< WTEP 2007, Jun. 29th, 2007 >

A Compositional Approach toward Dynamic Phrasal Thesaurus

Atsushi FUJITA, Shuhei KATO,
Naoki KATO, Satoshi SATO

Nagoya Univ., Japan

Computing Semantic Equivalence (SE)

- Fundamental in NLP
  - Recognition: IR, IE, QA
  - Generation: MT, TTS, Summarization
- Previous attempts used ...
  - Thesauri [So many work]
  - Tree kernels [Collins+, 01] [Takahashi, 05]
  - Statistical translation models [Barzilay+, 03] [Brockett+, 05]
  - Distributional similarity [Harris, 64] [Lin+, 01] [Weeds+, 05]
  - Syntactic patterns [Mel’cuk+, 87] [Dras, 99] [Jacquemin, 99]

Computing Semantic Equivalence (SE)

- Fundamental in NLP
  - Recognition: IR, IE, QA
  - Generation: MT, TTS, Summarization
- Previous attempts used ...
  - Thesauri
  - Tree kernels

  Words are not necessarily the unit of meaning (polysemous words, meaning of construction)

  Cannot generate paraphrases

  - Statistical translation models
  - Distributional similarity

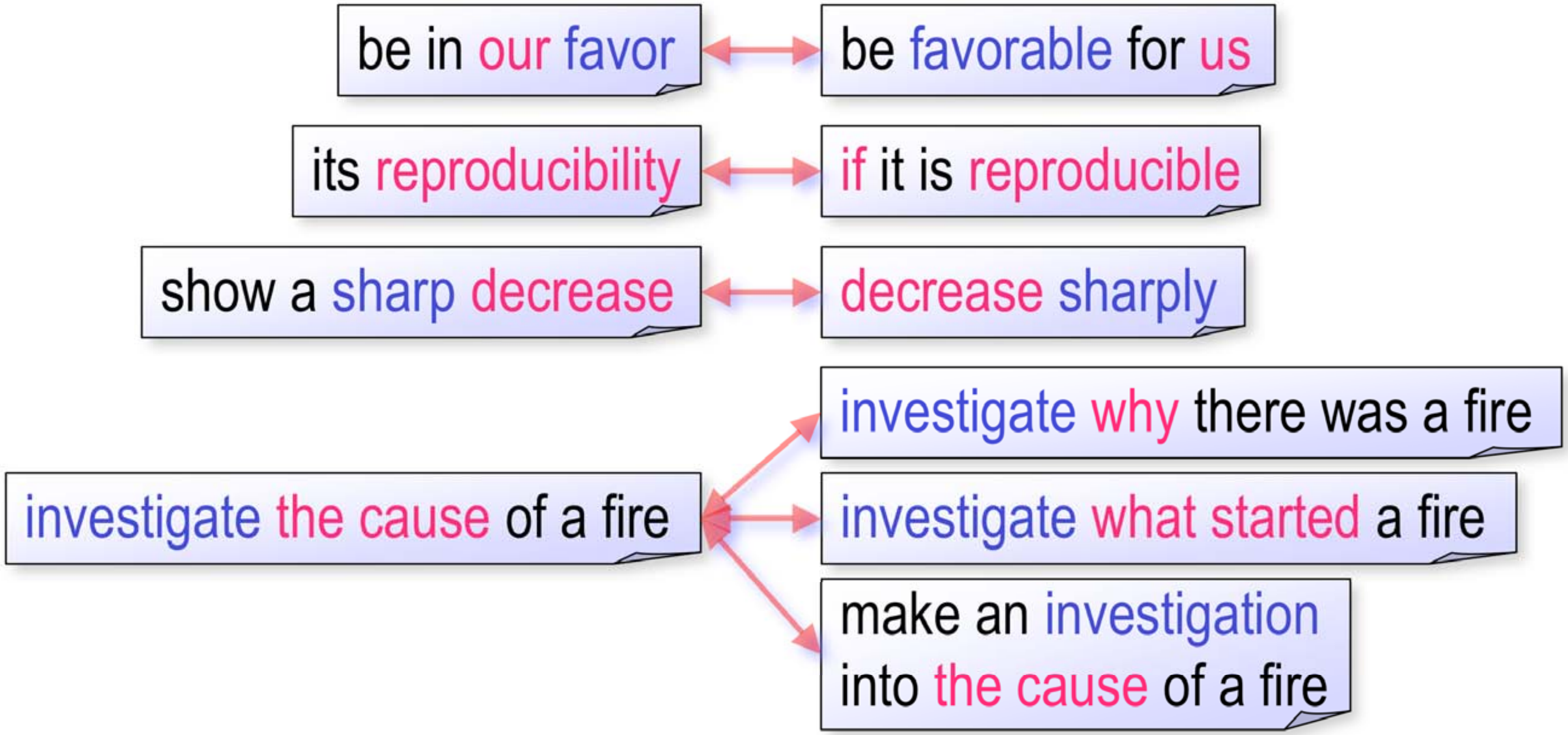
  Corpus is not almighty (data sparseness, cost)

  - Syntactic patterns

  No thorough list

Our Proposal

- Phrasal Thesaurus
  - A mechanism for directly computing SE between phrases



be in our favor ↔ be favorable for us

its reproducibility ↔ if it is reproducible

show a sharp decrease ↔ decrease sharply

investigate the cause of a fire ↔ investigate why there was a fire

investigate the cause of a fire ↔ investigate what started a fire

investigate the cause of a fire ↔ make an investigation into the cause of a fire

Aim

- Implement tools and resources
  - Application-independent module
  - Human aids: writing / reading texts
- Confirm phrase is appropriate unit for computing SE
  - Ambiguity of words >> Ambiguity of phrases
    (more suitable to handle)

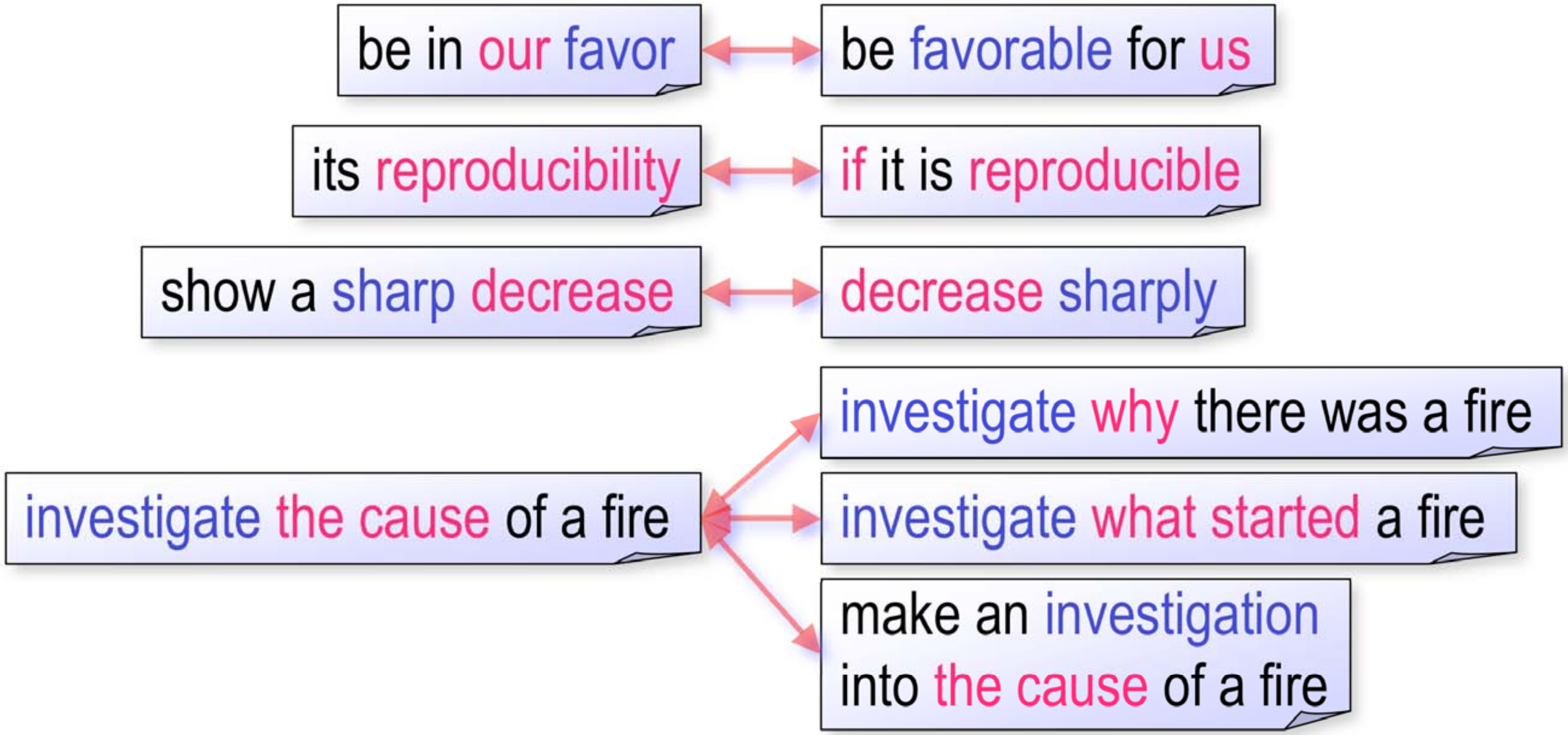
This is a preliminary progress report
(w/o concrete evaluation)

Outline

1. Motivation & Aim
2. Range of phenomena
3. System & implementation
4. Discussion
5. Conclusion

Towards Phrasal Thesaurus

- What sorts of phrases?
- How to handle a variety of expressions?



be in our favor ↔ be favorable for us

its reproducibility ↔ if it is reproducible

show a sharp decrease ↔ decrease sharply

investigate the cause of a fire ↔ investigate why there was a fire

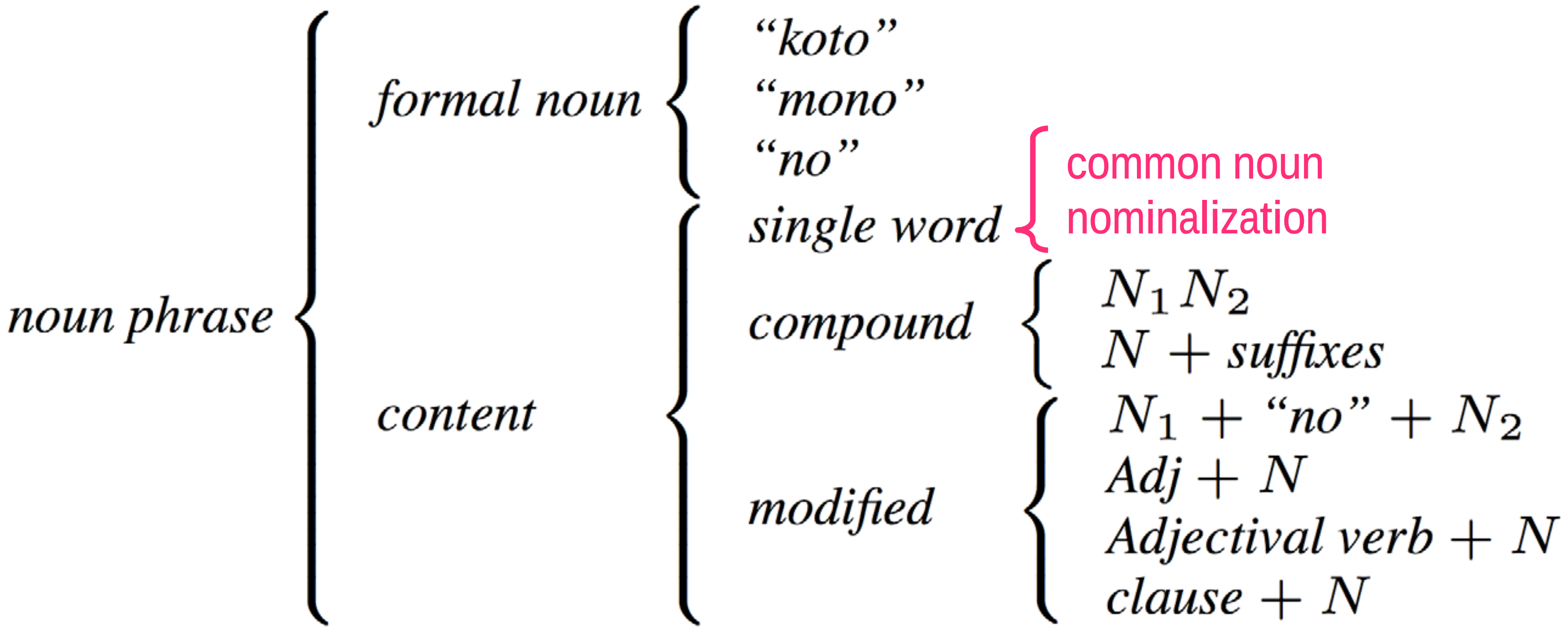
investigate the cause of a fire ↔ investigate what started a fire

investigate the cause of a fire ↔ make an investigation into the cause of a fire

Range of phrases

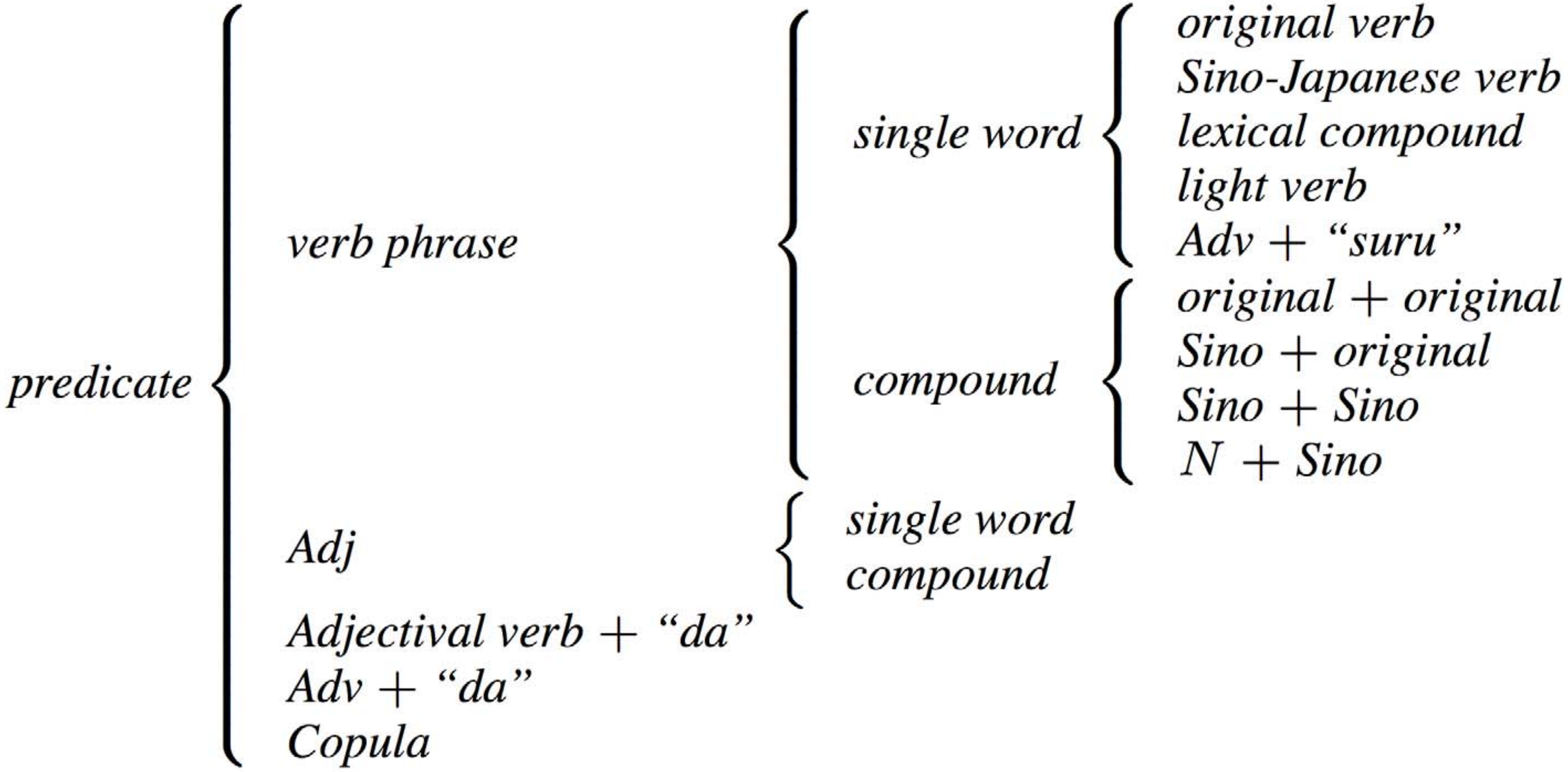
- Predicate phrase (cf. various exps. in RTE)
  - Reliably captured using recent technologies
  - Approx. corresponds to single event
    [Chklovski and Pantel, 2004] [Torisawa, 2006]
- Our target language: Japanese
  - noun phrase + case marker + predicate
    - Various noun phrases
    - Various predicates
    - Case markers indicate grammatical roles of noun phrases

Classification of noun phrases in Japanese



- *noun phrase*
  - *formal noun*
    - *“koto”*
    - *“mono”*
    - *“no”*
  - *content*
    - *single word*
      - common noun
      - nominalization
    - *compound*
      - $N_1 N_2$
      - N + *suffixes*
    - *modified*
      - N_1 + *“no”* + N_2
      - *Adj* + N
      - *Adjectival verb* + N
      - *clause* + N

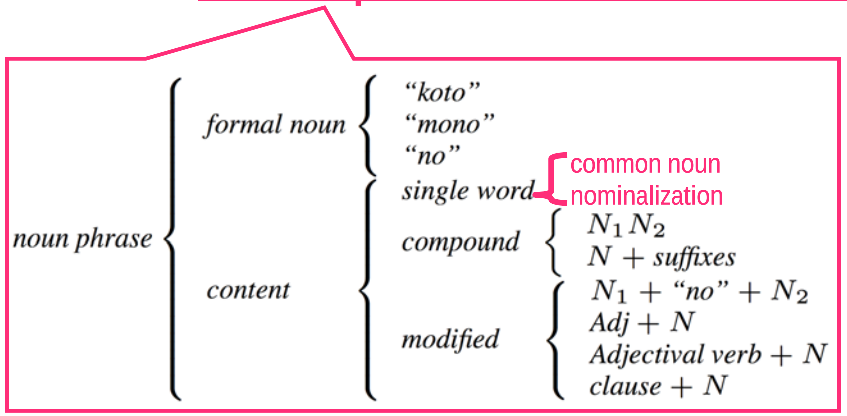
Classification of predicates in Japanese



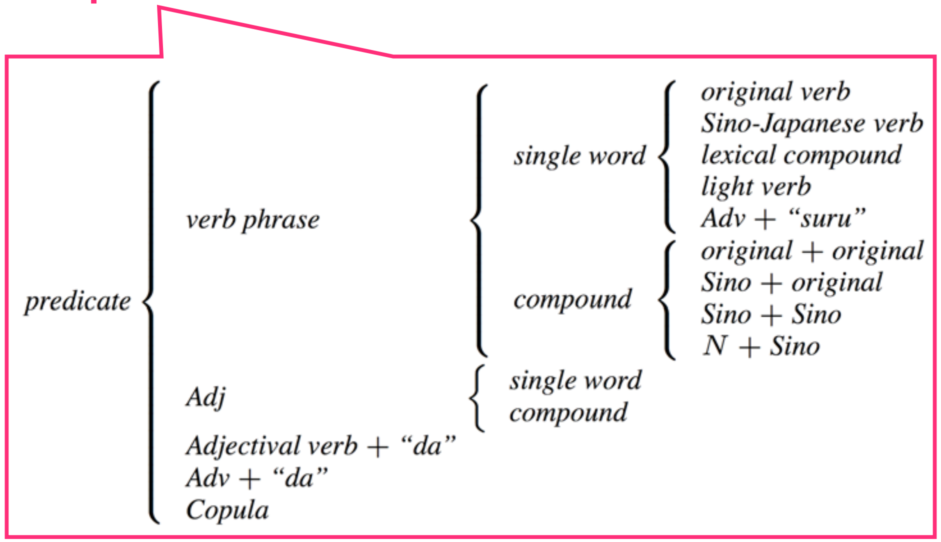
- predicate
  - verb phrase
    - single word
      - original verb
      - Sino-Japanese verb
      - lexical compound
      - light verb
      - Adv + “suru”
    - compound
      - original + original
      - Sino + original
      - Sino + Sino
      - N + Sino
  - Adj
    - single word
    - compound
  - Adjectival verb + “da”
  - Adv + “da”
  - Copula

Range of phrases

- Our target language: Japanese
  - noun phrase + case marker + predicate



- *noun phrase*
  - *formal noun*
    - *“koto”*
    - *“mono”*
    - *“no”*
  - *content*
    - *single word*
      - common noun
      - nominalization
    - *compound*
      - $N_1 N_2$
      - N + *suffixes*
    - *modified*
      - N_1 + *“no”* + N_2
      - *Adj* + N
      - *Adjectival verb* + N
      - *clause* + N



- *predicate*
  - *verb phrase*
    - *single word*
      - *original verb*
      - *Sino-Japanese verb*
      - *lexical compound*
      - *light verb*
      - *Adv* + *“suru”*
    - *compound*
      - *original* + *original*
      - *Sino* + *original*
      - *Sino* + *Sino*
      - N + *Sino*
  - *Adj*
    - *single word*
    - *compound*
  - *Adjectival verb* + *“da”*
  - *Adv* + *“da”*
  - *Copula*

- Variation of phrases >> Variation of words
  - Various combinations of open-class words

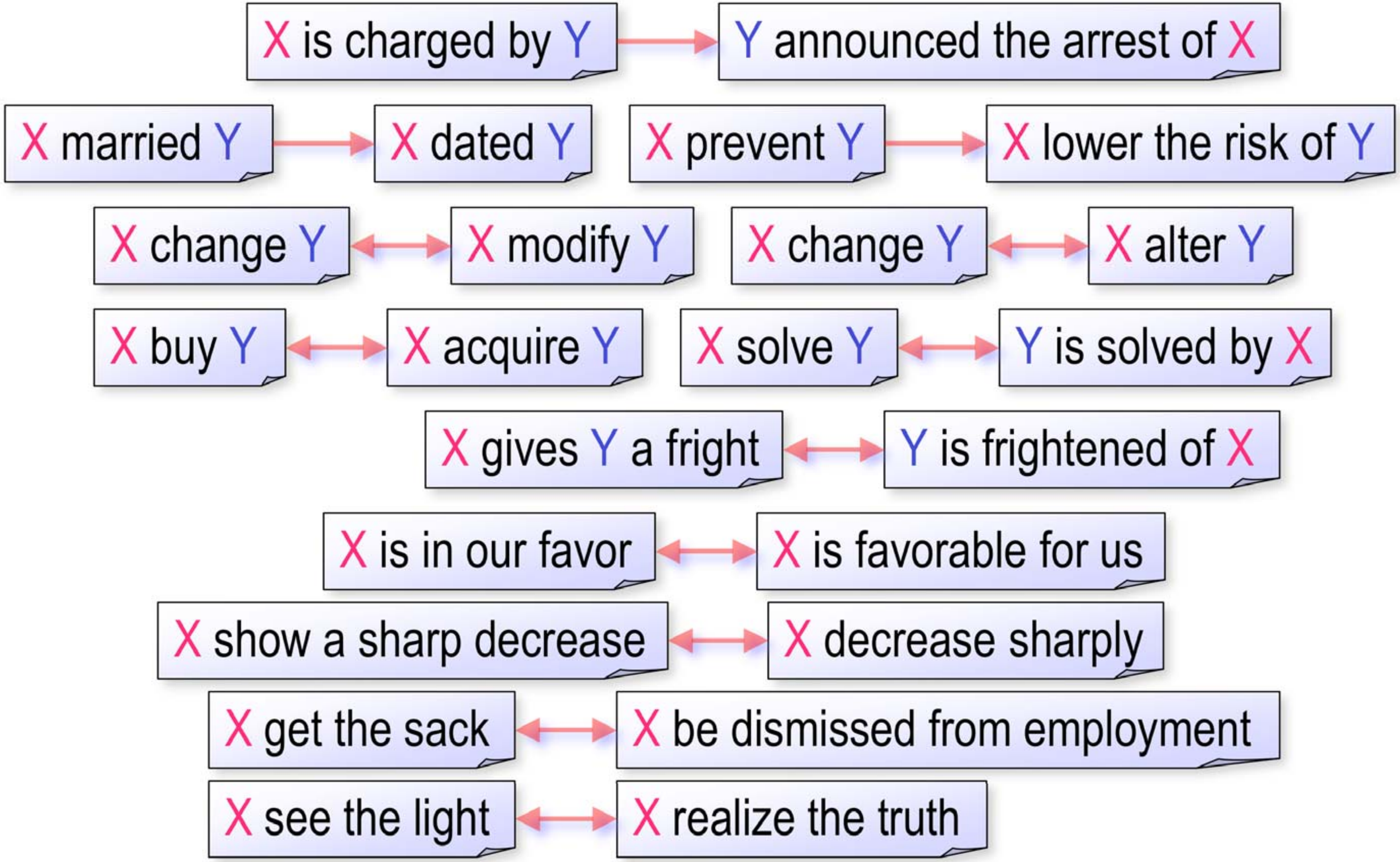
Range of phenomena

- Variation of paraphrases of phrases

  >> Variation of paraphrases of words

  - Difficult (hard?) to statically enumerate
  - No previous work explicitly collected:
    - “All verbs that can be passivized”
    - “All noun-verb pairs that compose light-verb constructions”
  - How to handle them?

Paraphrases of predicate phrases



X is charged by Y → Y announced the arrest of X

X married Y → X dated Y

X prevent Y → X lower the risk of Y

X change Y ↔ X modify Y

X change Y ↔ X alter Y

X buy Y ↔ X acquire Y

X solve Y ↔ Y is solved by X

X gives Y a fright ↔ Y is frightened of X

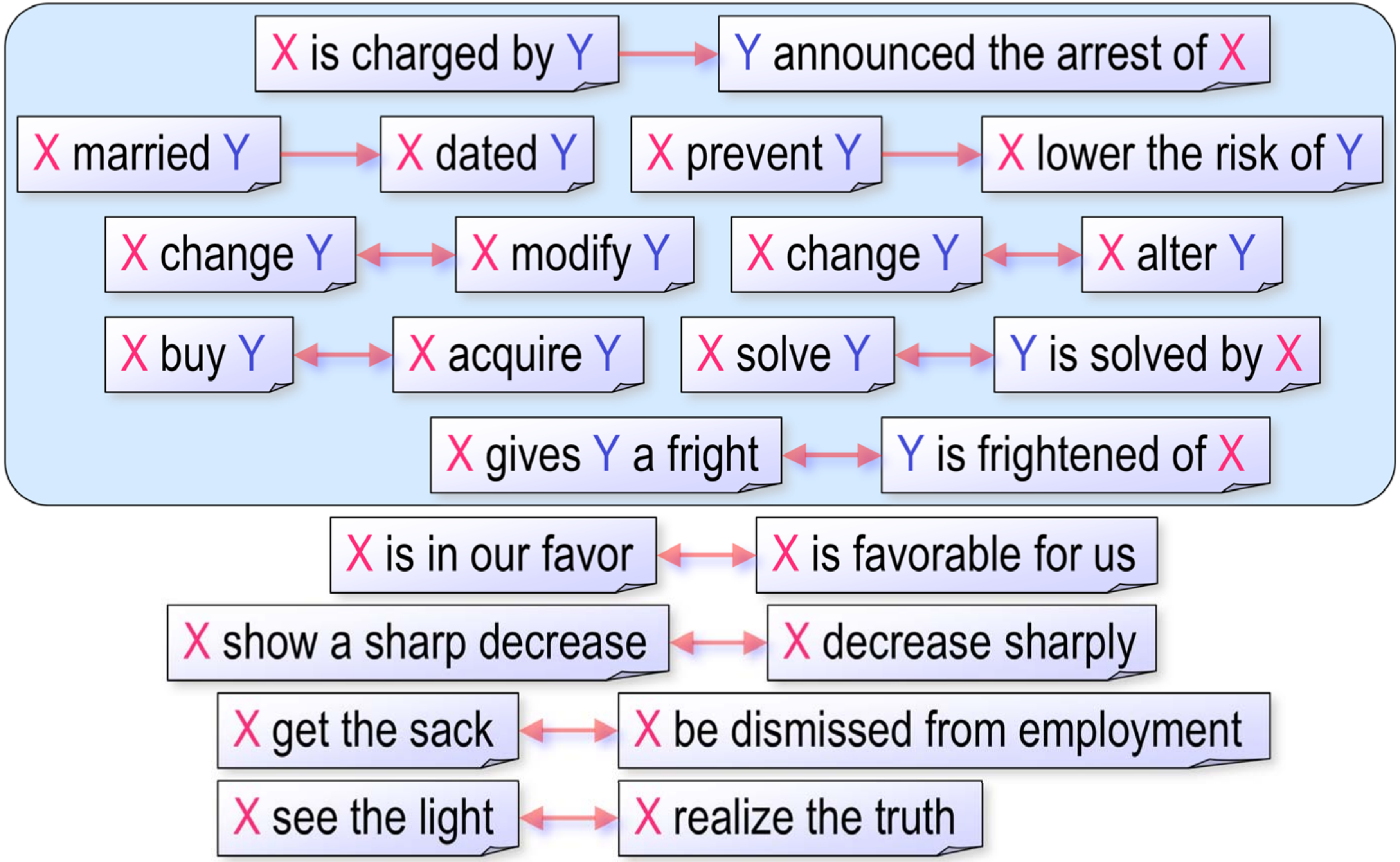
X is in our favor ↔ X is favorable for us

X show a sharp decrease ↔ X decrease sharply

X get the sack ↔ X be dismissed from employment

X see the light ↔ X realize the truth

Paraphrases of predicate phrases



X is charged by Y → Y announced the arrest of X

X married Y → X dated Y

X prevent Y → X lower the risk of Y

X change Y ↔ X modify Y

X change Y ↔ X alter Y

X buy Y ↔ X acquire Y

X solve Y ↔ Y is solved by X

X gives Y a fright ↔ Y is frightened of X

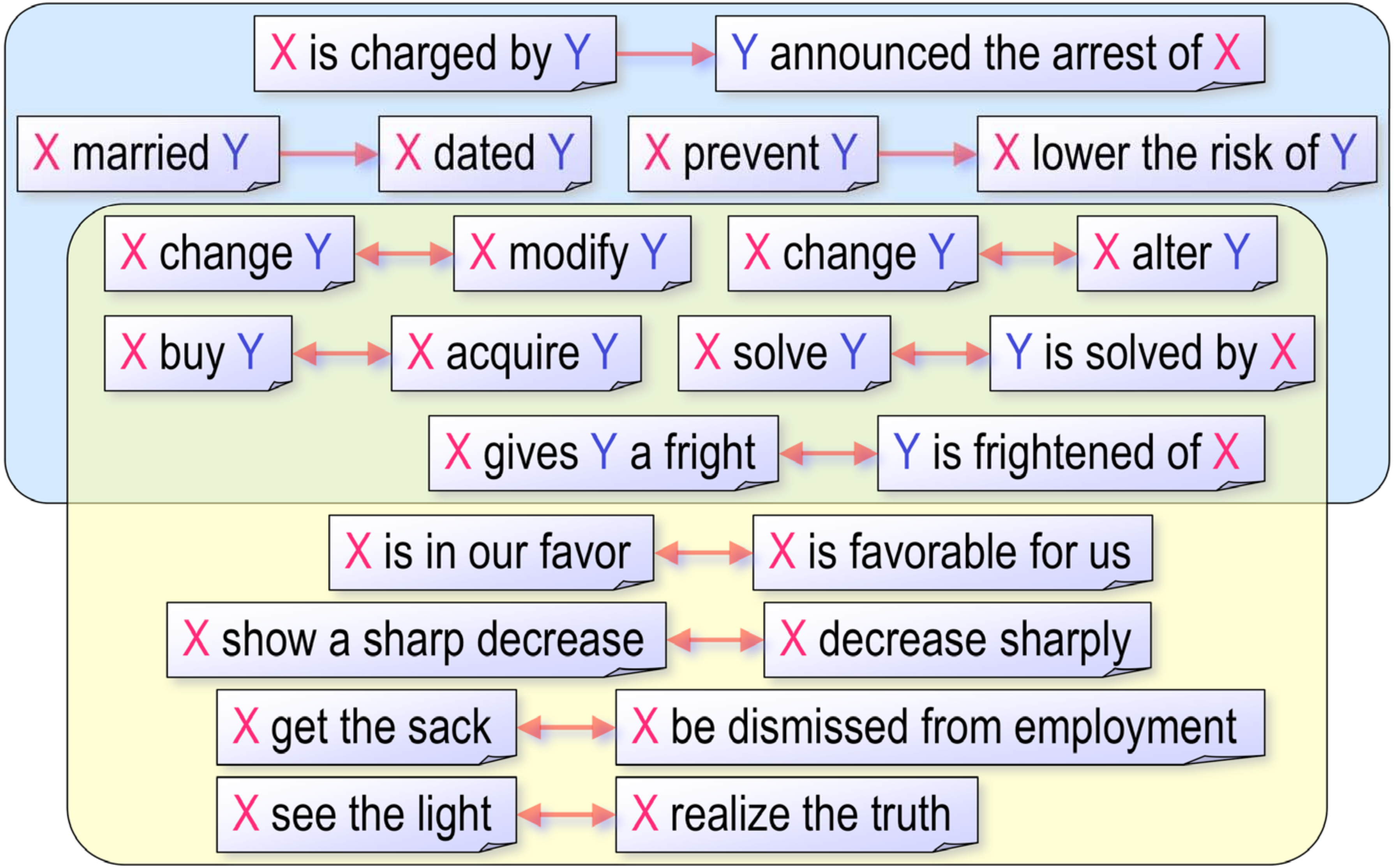
X is in our favor ↔ X is favorable for us

X show a sharp decrease ↔ X decrease sharply

X get the sack ↔ X be dismissed from employment

X see the light ↔ X realize the truth

Paraphrases of predicate phrases



X is charged by Y → Y announced the arrest of X

X married Y → X dated Y

X prevent Y → X lower the risk of Y

X change Y ↔ X modify Y

X change Y ↔ X alter Y

X buy Y ↔ X acquire Y

X solve Y ↔ Y is solved by X

X gives Y a fright ↔ Y is frightened of X

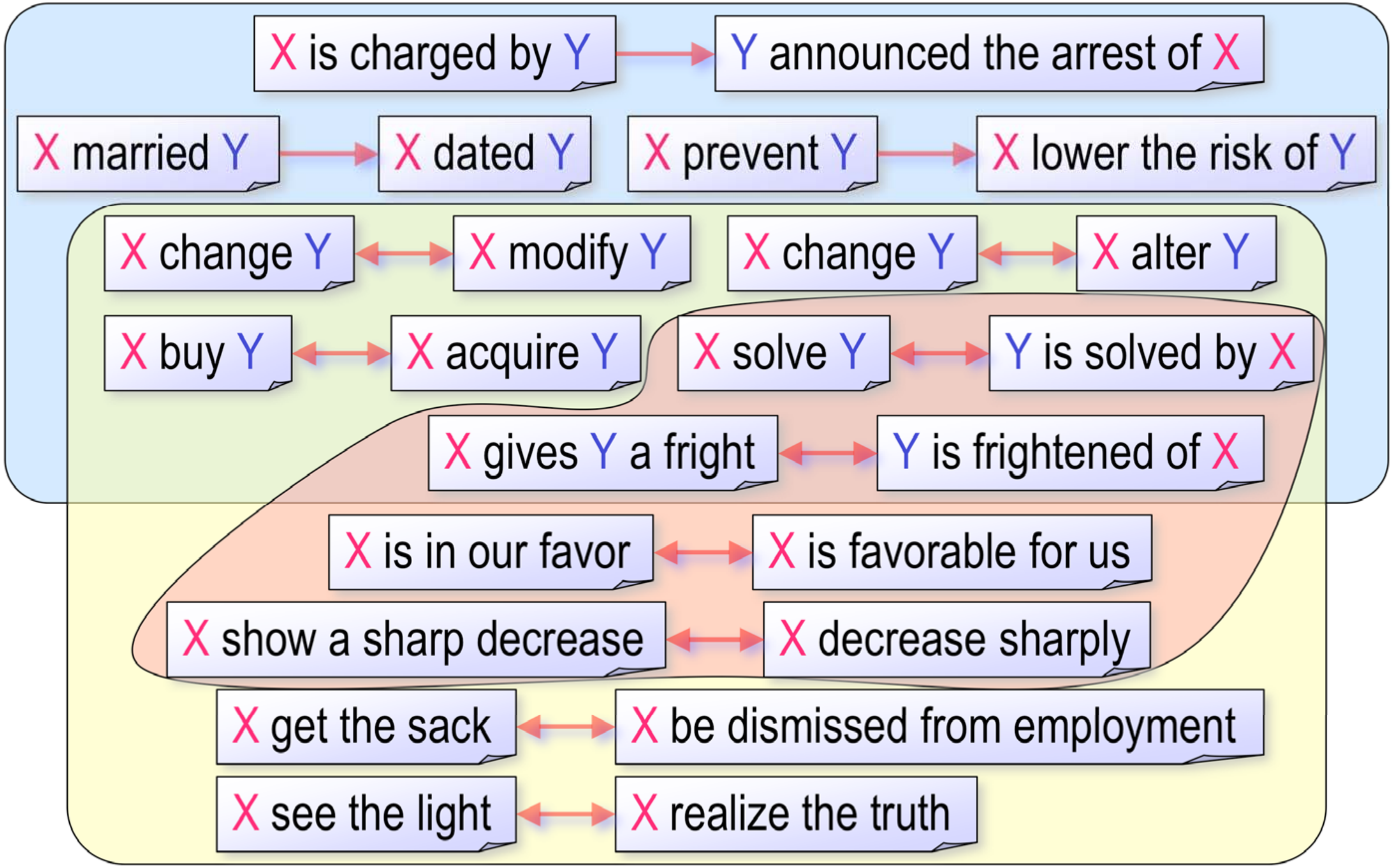
X is in our favor ↔ X is favorable for us

X show a sharp decrease ↔ X decrease sharply

X get the sack ↔ X be dismissed from employment

X see the light ↔ X realize the truth

Paraphrases of predicate phrases



X is charged by Y → Y announced the arrest of X

X married Y → X dated Y

X prevent Y → X lower the risk of Y

X change Y ↔ X modify Y

X change Y ↔ X alter Y

X buy Y ↔ X acquire Y

X solve Y ↔ Y is solved by X

X gives Y a fright ↔ Y is frightened of X

X is in our favor ↔ X is favorable for us

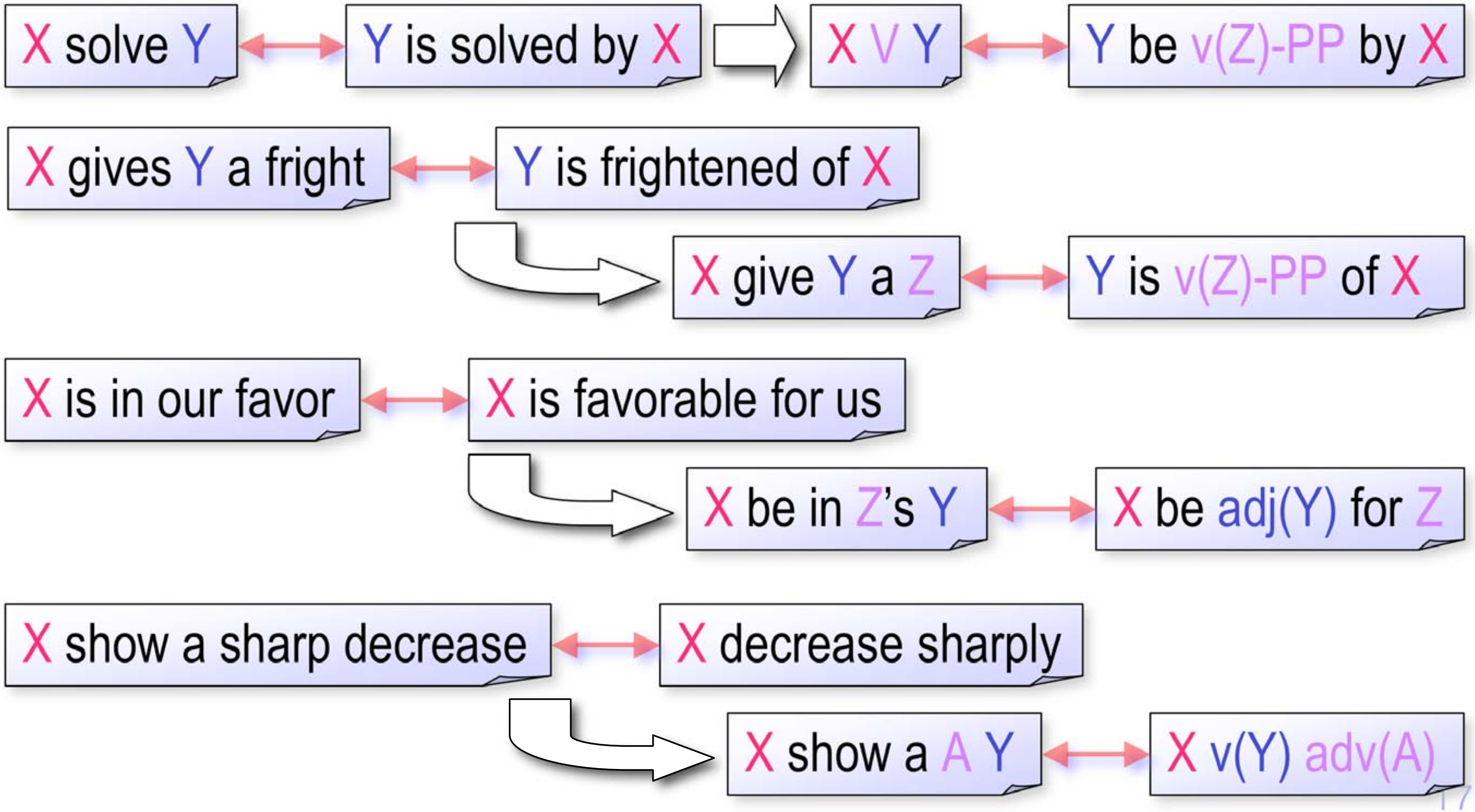
X show a sharp decrease ↔ X decrease sharply

X get the sack ↔ X be dismissed from employment

X see the light ↔ X realize the truth

Compositional paraphrases (syntactic variants)

- Syntactic transformation + Lexical derivation
  ⇒ Dynamic generation (Dynamic Phrasal Thesaurus)



X solve Y ↔ Y is solved by X ⇒ X V Y ↔ Y be v(Z)-PP by X

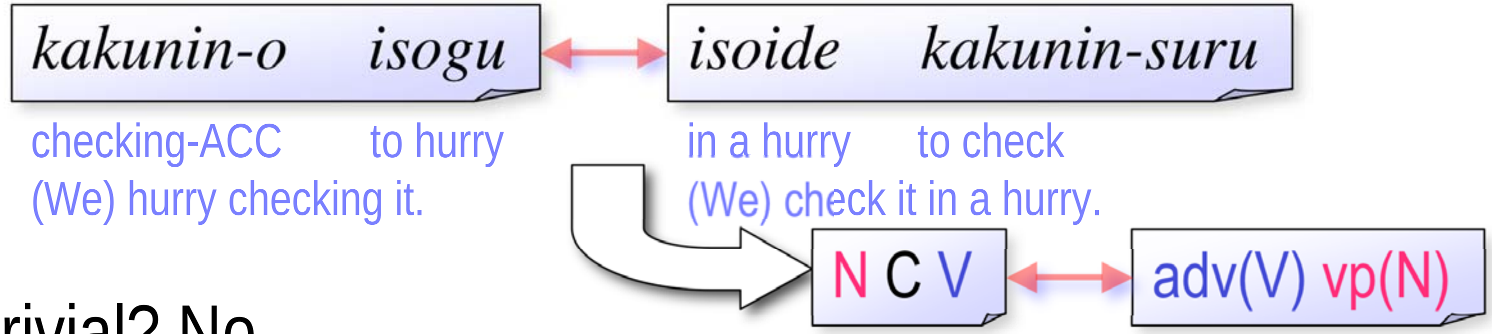
X gives Y a fright ↔ Y is frightened of X ⇒ X give Y a Z ↔ Y is v(Z)-PP of X

X is in our favor ↔ X is favorable for us ⇒ X be in Z's Y ↔ X be adj(Y) for Z

X show a sharp decrease ↔ X decrease sharply ⇒ X show a A Y ↔ X v(Y) adv(A)

Compositional paraphrases (syntactic variants)

- Syntactic transformation + Lexical derivation
  ⇒ Dynamic generation (Dynamic Phrasal Thesaurus)
  - Our target language: Japanese



kakunin-o *isogu*

checking-ACC to hurry

(We) hurry checking it.

isoide *kakunin-suru*

in a hurry to check

(We) check it in a hurry.

N C V

adv(V) vp(N)

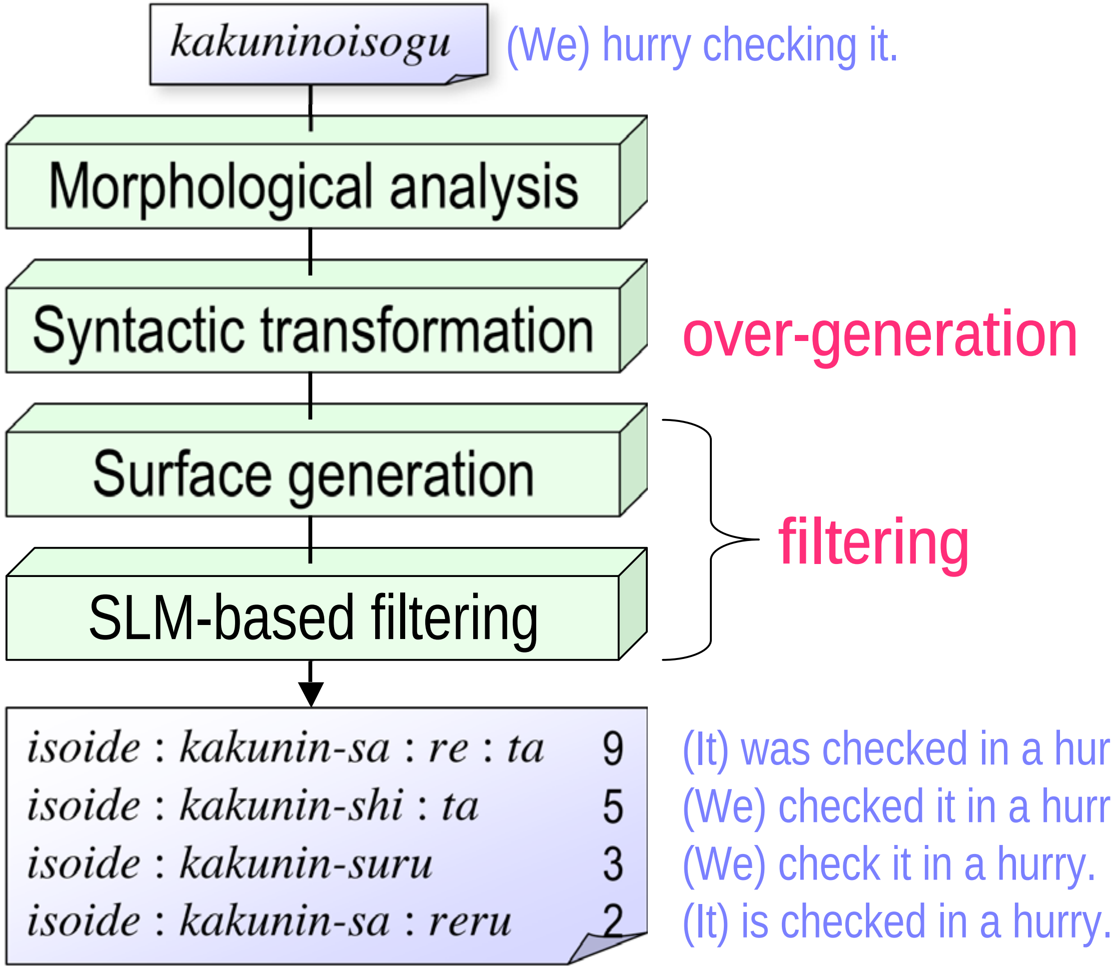
  - Trivial? No.
    - Not exhaustively explored
    - Beneficial [Dolan+, 04] [Romano+, 06]

Outline

1. Motivation & Aim
2. Range of phenomena
3. System & implementation
4. Discussion
5. Conclusion

System overview

- <u>Input</u>: Phrase (string)
- <u>Output</u>: List of paraphrases



kakuninoisogu (We) hurry checking it.

Morphological analysis

Syntactic transformation

Surface generation

SLM-based filtering

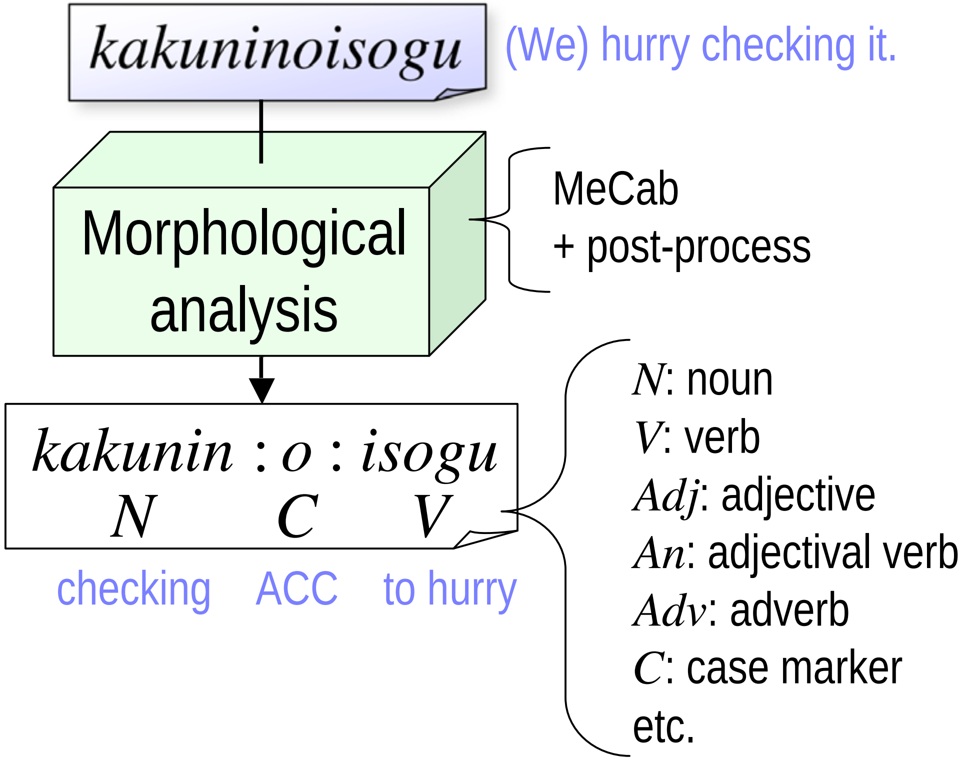
over-generation

filtering

isoide : kakunin-sa : re : ta	9	(It) was checked in a hurry.	(x)
isoide : kakunin-shi : ta	5	(We) checked it in a hurry.	(x)
isoide : kakunin-suru	3	(We) check it in a hurry.	(o)
isoide : kakunin-sa : reru	2	(It) is checked in a hurry.	(x)

1. Morphological analysis

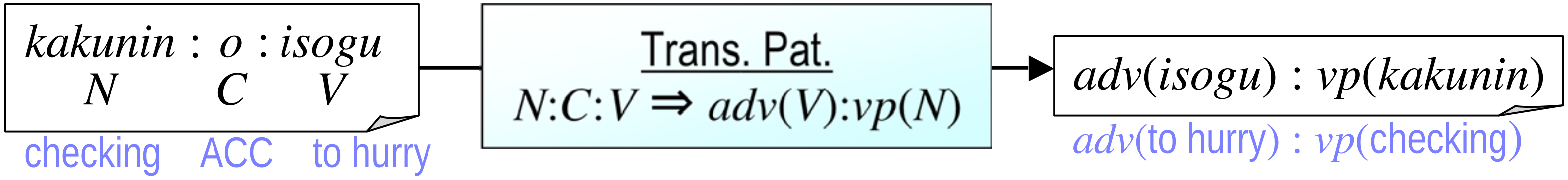
- Input: Phrase (string)
- Output: Array of morphemes w/ POS-tag
  - Using MeCab-0.91, a state-of-the-art morphological analyzer



N: noun
V: verb
Adj: adjective
An: adjectival verb
Adv: adverb
C: case marker
etc.

2. Syntactic transformation: knowledge used

- Transformation pattern
  - Generates skeletons of syntactic variants

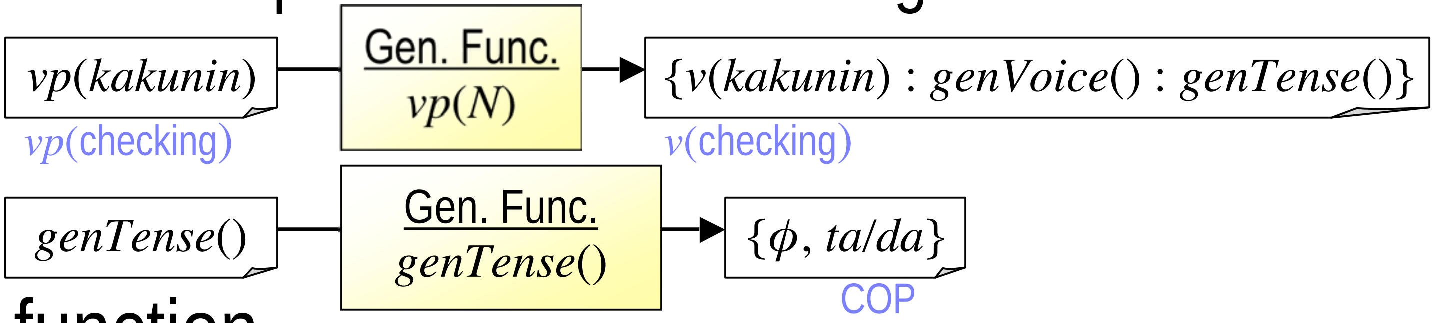


$kakunin : o : isogu$
$N \quad C \quad V$
checking ACC to hurry

Trans. Pat.
$N{:}C{:}V \Rightarrow adv(V){:}vp(N)$

$adv(isogu) : vp(kakunin)$
adv(to hurry) : vp(checking)

- Generation function
  - Enumerates expressions made of the given set of words



$vp(kakunin)$
vp(checking)

Gen. Func.
$vp(N)$

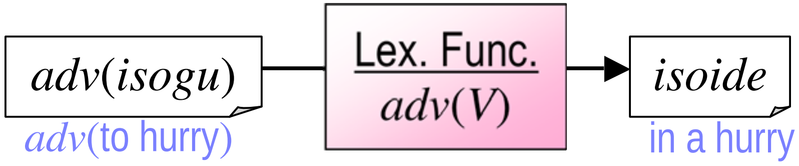
$\{v(kakunin) : genVoice() : genTense()\}$
v(checking)

$genTense()$

Gen. Func.
$genTense()$

$\{\phi, ta/da\}$
COP

- Lexical function
  - Generates different lexical items in certain relation

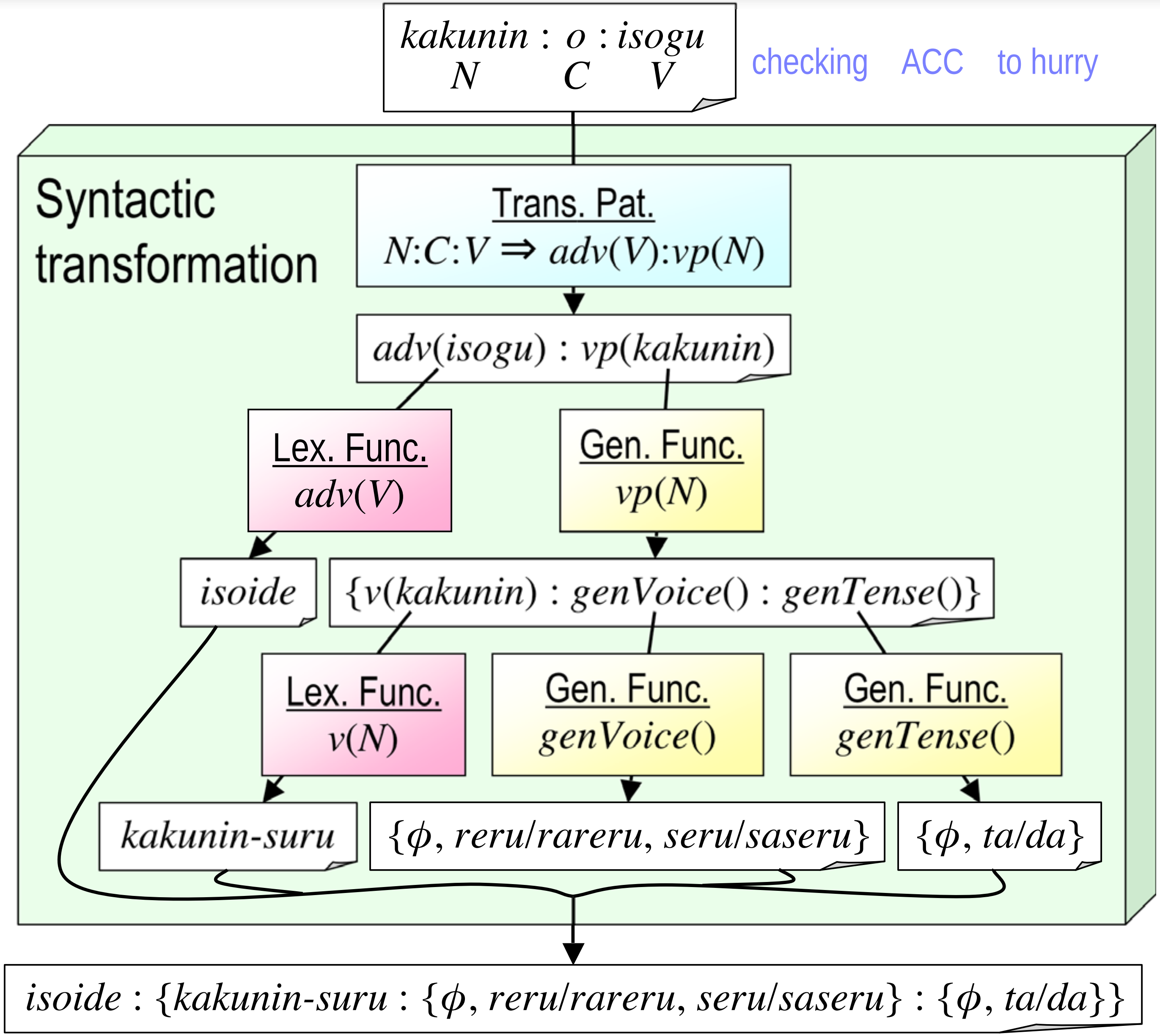


$adv(isogu)$
adv(to hurry)

Lex. Func.
$adv(V)$

$isoide$
in a hurry

2. Syntactic transformation: example



kakunin : *o* : *isogu*
N *C* *V*

checking ACC to hurry

Syntactic transformation

Trans. Pat.
$N{:}C{:}V \Rightarrow adv(V){:}vp(N)$

$adv(isogu) : vp(kakunin)$

Lex. Func.
$adv(V)$

Gen. Func.
$vp(N)$

isoide

$\{v(kakunin) : genVoice() : genTense()\}$

Lex. Func.
$v(N)$

Gen. Func.
$genVoice()$

Gen. Func.
$genTense()$

kakunin-suru

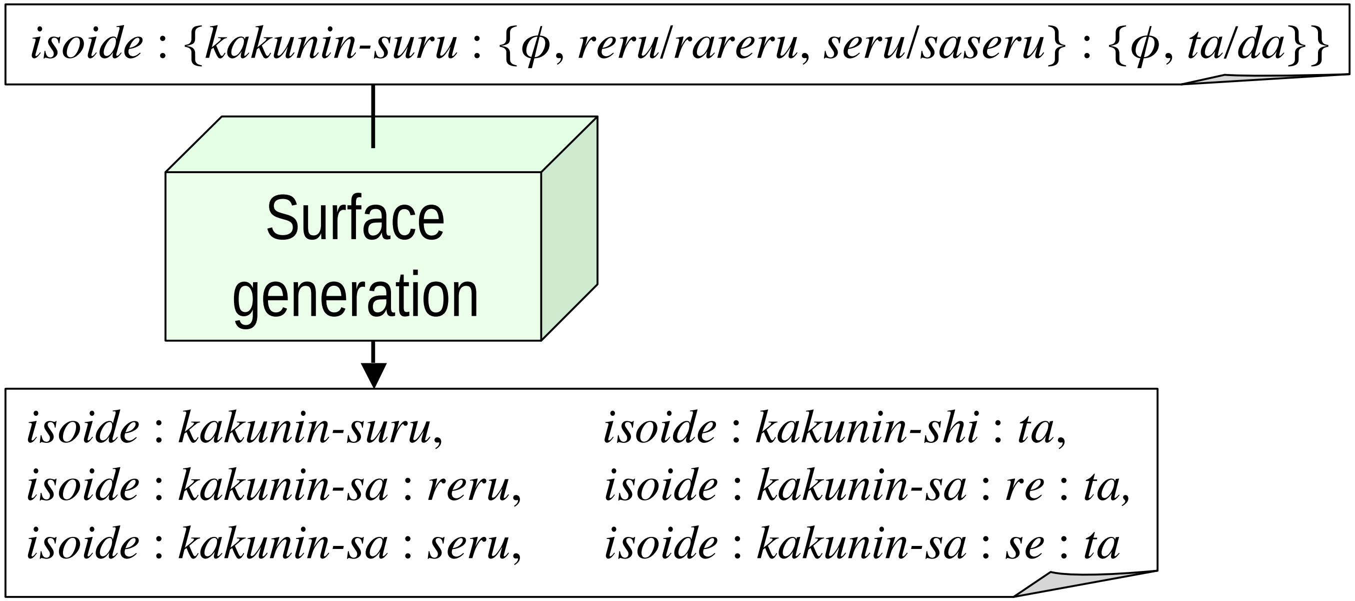
$\{\phi, reru/rareru, seru/saseru\}$

$\{\phi, ta/da\}$

$isoide : \{kakunin\text{-}suru : \{\phi, reru/rareru, seru/saseru\} : \{\phi, ta/da\}\}$

3. Surface generation

- Input: Bunch of candidate phrases
- Output: List of candidate phrases
  - 1. Unfolding
  - 2. Lexical choice (exclusively used auxiliaries)
  - 3. Conjugation



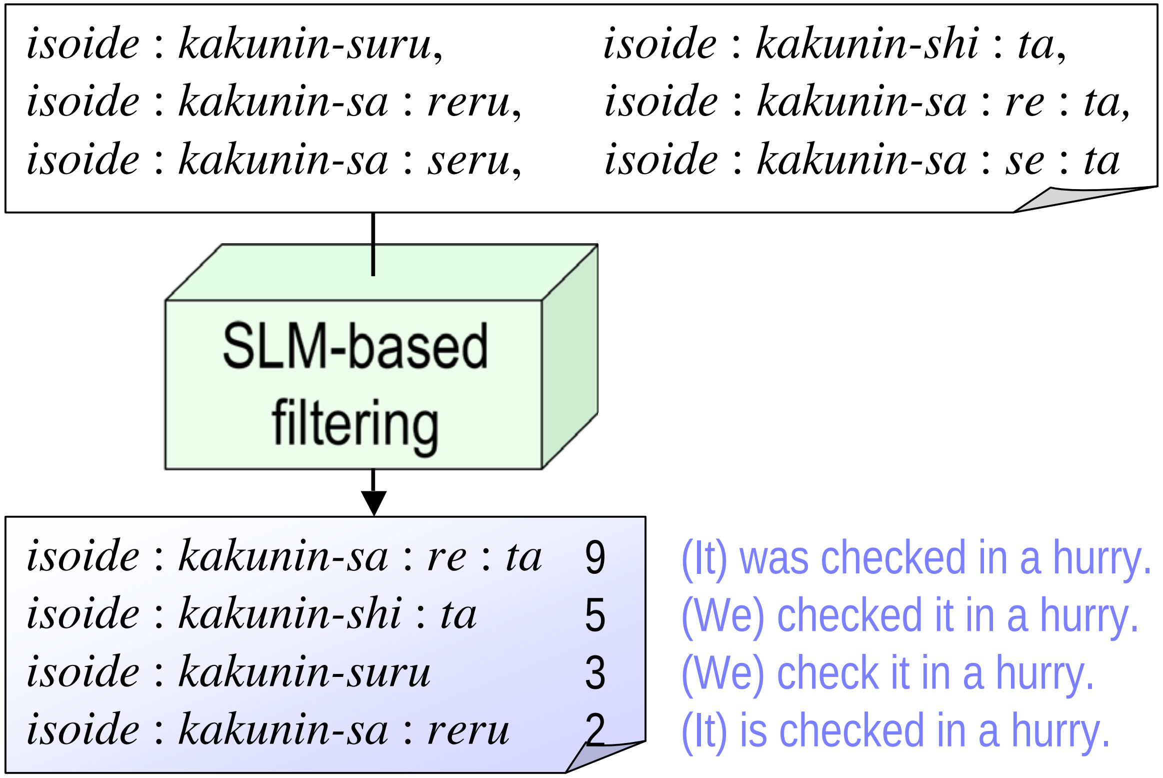
$\mathit{isoide} : \{\mathit{kakunin\text{-}suru} : \{\phi, \mathit{reru/rareru}, \mathit{seru/saseru}\} : \{\phi, \mathit{ta/da}\}\}$

Surface generation

$\mathit{isoide} : \mathit{kakunin\text{-}suru}$, $\mathit{isoide} : \mathit{kakunin\text{-}shi} : \mathit{ta}$,
$\mathit{isoide} : \mathit{kakunin\text{-}sa} : \mathit{reru}$, $\mathit{isoide} : \mathit{kakunin\text{-}sa} : \mathit{re} : \mathit{ta}$,
$\mathit{isoide} : \mathit{kakunin\text{-}sa} : \mathit{seru}$, $\mathit{isoide} : \mathit{kakunin\text{-}sa} : \mathit{se} : \mathit{ta}$

4. SLM-based filtering

- **Input**: List of candidate phrases
- **Output**: List of grammatical phrases
  - Grammaticality assessment
    - Initial model: if occur in Mainichi 1999-2005 (1.8GB)



isoide : *kakunin-suru*, *isoide* : *kakunin-shi* : *ta*,
isoide : *kakunin-sa* : *reru*, *isoide* : *kakunin-sa* : *re* : *ta*,
isoide : *kakunin-sa* : *seru*, *isoide* : *kakunin-sa* : *se* : *ta*

isoide : *kakunin-sa* : *re* : *ta*	9	(It) was checked in a hurry.	(x)
isoide : *kakunin-shi* : *ta*	5	(We) checked it in a hurry.	(x)
isoide : *kakunin-suru*	3	(We) check it in a hurry.	(o)
isoide : *kakunin-sa* : *reru*	2	(It) is checked in a hurry.	(x)

Knowledge development

- Paraphrase phenomena ⇒ Create patterns
  - Not necessarily from examples
  - Same manner as
    - MTT [Mel’cuk+, 1987]
    - STAG [Dras, 1999]
    - FASTR [Jacquemin, 1999]
    - KURA [Takahashi+, 2001]
- cf. FrameNet [Baker+, 1998]
  - Frame ⇒ Register various expressions

Comparison w/ previous work

- MTT [Mel’cuk+, 1987]
  - Paraphrasing rules at 7 levels
  - More than 60 Lexical functions

Trans. Pat.
$N{:}C{:}V \Rightarrow adv(V){:}vp(N)$

Lex. Func.
$adv(V)$

- FASTR [Jacquemin, 1999]
  - Structural transformations (Syntagma)
  - Semantic links (Paradigm)

Trans. Pat.
$N{:}C{:}V \Rightarrow adv(V){:}vp(N)$

Lex. Func.
$adv(V)$

- Ours
  - Transformation at SSynt level only (cf. MTT)
  - Predicate phrase, not technical term (cf. FASTR)
  - One-to-N generation by Gen.Func.

Gen. Func.
$vp(N)$

Current scale of knowledge

- Transformation pattern
  - Starting from N:C:V
    - $N_1{:}N_2{:}C{:}V$, $N{:}C{:}V_1{:}V_2$, ... : 37 patterns

Trans. Pat.
$N{:}C{:}V \Rightarrow adv(V){:}vp(N)$

- Generation function
  - As a by-product of generalizing transformation patterns
    - Content phrases (5): NPs, VPs
    - Functional expressions (4): Case markers, Auxiliaries

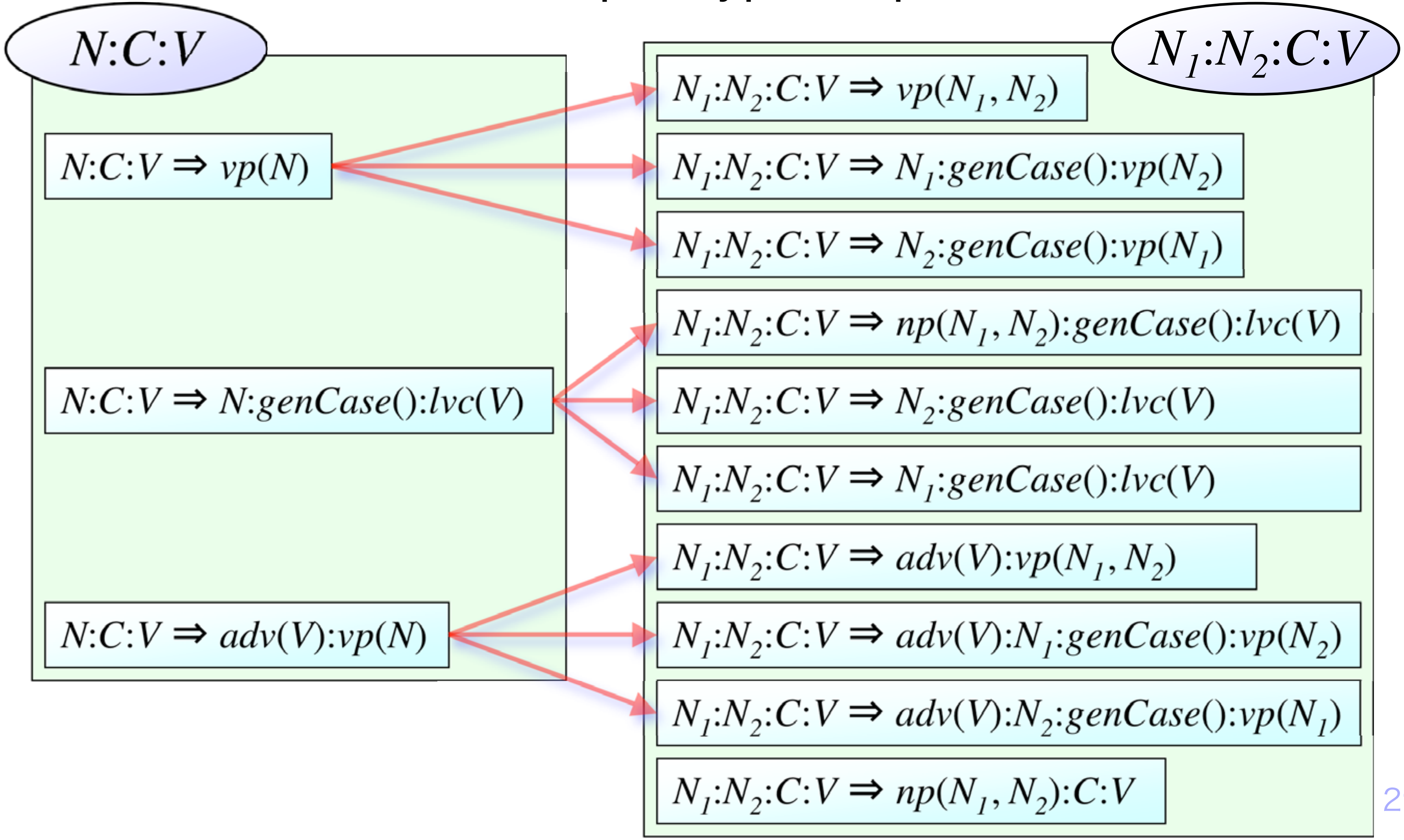
Gen. Func.
$vp(N)$

- Lexical function
  - Lexical derivation (10 dics, totally 6,322 word pairs)
  - Noun-to-interrogative (1)

Lex. Func.
$adv(V)$

To ensure coverage

1. Enumerate Trans. Pat. for N:C:V
2. Extend them for more complex types of phrases



$N{:}C{:}V$

$N{:}C{:}V \Rightarrow vp(N)$

$N{:}C{:}V \Rightarrow N{:}genCase(){:}lvc(V)$

$N{:}C{:}V \Rightarrow adv(V){:}vp(N)$

$N_1{:}N_2{:}C{:}V$

$N_1{:}N_2{:}C{:}V \Rightarrow vp(N_1, N_2)$

$N_1{:}N_2{:}C{:}V \Rightarrow N_1{:}genCase(){:}vp(N_2)$

$N_1{:}N_2{:}C{:}V \Rightarrow N_2{:}genCase(){:}vp(N_1)$

$N_1{:}N_2{:}C{:}V \Rightarrow np(N_1, N_2){:}genCase(){:}lvc(V)$

$N_1{:}N_2{:}C{:}V \Rightarrow N_2{:}genCase(){:}lvc(V)$

$N_1{:}N_2{:}C{:}V \Rightarrow N_1{:}genCase(){:}lvc(V)$

$N_1{:}N_2{:}C{:}V \Rightarrow adv(V){:}vp(N_1, N_2)$

$N_1{:}N_2{:}C{:}V \Rightarrow adv(V){:}N_1{:}genCase(){:}vp(N_2)$

$N_1{:}N_2{:}C{:}V \Rightarrow adv(V){:}N_2{:}genCase(){:}vp(N_1)$

$N_1{:}N_2{:}C{:}V \Rightarrow np(N_1, N_2){:}C{:}V$

The body of Lex. Func.

- IPADIC-2.7.0 + Mainichi 1999-2005 (1.8GB)

POS-pair	$\lvert D \rvert$	$\lvert C \rvert$	$\lvert D \cup C \rvert$	$\lvert J \rvert$	cleaning
noun - verb	3,431	-	3,431	3,431	
noun - adjective	308	667	906	475	done
noun - adjectival verb	1,579	-	1,579	1,579	
noun - adverb	271	-	271	271	
verb - adjective	252	-	252	192	done
verb - adjectival verb	74	-	74	68	done
verb - adverb	74	-	74	64	done
adjective - adjectival verb	66	95	159	146	done
adjective - adverb	33	-	33	26	done
adjectival verb - adverb	70	-	70	70	
Total	6,158	762	6,849	6,322	

Outline

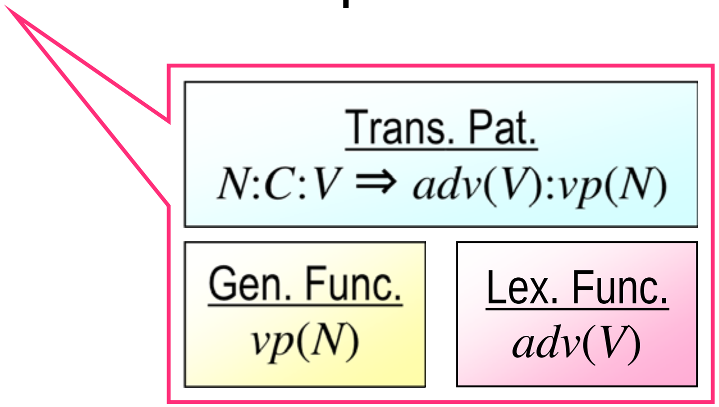
1. Motivation & Aim
2. Range of phenomena
3. System & implementation
4. Discussion
5. Conclusion

Discussion (≒ future work)

- Sufficient condition
  - Patterns does not ensure paraphrasability perfectly
  - Extensional definition of selectional preferences [Pantel+, 2007]
- Structured transformation
  - For flexible and accurate matching
  - Less impact due to short phrase
- Methodology of resource development
  - Modularization of Gen. Func. is inconsistent
  - Requires linguistic expertise
  - Simple KBs are preferable (cf. MTT)

Conclusion & Future work

- Notion of Phrasal Thesaurus is introduced
  - Compositional paraphrases of predicate phrases
  - Preliminary progress report of resource development
- Future work
  - Development
    - Resources
    - SLM (Structured, Web, etc.)
    - Applicability conditions
  - Intrinsic / extrinsic evaluation



Trans. Pat.

$N{:}C{:}V \Rightarrow adv(V){:}vp(N)$

Gen. Func.

$vp(N)$

Lex. Func.

$adv(V)$