



# The HINOKI (檜) Project

## ~Toward Text Understanding

---

Francis Bond\*, Sanae Fujita\*,  
Chikara Hashimoto\*\*, Eric Nichols\*\*\*,  
Takaaki Tanaka\*

\*NTT Communication science lab.

\*\*Kyoto University, \*\*\*NAIST

# HINOKI Project: Motivation

- Make computers clever
- Natural language processing based on both **structural** and **lexical** semantics.
- Ex) Information Retrieval, QA, etc...

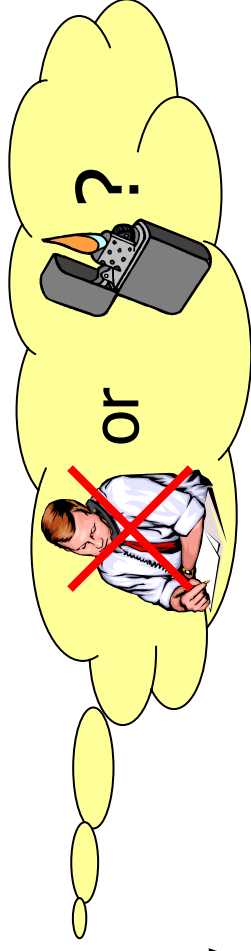
- ライターを雇いたい。

*I want to hire a **writer**.*



- ライターを買いたい。

*I want to buy a **lighter**.*

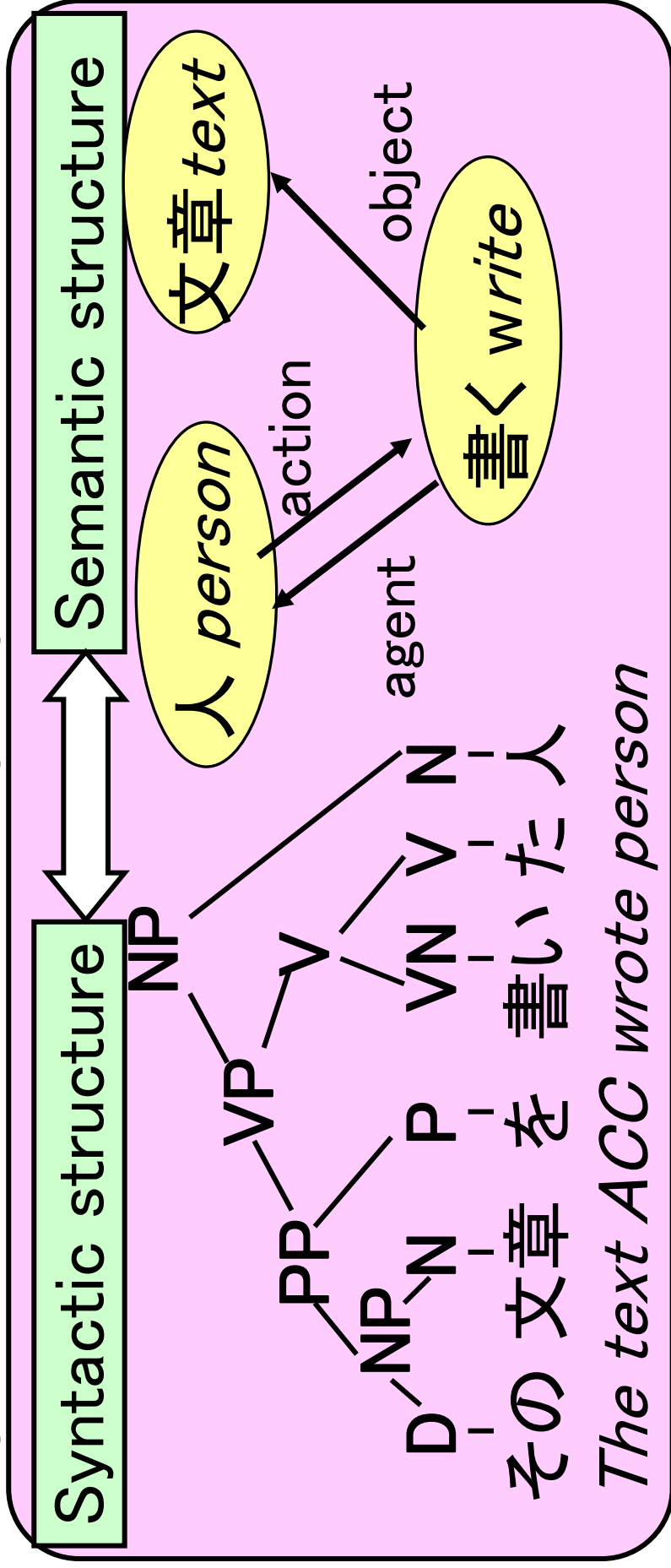


# What's needed? (1)

## ~Grammar (Jacy) & Treebank

- Using Grammar Jacy, DELPH-IN tools (LKB. PET, [incr tsdb()]).

Ex) Definition Sentences of ライター (*writer*)



# What's needed? (2)

## ~Dictionary (Lexeed)



- Lexeed (Seed of Lexicon)
- The most familiar words of Japanese
  - Familiarity is estimated by psychological experiments
- 28,000 words, 46,000 senses
- Covers 75% of tokens in a typical newspaper.
- Self-contained dictionary
- Has def. sentences and at least one ex. for each sense.

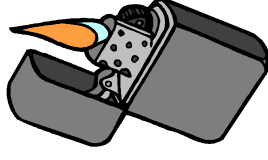
Ex) Definitions

ライター<sup>2</sup> (*writer*) [noun]  
(2-1) a **person** who has written the writing.  
author.



ライター<sup>1</sup> (*lighter*) [noun]

(1-1) light. igniter. a **device** for lighting or igniting fuel, especially cigarette.  
ex) He lit the cigarette with his lighter.



# What's needed? (3)

## ～Sensebank

- We tagged some corpora using the Lexseed senses.

ライターに依頼した。

*I asked a writer.*

ライターを買った。

*I bought a lighter.*

ライターが必要だ。

*I need a writer/lighter.*

ライター<sub>2</sub> (*writer*) [noun]

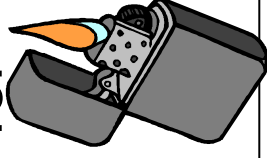
(2-1) a **person** who has written the writing.  
author.



Definition of Lexseed

ライター<sub>1</sub> (*lighter*) [noun]

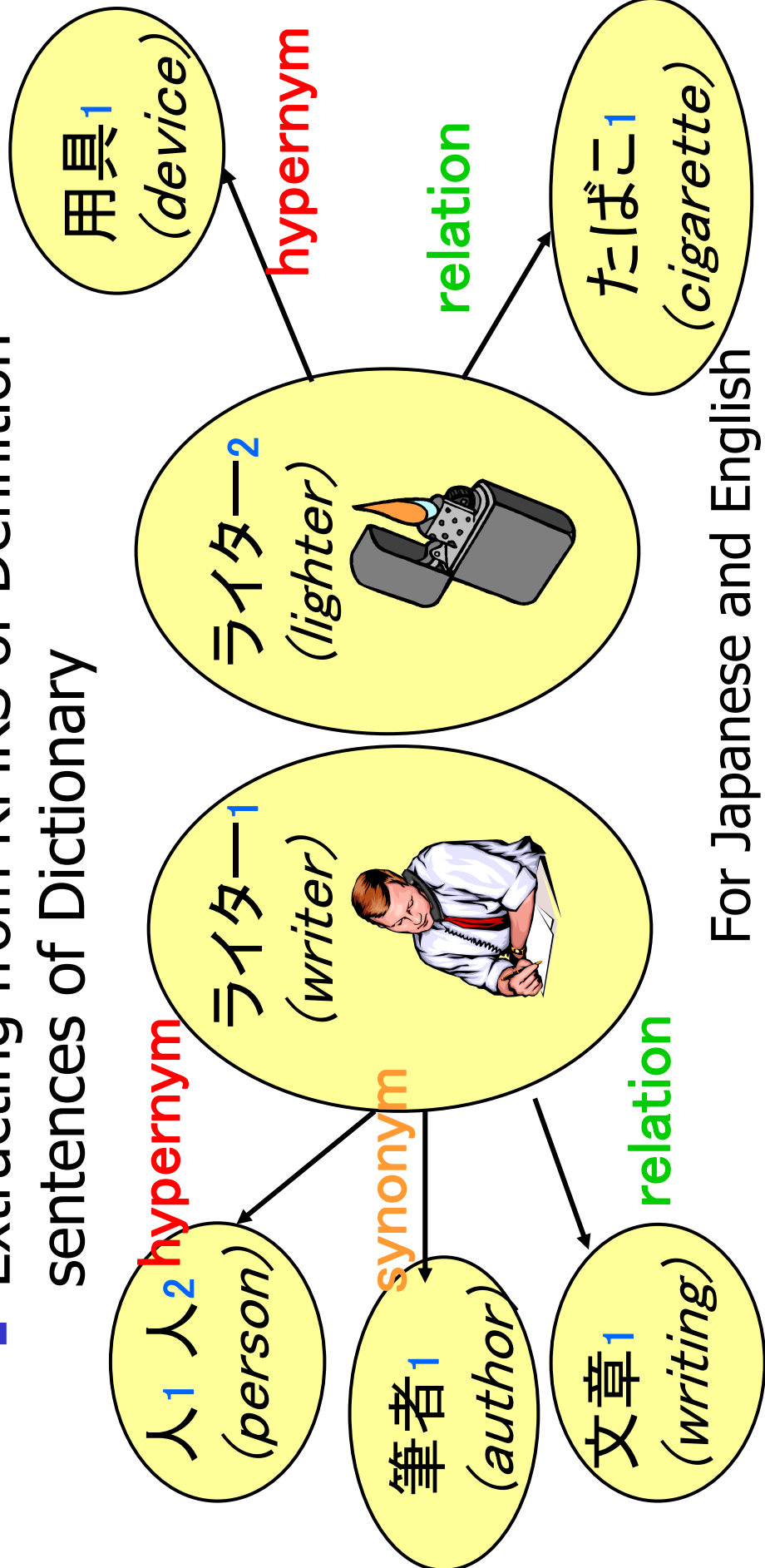
(1-1) light. igniter. a **device** for lighting or igniting fuel, especially cigarette.



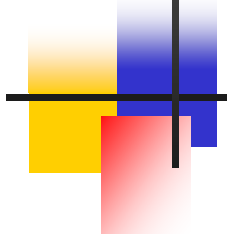
# What's needed? (4)

## ~Ontology

- Extracting from RMRS of Definition sentences of Dictionary

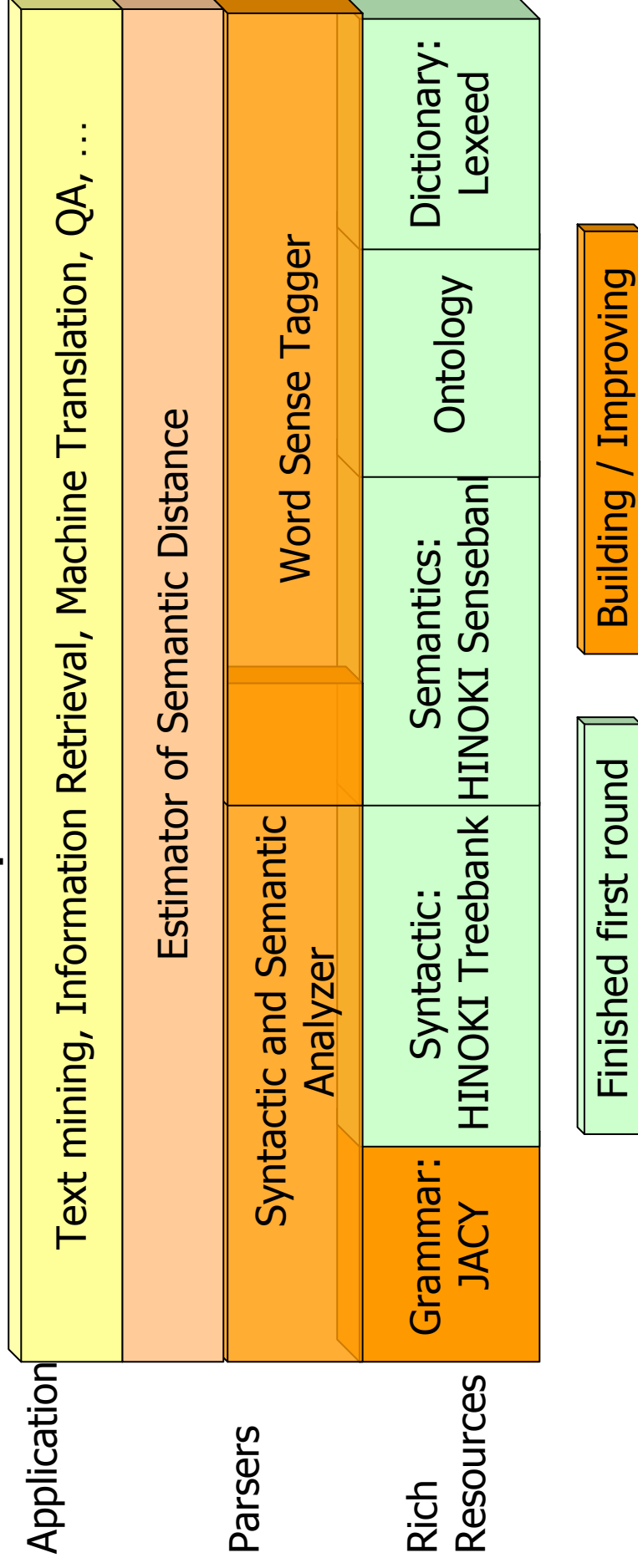


For Japanese and English



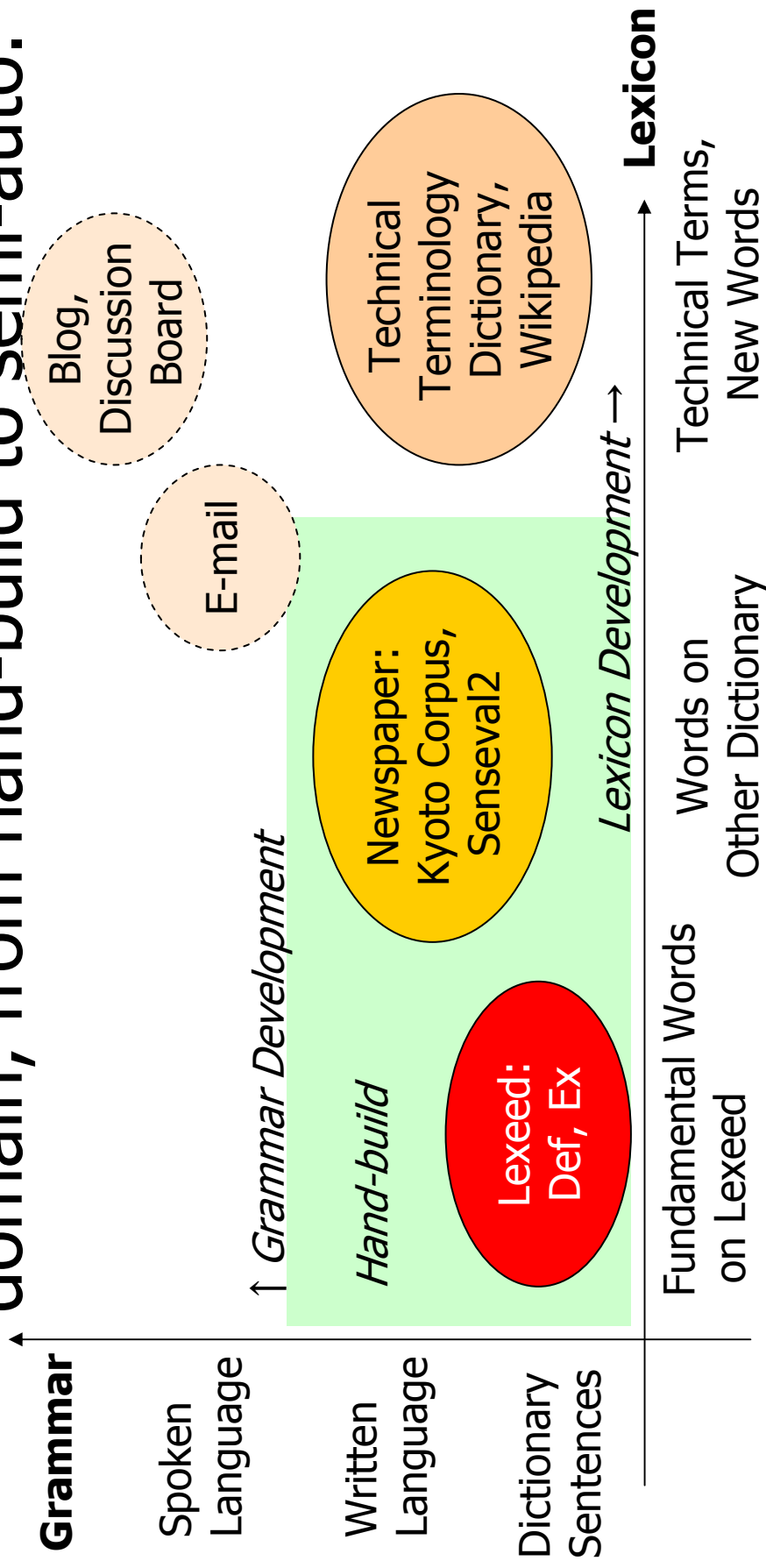
# HINOKI Project: Status

- Constructing resources and making fundamental parsers



# HINOKI Project: Domain

- Expanding from closed world to open domain, from hand-build to semi-auto.





# HINOKI Treebank: Status

- Finished first round for dictionary
  - 75% of def. and ex. sentences have correct trees
- Now annotating newspaper articles
  - 47% (Kyoto) and 25% (Senseval 2) are parsed, of those around 50% are correct
  - Need more Grammar and Lexicon development

| Corpus             | Sentences | Words   | have Correct Tree |
|--------------------|-----------|---------|-------------------|
| Lexceed Def.       | 75,000    | 691,000 | 54,000            |
| Lexceed Ex.        | 45,000    | 499,000 | 35,000            |
| News (Kyoto Univ.) | 38,000    | 969,000 | *4,600            |
| News (Senseval 2)  | 36,000    | 888,000 | *4,400            |

# HINOKI Sensebank: Status

- Biggest Japanese sensebank
  - Senseval2 (Japanese): 150,000
  - SemCor (English): 230,000

In total,  
819,000

| Corpus          | Words (No.) | Tagged  | agreement |
|-----------------|-------------|---------|-----------|
| Lexeed Def.     | 691,000     | 199,000 | 78.7      |
| Lexeed Ex.      | 499,000     | 127,000 | 82.0      |
| News (Kyoto)    | 969,000     | 224,000 | 83.2      |
| News(Senseval2) | 888,000     | 269,000 | 83.3      |



# HINOKI Ontology: Status

- Created from several dictionaries
- Compared with existing thesaurus
  - Consistency: ratio of extracted relations which are consistent with existing thesaurus (Goi-taikai, WordNet)

| Dictionary    | Lang. | Word No. | Sense No. | Link No. | consistency |
|---------------|-------|----------|-----------|----------|-------------|
| Lexeed        | Ja    | 28,000   | 46,000    | 77,000   | 57%*        |
| Iwanami(RWCP) | Ja    | 60,000   | 86,000    | 142,000  | 60%         |
| GCIDE         | En    | 130,000  | 171,000   | 78,000   | 36%**       |

\* : Evaluated by hand: 89%

\*\* : used only hypernym 100,000 sentences (12,000 words, 36,000 senses)



# Future Work

---

- Automate analysis
  - Word sense tagger (prototype is made)
  - Parse ranking
  - Combining them
- Expand the system and resources
  - Expand the coverage of the Grammar and Dictionary
  - Enable the system to treat unknown words