Lexical Knowledge Acquisition

Towards a Continuous and Flexible Representation of the Lexicon

Pierre Marchal and Thierry Poibeau

CNRS & INALCO

July 11, 2016

Problems with Static Representations of the Lexicon

Static representations of word meaning (like in traditional paper dictionaries) are not appropriate

- Semantic categories have fuzzy boundaries and thus the number of word meanings per lexical item is to a large extent arbitrary [Tuggy, 1993]
- Lexical entries in traditional dictionaries overlap and different word meanings can be associated with a same example [Erk and McCarthy, 2009]

Continuous Representations of Word Meaning

- Continuous models encode intuitively valid and cognitively plausible principles
  - Semantic similarity is relative
  - It is context-sensitive
  - It depends on multiple-cue integration
- However, these models have not been used for representing meaning in dictionaries written for humans because of their complexity

Our Framework

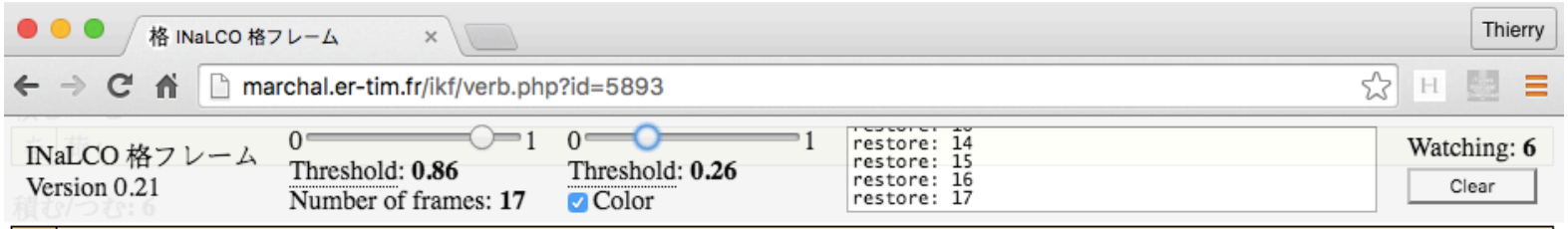
We propose to develop continuous representations that are usable by humans

- We focus on verbs since verbs offer the most complex syntactic and semantic behaviors
- We focus on Japanese because of its complex case system
- The system could be easily adapted to English by substituting prepositions to case markers

Our Main Hypotheses

- Lexical entries have fuzzy boundaries $\Rightarrow$ Continuum between ambiguity and vagueness [Tuggy 1993]
- Argument / modifier distinction also has fuzzy boundaries $\Rightarrow$ Continuum between arguments and modifiers [Manning 2003]
- The end-user must be able to navigate the data through these two dimensions
- Cf. previous works in lexical acquisition [Korhonen 2002 ; Fabre et Bourigault 2008].

A Snapshot of the System



を	土嚢；引当金；徳；発射機；レンガ；調教；スクラップ；核弾頭；頭金；用材；
が	オルフェーブル；

積む/つむ: 7

を	経験；実績；体験；ノウハウ；
で	石清水八幡宮；ドイツ；ホテル；クラブ；遠征；豪州；欧州；部門；コース；サンダース；ヘルタ；海外；下；
から	出会い；
として	社員；プログラマー；産婦人科医；
に	記者共；

積む/つむ: 8

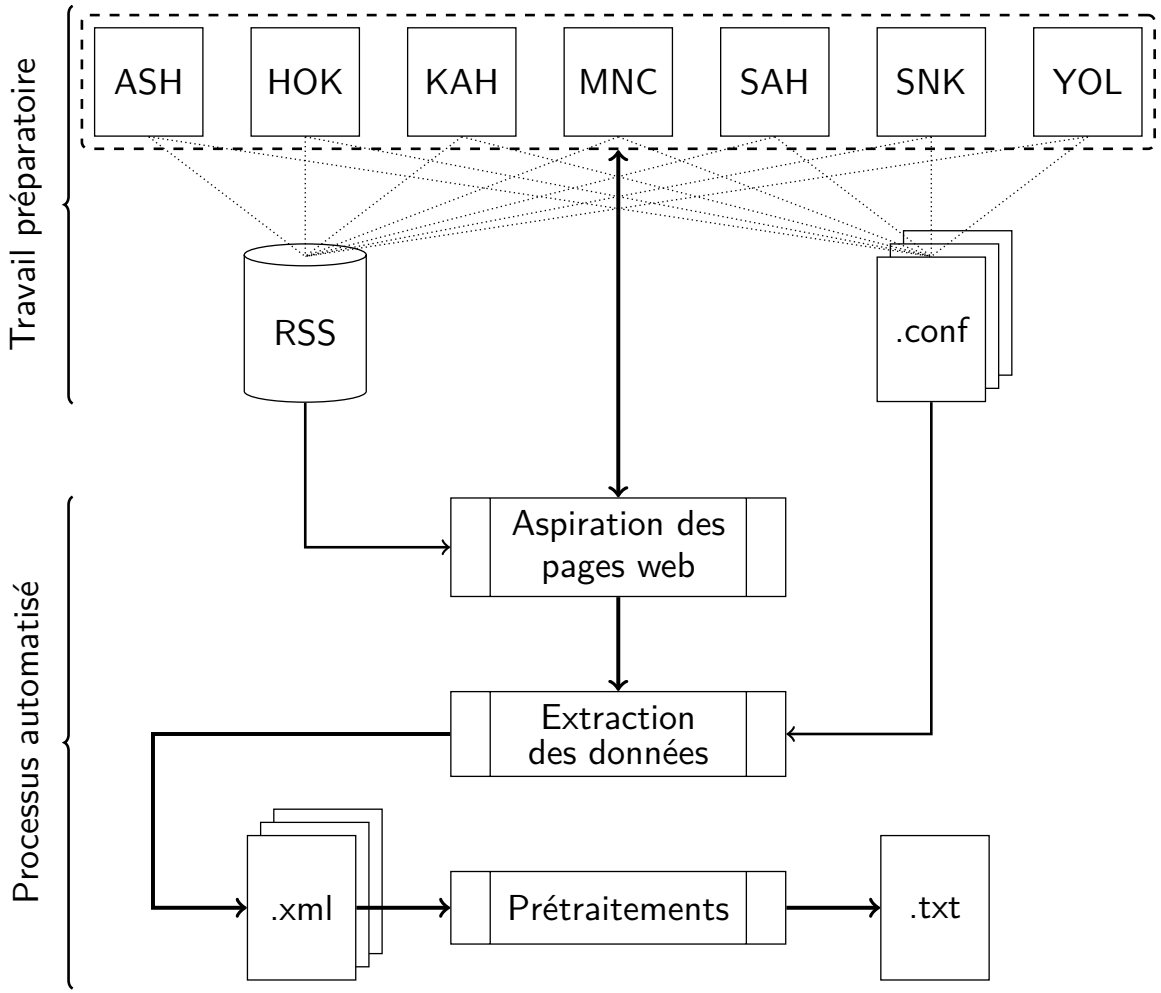
を	研鑽；修業；研さん；鍛錬；トレーニング；修行；キャリア；練習；稽古；訓練；修練；レッスン；特訓；研修；実戦；自主トレ；努力；リハビリ；勉強；
で	学問所；延暦寺；音楽院；リンク；ドイツ；ホテル；豪州；米国；欧州；ジム；コース；ブートキャンプ；海事局；トリニティ・カレッジ・オブ・ミュージック；東福寺；スポ；下；塗装店；
として	弟子；デザイナー；書評家；坊さん；船大工；
と	アリスター・オーフレイムら；鷲ら；
とともに	ゆか；

Description of the Analysis Process

The different steps of the process

- Collect a large corpus from the Web
- Parse the corpus with a non lexicalized parser
- Extract verbs along with their subcategorization frames (complements and case markers)
- Calculate verb similarity based on their usage

Corpus collection



Acquisition and pre-processing of textual data

Collection of newspaper articles

- From 7 newspaper websites
- during more than 32 months (Nov. 2011–Aug 2014)
- $\sim$ 787 000 articles

Size of the corpus

- 7 million sentences
- 186 millions tokens
- 294 millions characters

Subcategorization Frame Acquisition

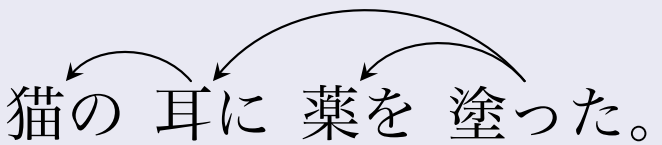
Input

猫の 耳に 薬を 塗った。

Output

Subcategorization Frame Acquisition

Input

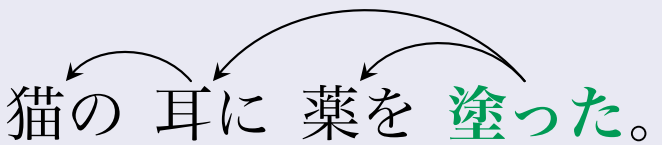


Output

1. Syntactic analysis (CaboCha 0.68)

Subcategorization Frame Acquisition

Input



Output

塗る
apply

1. Syntactic analysis (CaboCha 0.68)
2. Extraction of verbal predicative structures
   - Verb identification

Subcategorization Frame Acquisition

Input

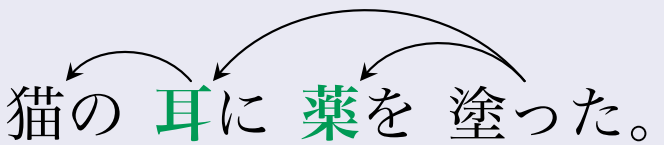
Output

塗る	:に	:を
apply	:*ni*	:*wo*

1. Syntactic analysis (CaboCha 0.68)
2. Extraction of verbal predicative structures
   - Verb identification
   - Case marker identification

Subcategorization Frame Acquisition

Input



Output

塗る　耳:に　薬:を
apply　ear:*ni*　ointment:*wo*

1. Syntactic analysis (CaboCha 0.68)
2. Extraction of verbal predicative structures
   - Verb identification
   - Case marker identification
   - Lexical head identification

Subcategorization Frame Acquisition

Input	Output
猫の 耳に 薬を 塗った。	塗る　耳:に　薬:を apply　ear:*ni*　ointment:*wo*

1. Syntactic analysis (CaboCha 0.68)
2. Extraction of verbal predicative structures
   - Verb identification
   - Case marker identification
   - Lexical head identification
3. Filtering (errors, etc.)

Grouping related subcategorization frames

Input

切る　クリスマス:まで　<NUM>カ月:を	切る　美容師:が　髪:を
切る　ボタン:で　スイッチ:を	切る　業者:が　電源:を
切る　ガラス:で　従業員:が　頭部:を	切る　告示:まで　<NUM>カ月:を
切る　髪:を	切る　ガラス:で　手:を
切る　スイッチ:を	切る　教諭:が　髪:を　ハサミ:で

Output

Grouping related subcategorization frames

Input

切る　クリスマス:まで　<NUM>カ月:を	切る　美容師:が　髪:を
切る　ボタン:で　スイッチ:を	切る　業者:が　電源:を
切る　ガラス:で　従業員:が　頭部:を	**切る　告示:まで　<NUM>カ月:を**
切る　髪:を	切る　ガラス:で　手:を
切る　スイッチ:を	切る　教諭:が　髪:を　ハサミ:で

Output

切る　{クリスマス, 告示}:まで　{<NUM>カ月}:を

Grouping related subcategorization frames

Input

切る　クリスマス:まで　<NUM>カ月:を	切る　美容師:が　髪:を
切る　ボタン:で　スイッチ:を	切る　業者:が　電源:を
切る　ガラス:で　従業員:が　頭部:を	切る　告示:まで　<NUM>カ月:を
切る　髪:を	切る　ガラス:で　手:を
切る　スイッチ:を	切る　教諭:が　髪:を　ハサミ:で

Output

切る　{クリスマス, 告示}:まで　{<NUM>カ月}:を

切る　{ボタン}:で　{スイッチ}:を

Grouping related subcategorization frames

Input

切る　クリスマス:まで　<NUM>カ月:を	切る　美容師:が　髪:を
切る　ボタン:で　スイッチ:を	**切る　業者:が　電源:を**
切る　ガラス:で　従業員:が　頭部:を	切る　告示:まで　<NUM>カ月:を
切る　髪:を	切る　ガラス:で　手:を
切る　スイッチ:を	切る　教諭:が　髪:を　ハサミ:で

Output

切る　{クリスマス, 告示}:まで　{<NUM>カ月}:を

切る　{ボタン}:で　{スイッチ}:を

切る　{業者}:が　{電源}:を

Grouping related subcategorization frames

Input

切る　クリスマス:まで　<NUM>カ月:を	切る　美容師:が　髪:を
切る　ボタン:で　スイッチ:を	切る　業者:が　電源:を
切る　ガラス:で　従業員:が　頭部:を	切る　告示:まで　<NUM>カ月:を
切る　髪:を	**切る　ガラス:で　手:を**
切る　スイッチ:を	切る　教諭:が　髪:を　ハサミ:で

Output

切る　{クリスマス, 告示}:まで　{<NUM>カ月}:を

切る　{ボタン}:で　{スイッチ}:を

切る　{業者}:が　{電源}:を

切る　{ガラス}:で　{従業員}:が　{頭部, 手}:を

Grouping related subcategorization frames

Input

切る　クリスマス:まで　<NUM>カ月:を
切る　ボタン:で　スイッチ:を
切る　ガラス:で　従業員:が　頭部:を
切る　髪:を
切る　スイッチ:を

切る　美容師:が　髪:を
切る　業者:が　電源:を
切る　告示:まで　<NUM>カ月:を
切る　ガラス:で　手:を
切る　教諭:が　髪:を　ハサミ:で

Output

切る　{クリスマス, 告示}:まで　{<NUM>カ月}:を
切る　{ボタン}:で　{スイッチ}:を
切る　{業者}:が　{電源}:を
切る　{ガラス}:で　{従業員}:が　{頭部, 手}:を
切る　{美容師, 教諭}:が　{髪}:を　{ハサミ}:で

Grouping related subcategorization frames

Input

切る　クリスマス:まで　<NUM>カ月:を	切る　美容師:が　髪:を
切る　ボタン:で　スイッチ:を	切る　業者:が　電源:を
切る　ガラス:で　従業員:が　頭部:を	切る　告示:まで　<NUM>カ月:を
切る　髪:を	切る　ガラス:で　手:を
切る　スイッチ:を	切る　教諭:が　髪:を　ハサミ:で

Output

切る　{クリスマス, 告示}:まで　{<NUM>カ月}:を

切る　{ボタン}:で　{スイッチ}:を

切る　{業者}:が　{電源}:を

切る　{ガラス}:で　{従業員}:が　{頭部, 手}:を

切る　{美容師, 教諭}:が　{髪}:を　{ハサミ}:で

Grouping related subcategorization frames

Input

切る　クリスマス:まで　<NUM>カ月:を	切る　美容師:が　髪:を
切る　ボタン:で　スイッチ:を	切る　業者:が　電源:を
切る　ガラス:で　従業員:が　頭部:を	切る　告示:まで　<NUM>カ月:を
切る　髪:を	切る　ガラス:で　手:を
切る　スイッチ:を	切る　教諭:が　髪:を　ハサミ:で

Output

切る　{クリスマス, 告示}:まで　{<NUM>カ月}:を

- 切る　{ボタン}:で　{スイッチ}:を
- 切る　{業者}:が　{電源}:を

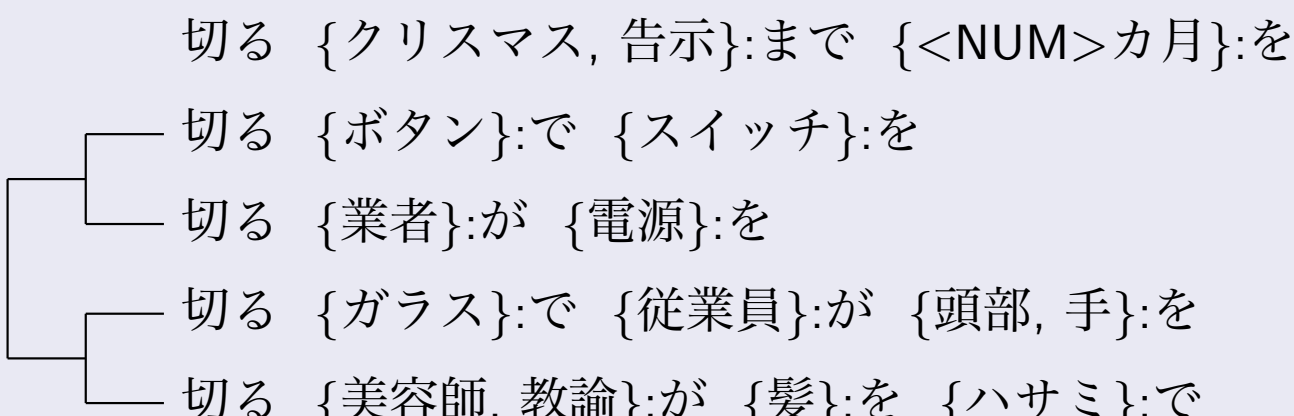
- 切る　{ガラス}:で　{従業員}:が　{頭部, 手}:を
- 切る　{美容師, 教諭}:が　{髪}:を　{ハサミ}:で

Grouping related subcategorization frames

Input

切る　クリスマス:まで　<NUM>カ月:を	切る　美容師:が　髪:を
切る　ボタン:で　スイッチ:を	切る　業者:が　電源:を
切る　ガラス:で　従業員:が　頭部:を	切る　告示:まで　<NUM>カ月:を
切る　髪:を	切る　ガラス:で　手:を
切る　スイッチ:を	切る　教諭:が　髪:を　ハサミ:で

Output



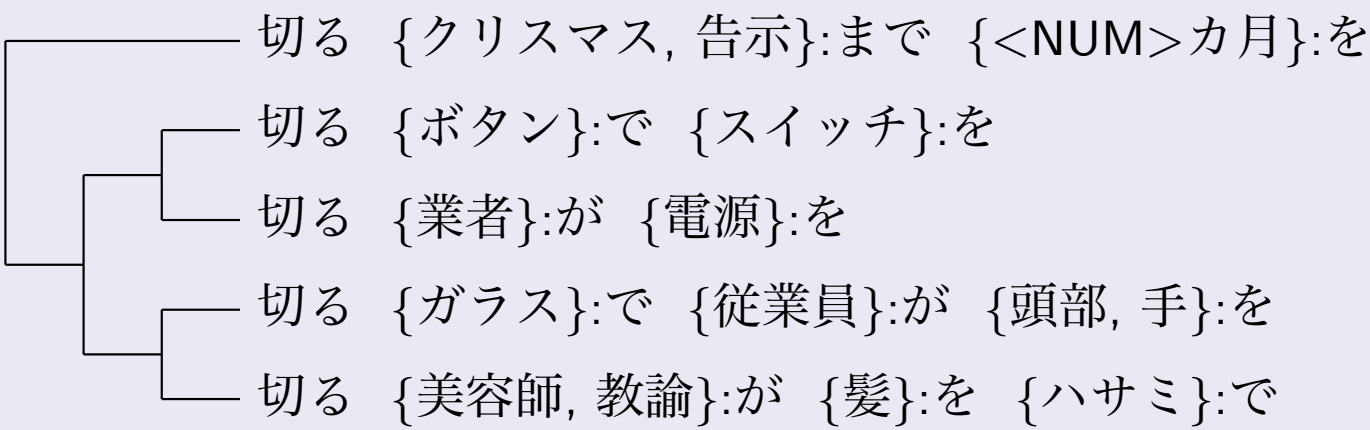
Grouping related subcategorization frames

Input

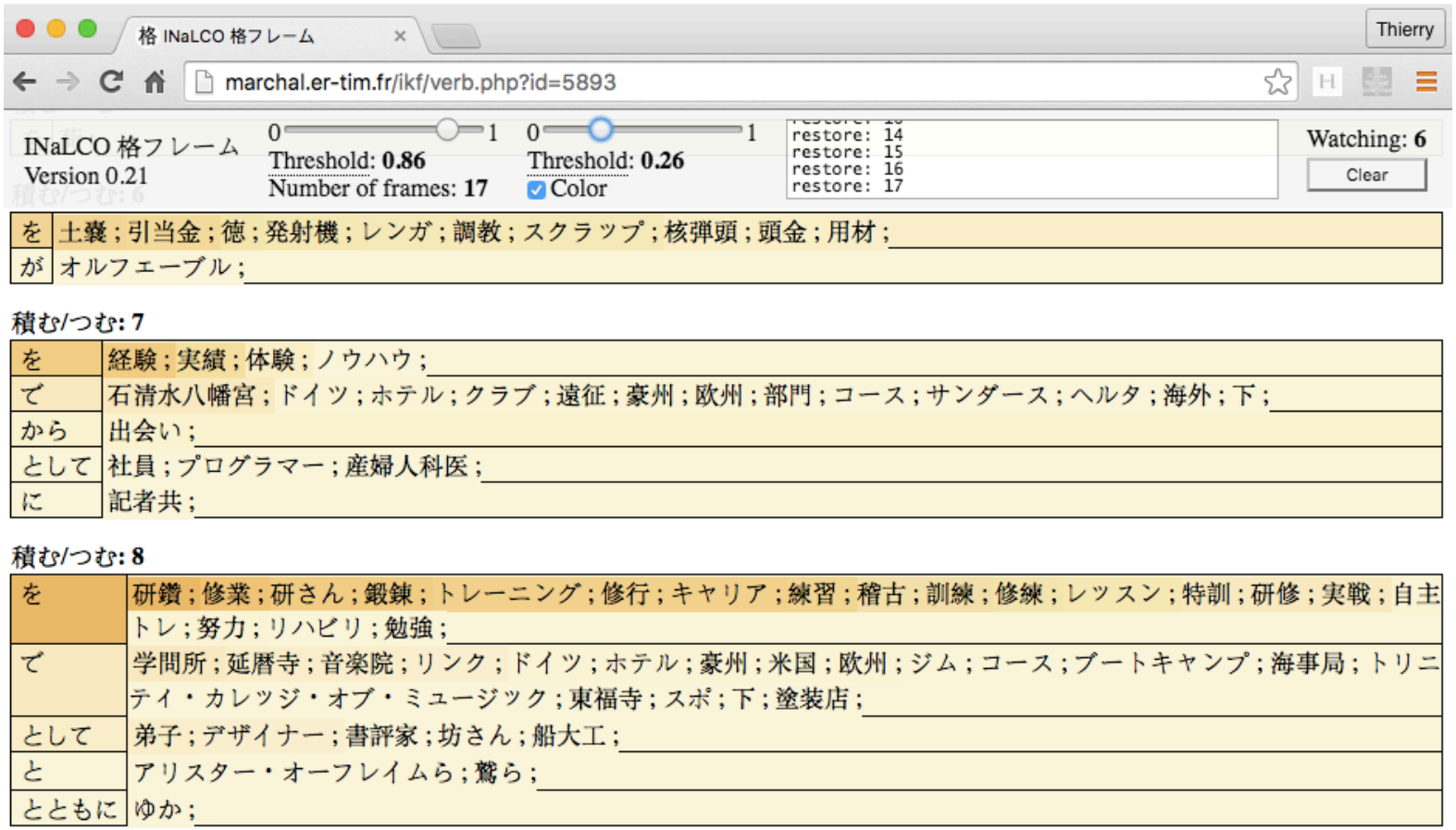
切る　クリスマス:まで　<NUM>カ月:を
切る　ボタン:で　スイッチ:を
切る　ガラス:で　従業員:が　頭部:を
切る　髪:を
切る　スイッチ:を

切る　美容師:が　髪:を
切る　業者:が　電源:を
切る　告示:まで　<NUM>カ月:を
切る　ガラス:で　手:を
切る　教諭:が　髪:を　ハサミ:で

Output



Graphical Interface



See `http://marchal.er-tim.fr/ikf/`

Assessment of the Resource by Lexicographers

What is the interface used for ?

- Observe linguistic phenomena "emerging" from the data (diathesis alternations, etc)
- Verify prior assumptions, especially concerning the status of complements (i.e. arguments versus adjuncts)

Main result $\Rightarrow$ Fine grained observation of specific phenomena

- Transitive versus intransitive use of the verbs
- Broad semantic classes of the complements (e.g. literal versus figurative)

Conclusion and Perspectives

- Continue experiments with lexicographer
- Perform a practical evaluation through the integration of this resource into specific natural language applications (e.g. Machine translation)
- Evaluate the approach on other languages (e.g. Finnish)

Thank you !