tired to go hike, she didn't want to go hiking.
Because Abigail was too tired to go hike, she didn't want to go.
Because Abigail was too tired to go hike, she didn't want to hike.
Because Abigail was too tired to go hike, she didn't want to.
Because Abigail was too tired to go hike, she didn't.
Because Abigail was too tired to go hiking, she didn't want to go hike.
Because Abigail was too tired to go hiking, she didn't want to go hiking.
Because Abigail was too tired to go hiking, she didn't want to go.
Because Abigail was too tired to go hiking, she didn't want to hike.
Because Abigail was too tired to go hiking, she didn't want to.
Because Abigail was too tired to go hiking, she didn't.
Because Abigail was too tired to go, she didn't want to go hike.
Because Abigail was too tired to go, she didn't want to go hiking.
Because Abigail was too tired to go, she didn't want to hike.
Because Abigail was too tired to go, she didn't want to.
Because Abigail was too tired to go, she didn't.
Because Abigail was too tired to hike, she didn't want to go hike.
Because Abigail was too tired to hike, she didn't want to go hiking.
Because Abigail was too tired to hike, she didn't want to go.
Because Abigail was too tired to hike, she didn't want to hike.
Because Abigail was too tired to hike, she didn't want to.
Because Abigail was too tired to hike, she didn't.
Because Abigail was too tired to, she didn't want to go hike.
Because Abigail was too tired to, she didn't want to go hiking.
Because Abigail was too tired to, she didn't want to go.

There are quite a few possible answers

Because she was tired, Abigail didn't want to go hike.
Because she was tired, Abigail didn't want to go hiking.
Because she was tired, Abigail didn't want to go.
Because she was tired, Abigail didn't want to hike.
Because she was tired, Abigail didn't want to.
Because she was tired, Abigail didn't.
Because she was too tired to go hike, Abigail didn't want to go hike.
Because she was too tired to go hike, Abigail didn't want to go hiking.
Because she was too tired to go hike, Abigail didn't want to go.
Because she was too tired to go hike, Abigail didn't want to hike.
Because she was too tired to go hike, Abigail didn't want to.
Because she was too tired to go hike, Abigail didn't.
Because she was too tired to go hiking, Abigail didn't want to go hike.
Because she was too tired to go hiking, Abigail didn't want to go hiking.
Because she was too tired to go hiking, Abigail didn't want to go.
Because she was too tired to go hiking, Abigail didn't want to hike.
Because she was too tired to go hiking, Abigail didn't want to.
Because she was too tired to go hiking, Abigail didn't.
Because she was too tired to go, Abigail didn't want to go hike.
Because she was too tired to go, Abigail didn't want to go hiking.
Because she was too tired to go, Abigail didn't want to hike.
Because she was too tired to go, Abigail didn't want to.
Because she was too tired to go, Abigail didn't.
Because she was too tired to hike, Abigail didn't want to go hike.
Because she was too tired to hike, Abigail didn't want to go hiking.
Because she was too tired to hike, Abigail didn't want to go.
Because she was too tired to hike, Abigail didn't want to hike.
Because she was too tired to hike, Abigail didn't want to.
Because she was too tired to hike, Abigail didn't.
Because she was too tired to, Abigail didn't want to go hike.
Because she was too tired to, Abigail didn't want to go hiking.
Because she was too tired to, Abigail didn't want to go.

And this isn't all of them

Because Abigail was tired, Abigail didn't.
Because Abigail was too tired to, Abigail didn't.
Because Abigail was too tired to go, Abigail didn't.
Because Abigail was too tired to go hike, Abigail didn't.
Because Abigail was too tired to go hiking, Abigail didn't.
Because Abigail was too tired to hike, Abigail didn't.
Because Abigail was tired, Abigail didn't want to.
Because Abigail was too tired to, Abigail didn't want to.
Because Abigail was too tired to go, Abigail didn't want to.
Because Abigail was too tired to go hike, Abigail didn't want to.
Because Abigail was too tired to go hiking, Abigail didn't want to.
Because Abigail was too tired to hike, Abigail didn't want to.
Because Abigail was tired, Abigail didn't want to go.
Because Abigail was too tired to, Abigail didn't want to go.
Because Abigail was too tired to hike, Abigail didn't want to go.
Because Abigail was too tired to go hike, Abigail didn't want to go.
Because Abigail was too tired to go hiking, Abigail didn't want to go.
Because Abigail was tired, Abigail didn't want to go hike.
Because Abigail was too tired to, Abigail didn't want to go hike.
Because Abigail was too tired to go, Abigail didn't want to go hike.
Because Abigail was too tired to go hike, Abigail didn't want to go hike.
Because Abigail was too tired to go hiking, Abigail didn't want to go hike.
Because Abigail was too tired to hike, Abigail didn't want to go hike.
Because Abigail was tired, Abigail didn't want to go hiking.
Because Abigail was too tired to, Abigail didn't want to go hiking.
Because Abigail was too tired to go, Abigail didn't want to go hiking.
Because Abigail was too tired to go hike, Abigail didn't want to go hiking.
Because Abigail was too tired to go hiking, Abigail didn't want to go hiking.
Because Abigail was too tired to hike, Abigail didn't want to go hiking.
Because Abigail was tired, Abigail didn't want to hike.
Because Abigail was too tired to, Abigail didn't want to hike.
Because Abigail was too tired to go, Abigail didn't want to hike.

How can we mark these?

- Too many to enumerate by hand
- A probabilistic grammar does not give clear, correct judgments.
- (Re)Use a hand written, computational grammar (DELPH-IN www.delph-in.net)
 - ▶ English Resource Grammar (ERG: Flickinger, 2000, 2011)
 - ▶ New demo here (use UW server):
<http://delph-in.github.io/delphin-viz/demo/>
- Adapt [mal-rule](#) approach to accept mild ungrammaticality (Schneider and McCoy, 1998; Bender et al., 2004)
- Parse each novel input and return its derivation tree
- Check for root [robust](#):
 - ▶ If not, then it is grammatical: [Well Done!](#)
 - ▶ If robust then look up robustness symbol in error code table (grade-specific)
- Present appropriate message to student

English Resource Grammar (ERG)

- 7,000 types in multiple-inheritance monotonic hierarchy
- 975 leaf lexical types
- 35,000 manually constructed lexemes
- 200 syntactic rules
- 70 morphological rules (inflection and derivation)
- Online demo: lingo.stanford.edu/erg

Changes to the general-purpose grammar

- Extensions

- ▶ Mal-rules for inflection, syntax
e.g. bare singular NP, or bare 3sg-present verb
- ▶ Mal-types for lexicon
e.g. subj-equi with base VP: **Ricky likes go to the park*

- Reductions to avoid some ambiguity

- ▶ Rules
For example, block noun-noun compound rules
**Ricky's closet toys are in the closet*
- ▶ Lexicon
For example, block the verb flower
Art and science flowered during the 17th century.

Sample Errors

Answer is not grammatical.
Your answer is not a complete sentence.
Your answer is grammatical but awkward.
Your answer cannot be a question.
You are missing an article before the word *X*.
Remember to use “an” only before a vowel.
Don’t use “a” before a vowel.
Don’t put “the” before a name.
You are missing the preposition “on” before *X*.
You are missing “to” before *X*.
Don’t put “to” before *X*.
The verb *X* needs an object.
You are missing a noun.
You have an extra noun in your answer.
Use an adverb like “well” or “poorly” instead of “bad”.
Use “its” instead of “it’s” to show ownership.
Remember to use “this” only before a singular noun.
Don’t use “did/does/do” in your answer.
You have the wrong form of the verb.
Your subject doesn’t agree with the verb *X*.



He are be was dog
In the park
They saw in the lake a duck
Did she go to the beach
She went to house
He ate an sandwich
She saw a owl
The Katherine eats breakfast
They watched movies Tuesday
She told her brother borrow her book
She let her brother to borrow her book
She let borrow her book
His chased the cat
The children ate a lunch pizza
He did bad on his test
The dog is in it’s house
He said he like this mittens
The dog did go for a walk
The boy be late for school
Alex write a letter

- Large scale testing found using the system improved student's scores
- Now being developed by a educational company
- Core technology open-source
- Similar on-line system for Norwegian:
A Norwegian Grammar Sparrer (Hellan et al., 2013)
 - ▶ https://typecraft.org/tc2wiki/A_Norwegian_Grammar_Sparrer
 - ▶ Feedback in multiple languages

Second Language Learning (work at NTU)

- Multilingual (Chinese for English speakers)
- Targeting university level L2 learners
- Integrated with wordnets (gradual introduction of vocabulary)
- Goal: provide drill and feedback
 - ▶ Randomized examples with controlled vocabulary
 - ▶ Gradual increase in difficulty
- Pairwise Semantic-based Translation/Disambiguation to pinpoint errors
- Automatic collection of Learner corpora
(Ungrammatical input + tree-banked, sense-tagged corrected sentence)
Controlled experiment using NTU students of Mandarin L2

The Intelligent Tutor (Example)

Student:	That dog like the cat happy.
	<p>Hmm... something is wrong with your sentence. Did you mean any of these?</p> <p>A. 那只狗和猫一样高兴。 [That dog, like the cat, is happy.]*</p> <p>B. 那只狗喜欢猫高兴。 [That dog likes the cat happy.]*</p> <p>C. 那只狗喜欢高兴的猫。 [That dog likes the happy cat.]*</p>
Student:	C. 那只狗喜欢高兴的猫。
	<p>Ok! Then I believe you forgot to conjugate the verb 'to like'. Also, remember that an adjective must come before the noun it's modifying. Please try again!</p>

* The English translation is what the system thinks is correct, but it is not shown.

Figure: Ungrammatical Ambiguity for Chinese speakers learning English

An *Intelligent* Language Tutor

- *Intelligent* because... (NLP enabled)
 - ▶ bilingual (Semantic-based MT between English and Mandarin)
 - ▶ rich syntactic knowledge about both languages (Comp. Grammars)
 - ▶ knowledge about common ungrammatical structures (and why!)
 - ▶ knows how to correct ungrammatical sentences
- Language Tutor
 - ▶ knows the language curriculum
 - ★ step-by-step lexicon introduction
 - ★ step-by-step grammar introduction
 - ▶ knows each and every student (individually & by class)
 - ★ their strengths and weaknesses
 - ★ their progress through the syllabus
 - ▶ coaches students in their sentence composition
 - ★ giving tips on how to correct ungrammatical sentences

Our Focus: Chinese Learners (English natives)

Behind an *Intelligent* Chinese Tutor:

- survey syllabus used by early levels of Chinese L2
 - ▶ word meanings and syntactic structures (textbooks, official exams)
- survey most common writing mistakes made by Chinese L2 learners
- integrate this knowledge in existing NLP tools
 - ▶ build on our current research
 - ▶ use a small but precise Chinese grammar [Zhong](#) (Fan et al., 2015)
- develop and test an online tutor (i.e. web application)

Evaluation:

- **intrinsic**: ability to diagnose and correct naturally occurring grammatical mistakes by learners of Chinese L2 (Corpus Study)
- **extrinsic**: controlled experiment of a blended learning environment of Chinese L2 (Centre for Modern Languages, NTU: 120 students)

Our Focus: Chinese Learners by English Speakers

What we are building:

- Chinese-English bilingual web-based tutor for learners of Chinese
- Learner Corpus, produced and annotated automatically with syntactic and semantic information (from the student interaction)
- Written Text Difficulty Grader, that uses information on lexicon and syntactic complexity to assert a language difficulty level for Chinese texts (spin-off, prototype)

Future Work:

- Scale-up: Our design will allow the tutor to be expanded to other levels and languages (Japanese, Indonesian, ...)
- Personalize more: topicalize examples based on interests

Summary

- We are developing a system that will make it possible for the computer to drill students and give feedback
 - ▶ The student can practice as much as they want
 - ▶ The exercises will practice the most needed phenomena
 - ▶ The teacher can see where the problems are
 - ▶ The teacher can spend more time on higher level issues
- Acknowledgments This research is partially supported by the Singapore MOE TRF grant *Syntactic Well-Formedness Diagnosis and Error-Based Coaching in Computer Assisted Language Learning using Machine Translation Technology*.

References I

- Emily M Bender, Dan Flickinger, Stephan Oepen, Annemarie Walsh, and Timothy Baldwin. 2004. Arboretum: Using a precision grammar for grammar checking in call. In *Instil/icall symposium 2004*.
- Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352–1362. Sofia. URL <http://aclweb.org/anthology/P13-1133>.
- Francis Bond, Hitoshi Isahara, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2010. Japanese WordNet 1.0. In *16th Annual Meeting of the Association for Natural Language Processing*, pages A5–3. Tokyo.
- 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.

References II

- Ulrich Callmeier. 2000. PET - a platform for experimentation with efficient HPSG processing techniques. *Natural Language Engineering*, 6(1):99–108.
- Zhenzhen Fan, Sanghoun Song, and Francis Bond. 2015. Building Zhong [|], a Chinese HPSG meta-grammar. In *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar (HPSG 2015)*, pages 97–110.
- Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
- Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28. (Special Issue on Efficient Processing with HPSG).
- Dan Flickinger. 2011. Accuracy vs. robustness in grammar engineering. In Emily M. Bender and Jennifer E. Arnold, editors, *Language from a Cognitive Perspective: Grammar, Usage, and Processing*, pages 31–50. CSLI, Stanford.

References III

- Lars Hellan, Tore Bruland, Elias Aamot, and Mads Hustad Sandøy. 2013. A grammar sparrer for norwegian. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. URL <http://emmtee.net/oe/nodalida13/conference/83.pdf>.
- Suppes Pat, T. Liang, E. E. Macken, and Dan Flickinger. 2014. Positive technological and negative pre-test-score effects in a four-year assessment of low socioeconomic status K-8 student learning in computer-based math and language arts courses. *Computers & Education*, 71:23–32.
- Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. 2011. Methodology and construction of the Basque wordnet. *Language Resources and Evaluation*, 45(2):121–142.

References IV

- David Schneider and Kathleen F. McCoy. 1998. Recognizing syntactic errors in the writing of second language learners. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics: COLING/ACL-98*, pages 1198–1204. URL <http://dx.doi.org/10.3115/980432.980765>.
- 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. URL <http://www.lrec-conf.org/proceedings/lrec2010/summaries/848.html>.

References V

- 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.
- Jeanette Yi Wen Tan. 2012. *Automatic Generation of Multilingual Crossword Puzzles with WordNet*. Final year project, Linguistics and Multilingual Studies, Nanyang Technological University.

- 頭脳循環を加速する戦略的国際研究ネットワーク推進プログラム
危機言語・少数言語を中心とする循環型調査研究のための
機動的国際ネットワーク構築