

< CBA to Paraphrasing & Nominalization, Dec. 2nd, 2010 >

Typology of Paraphrases and Approaches to Compute Them

Atsushi FUJITA



Future University Hakodate, JAPAN

<http://paraphrasing.org/~fujita/>

What's paraphrase?

□ Intentional definition

- e.g., LDOCE

(v) to express in a shorter, clearer, or **different way** what someone has said or written

(n) a statement that expresses in a shorter, clearer, or **different way** what someone has said or written

What's paraphrase?

□ Extensional definition

- lexical, phrasal, sentential, discourse-level, ...
- covered all? well-organized?

The riddle is solved by me.

I solved the riddle.

Employment showed a sharp decrease.

Employment decreased sharply.

Emma burst into tears and he tried to comfort her.

Emma cried, and he tried to console her.

□ Scope / boundary

- Not precisely defined

I want some fresh air.

Could you open the window?

My son eats eggplants.

My son likes eggplants.

Typology of paraphrases

□ Axes

- Structure
- Required knowledge
- Application
- Sameness and difference of meaning

□ Guidepost

- To clarify how human beings process paraphrases
- To automate paraphrases (steadily)
 - ◆ Clarify required resources for each type
 - ◆ Modularize each type for selective use
- Artificial, so not be crazy

Goal of this talk

□ A survey

- Share the idea
- Discuss the way of creating typology
 - ◆ e.g., Axes
- Involve people for creating typologies
 - ◆ e.g., <http://paraphrasing.org/paraphrase.html>

Outline

- ▶ 1. Sameness of meaning
- 2. Linguistically-motivated typology
- 3. Paraphrases in apps
- 4. Computation
- 5. Future directions

Meaning of linguistic expressions

□ Semantics

- Formal semantics
- Situation semantics
 - ◆ Discourse representation theory [Kamp, 81]
 - ◆ Mental-space theory [Fauconnier, 85]
- Lexical semantics
 - ◆ Frame semantics [Fillmore, 76]
 - ◆ Lexical Conceptual Structure [Jackendoff, 90]
 - ◆ Generative Lexicon [Pustejovsky, 95]

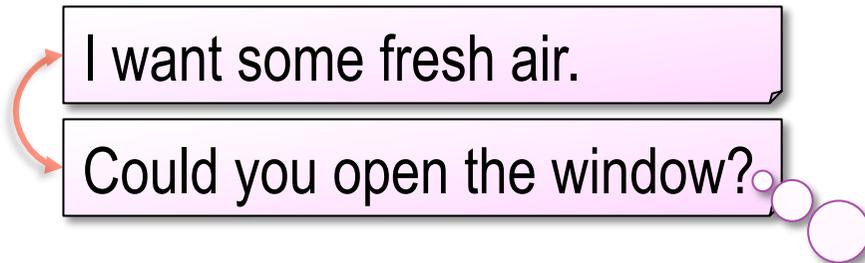
Paraphrase in semantics

- A good subject
 - To think of equality
 - Toward semantic computing
 - ◆ How to drive semantic frameworks

- Levels of sameness [Sato, 99]
 - Pragmatic meaning
 - Referential meaning
 - Denotation

Sameness of pragmatic meaning

□ Illocutionary / perlocutionary acts



- Various interpretation
 - ◆ But, only the speaker knows truth

Sameness of referential meaning

□ Coreference

Barça's #10 scored no goal in the last El Clásico.

Lionel Messi scored no goal in the last match against Real Madrid.

in 2008-2011

Barça's eye view

- May not true in the other situation
 - ◆ e.g., Ronaldinho, Riquelme, Rivaldo, ...
 - ◆ e.g., against Barça, between Barça and Real
- Discourse-level
 - ◆ incl. exophora
 - ◆ Cognitive meaning [Milićević, 07]

Sameness of denotation

□ Truth-value semantics



- Can be carried out
 - ◆ Without referring to the communicative situation
 - ◆ With linguistic knowledge
 - ◆ (With world knowledge)
- Have different connotation [Edmonds, 99][Inkpen+, 06]
 - ◆ Theme / Rheme
 - ◆ Formality
 - ◆ Emotion (attitude)

Remind the definition

□ It supposes **some differences**

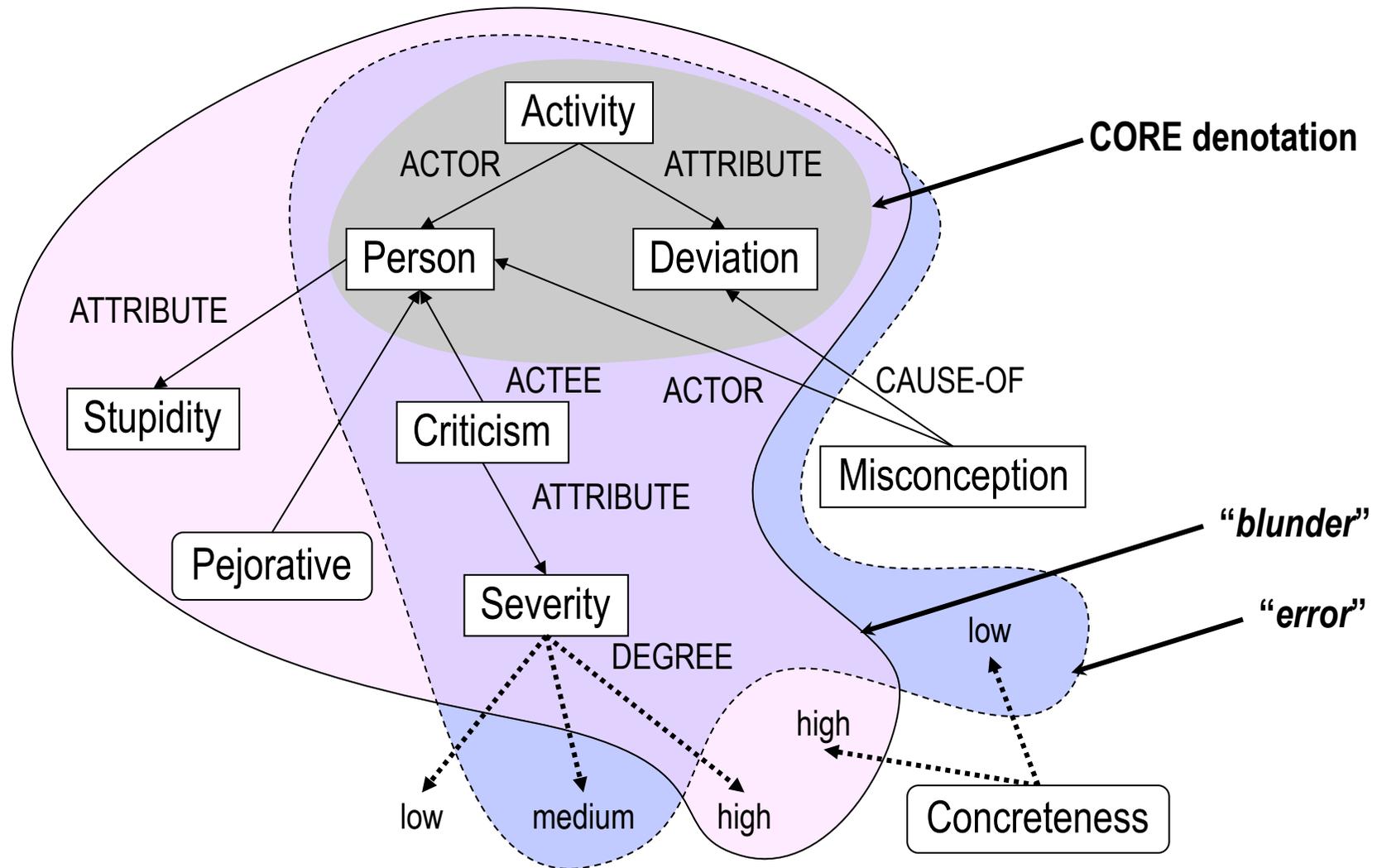
(v) to express in a **shorter, clearer**, or different way what someone has said or written

(n) a statement that expresses in a **shorter, clearer**, or different way what someone has said or written

- Not exactly same meaning (synonym) [Clark, 92]
- But near-synonym [Edmonds, 99]

Comparison of “blunder” and “error”

[Edmonds, 99]



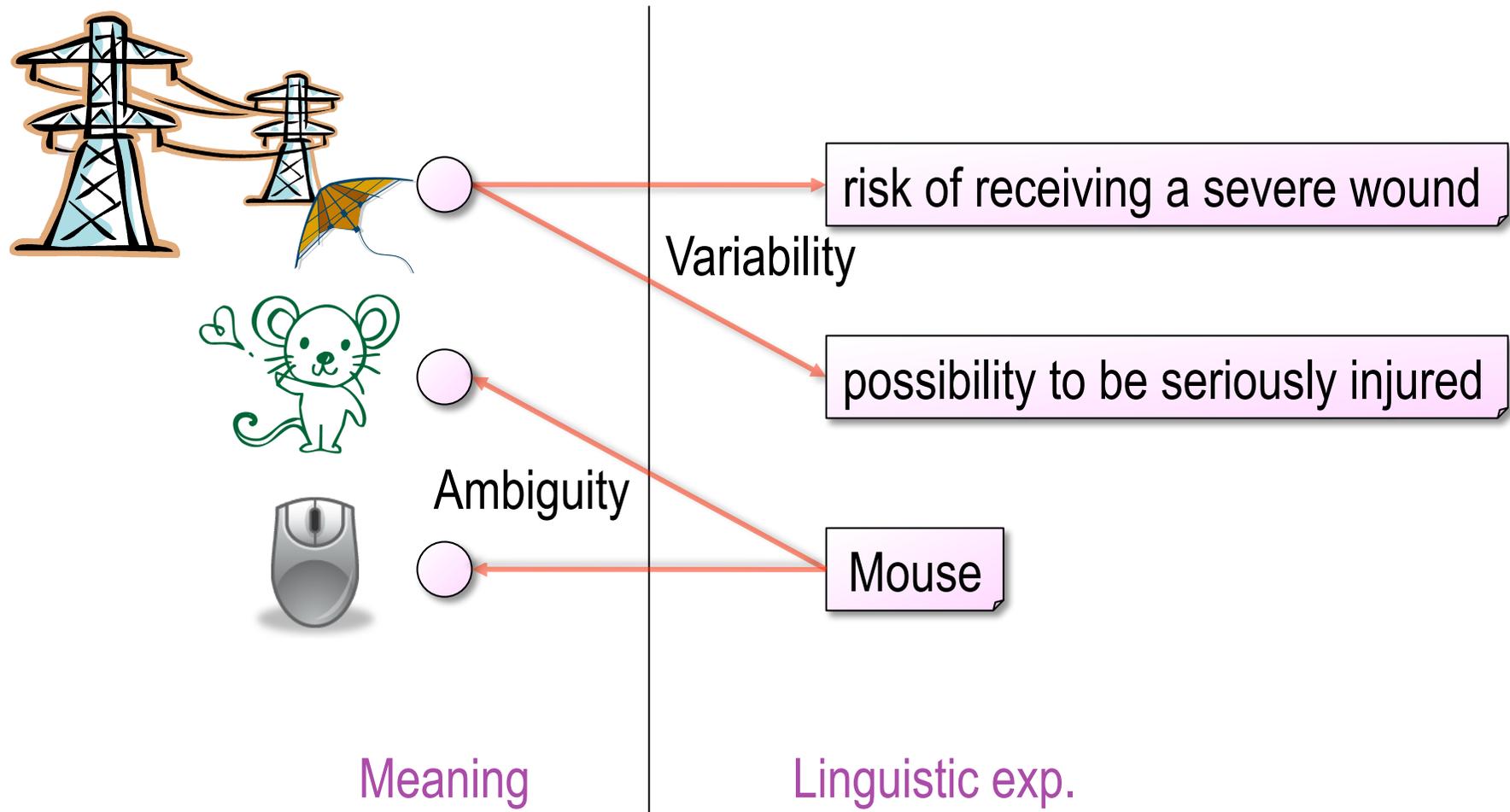
Aim of making difference

- What's changed?
 - complex → simple
 - verbose → clear
 - marked → unmarked
 - emotional → neutral

- Reasons why we paraphrase
 - To facilitate communication
 - ◆ For confirmation
 - ◆ For accelerating understanding
 - To strengthen the solidarity in a community

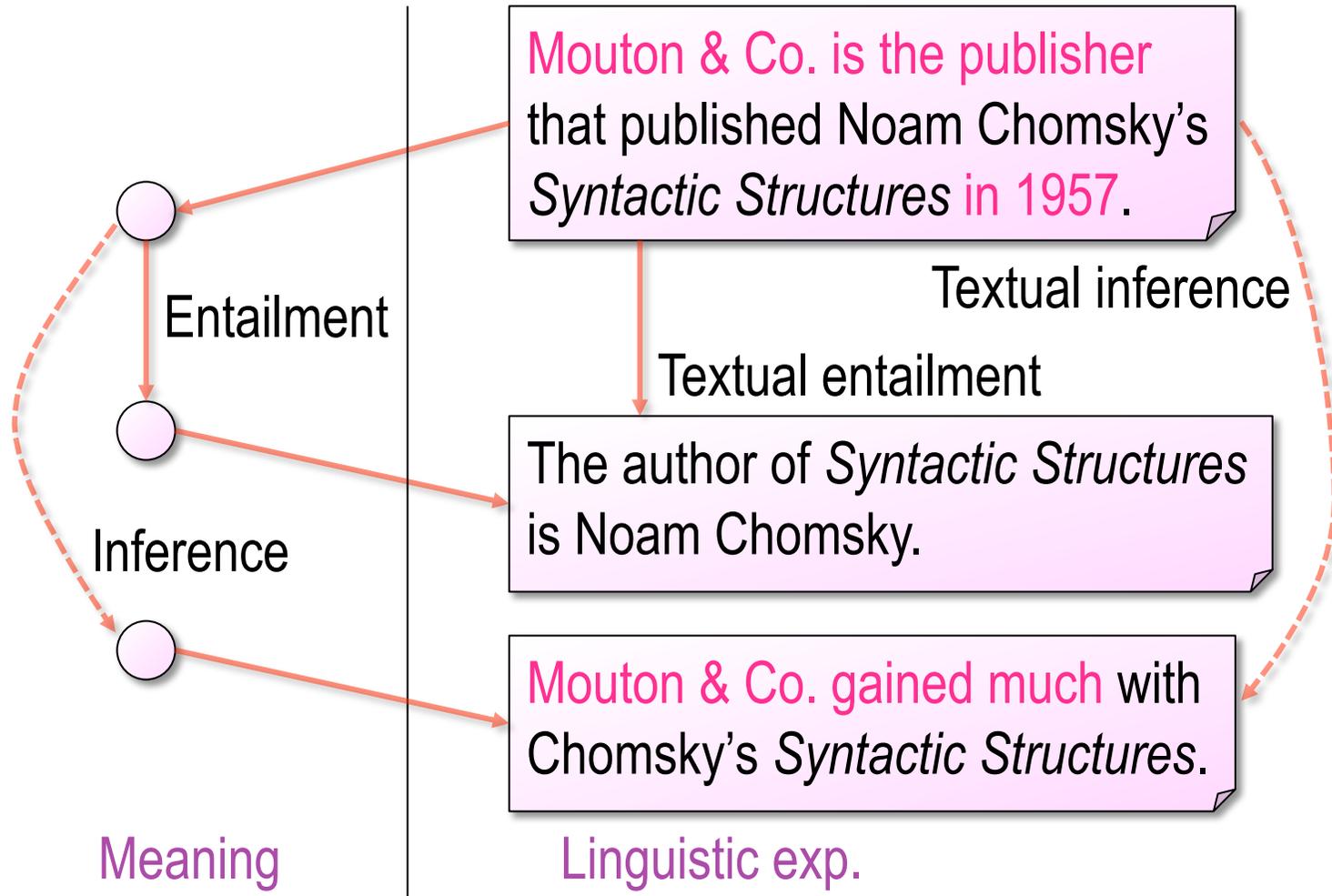
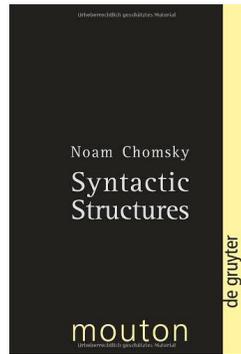
Paraphrase

- Linguistic variability in conveying a meaning



Entailment / Inference

- Relation between different meanings



Entailment

□ Not necessarily same meaning

○ $X \rightarrow Y$

Mouton & Co. is the publisher that published Noam Chomsky's *Syntactic Structures* in 1957.

The author of *Syntactic Structures* is Noam Chomsky.

○ e.g., lexical entailment in WordNet [Miller+, 85]

march → walk Troponymy

forget → know Backward presupposition

○ has started → started Temporal

Inference

- Not ensure even truth

Mouton & Co. is the publisher that published Noam Chomsky's *Syntactic Structures* in 1957.

Mouton & Co. gained much with Chomsky's *Syntactic Structures*.

- But useful in some situations [Pantel+, 07]

My son eats eggplants.

My son likes eggplants.

Everything is imported to Japan.

Everything is eaten in Japan.

Summary

- Levels of sameness [Sato, 99]
 - Pragmatic meaning
 - Referential meaning
 - Denotation

- Related concepts
 - Entailment: paraphrase \Leftrightarrow bi-directional entailment
 - Inference: entailment \supset always-true inference

Outline

1. Sameness of meaning
- ▶ 2. Linguistically-motivated typology
3. Paraphrases in apps
4. Computation
5. Future directions

Rough classification

□ Names used in papers

- Lexical / Phrasal
- Syntactic
- Sentential

□ Classification in [IWP, 2005]

- Phrase-level
- Sentence-level
- Discourse-level

Not necessarily atomic, because methods and results are centered

Our linguistically-motivated typology

- Focused on denotation
 - Explainable referring to
 - ◆ The given context
 - ◆ Linguistic knowledge
 - Ignored differences in connotation

- 5 types based on
 - Influenced scope
 - Generality (or productivity)

[A] Extra-sentential paraphrase

□ Clause separation (relative clause)

Småland, **which is** located to the south-west of Stockholm, **is** called “The Kingdom of Glass”. The reason is that there are sixteen glass manufacturers in this area.

Småland **is** located to the south-west of Stockholm. **It is** called “The Kingdom of Glass”. The reason is that there are sixteen glass manufacturers in this area.

□ Conjunction replacement

Note down the number. **Otherwise**, you may forget it.

Note down the number. **If not**, you may forget it.

[B] Extra-clausal paraphrase

□ Cleft → non-cleft

It was his best suit that John wore to the dance last night.
John wore his best suit to the dance last night.

□ Head-switch (clausal complement ↔ modifier)

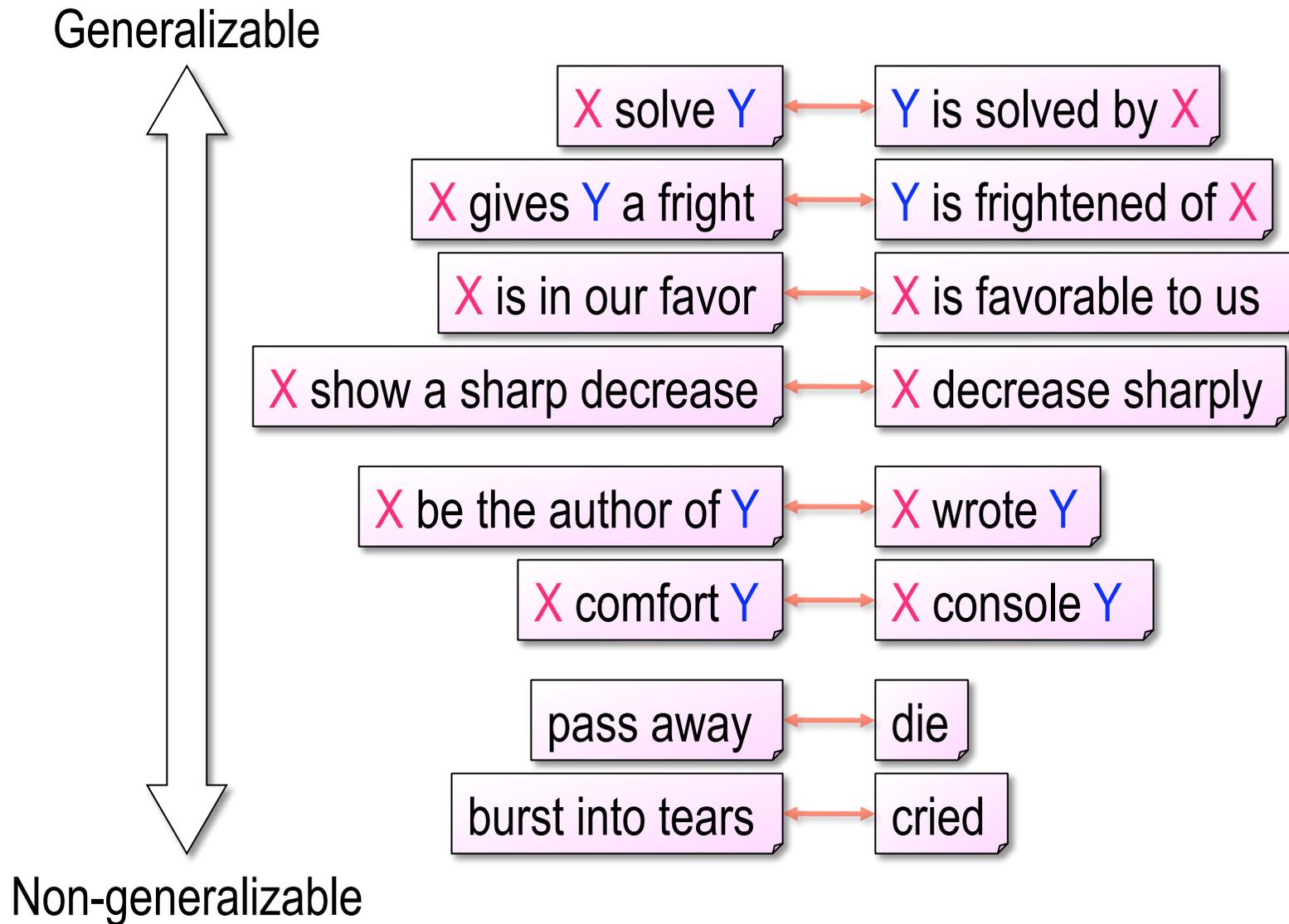
The conference venue is the building whose roof is red.
The conference venue is the building with red roof.

□ Move of negation

Your application is canceled if you do not reply.
Your application is not canceled if you reply.

□ Embedded ↔ coordinate, reordering, etc.

[C, D, E] Intra-clausal paraphrases



[C] Pure syntactic paraphrase

□ Inversion

If I **had** money enough, ...

Had I money enough, ...

Independent of the succeeding clause

□ Move of adverb

She can speak English **fluently**.

She can **fluently** speak English.

□ Paraphrase of negation

He **drank nothing but** famous **spirits**.

All he drank were famous **spirits**.

□ Less variation

[E] Lexical paraphrase

□ Not generalized at all

- → Need to collect thoroughly

There's a **risk** of receiving a **severe wound**.

N, Adj

There's a **possibility** of receiving **serious injure**.

Emma **burst into tears** and he tried to **comfort** her.

V, VP

Emma **cried**, and he tried to **console** her.

- Regards this as lexical?

- ◆ It's indecomposable any more

Real Sociedad **snapped a two-game losing streak**.

large VP

Real Sociedad **got points for the first time in three games**.

[D] Morpho-syntactic paraphrase

- Seems to be syntactic paraphrase
 - But have lexical constraints to some degree

John smeared **paint** on **the wall**.

John smeared **the wall** with **paint**.

Employment **showed a decrease**.

Employment **decreased**.

- Required information
 - ◆ Lexico-semantic information
 - Fine-grained argument structure
 - Lexical derivation, antonym, etc.
 - ◆ Selectional preference, collocation

Kinds of [D]: Verb alternation

[Levin, 93]

□ Passive to active

The riddle **is solved by** him.

He **solved** the riddle.

□ Dative alt.

Bill sold **a car** to **Tom**.

Bill sold **Tom** **a car**.

□ Locative alt.

John smeared **paint** on **the wall**.

John smeared **the wall** with **paint**.

□ Source alt.

The well gushed **oil**.

Oil gushed from **the well**.

□ Reciprocal alt.

The car collided **with the bicycle**.

The car and the bicycle collided.

□ Transitivity alt. (entailment)

Janet broke **the cup**.

The cup broke.

Kinds of [D]: Category shift

- Light-verb construction (N ⇔ V), A ⇔ Adv

Employment **showed** a **sharp** decrease.

Employment **decreased** **sharply**.

- Adj ⇔ V

I visited a priest in the **old** temple.

I visited a priest in the **olden(ed)** temple.

- Adj ⇔ N

I feel **drowsy**.

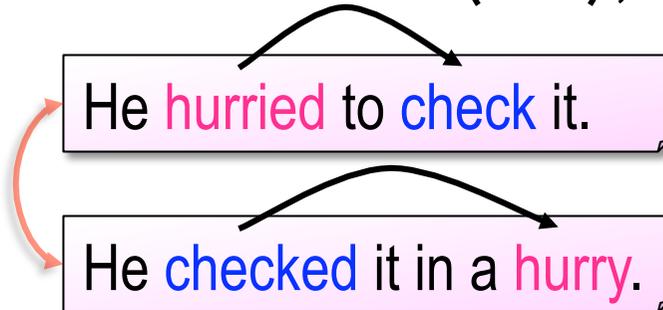
I have a **drowsiness**.

Kinds of [D]: Structural alternation

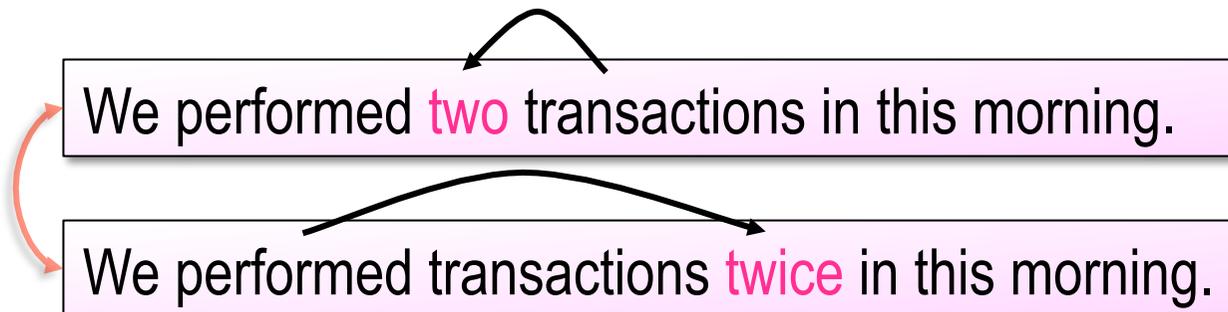
- Head-switch (NP), N \leftrightarrow V



- Head-switch (VP), V \leftrightarrow Adv, N \leftrightarrow V



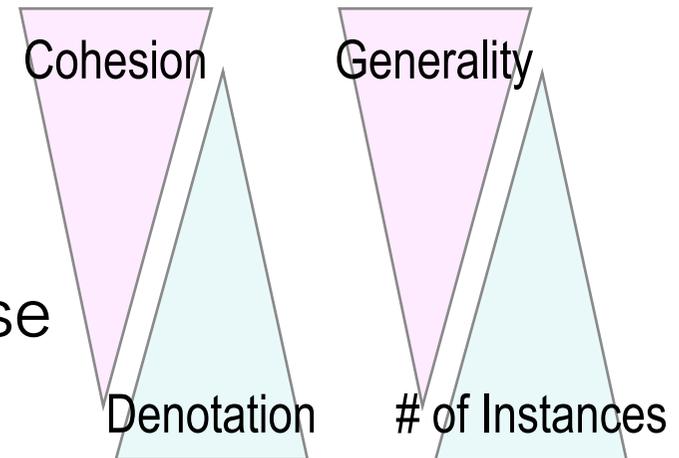
- Move of quantifier



Summary

□ A linguistically motivated typology

- [A] Extra-sentential
- [B] Extra-clausal
- [C] Pure syntactic
- [D] Morpho-syntactic paraphrase
- [E] Lexical (word, phrasal)



□ Focused on denotation

- Atomicity
- Scope
- Generality

Discussion and issues

□ On the typology

- Less [C] Pure syntactic paraphrases
 - ◆ After all, inter-clausal vs intra-clausal (within a VP)
- Treatment of indecomposable ones

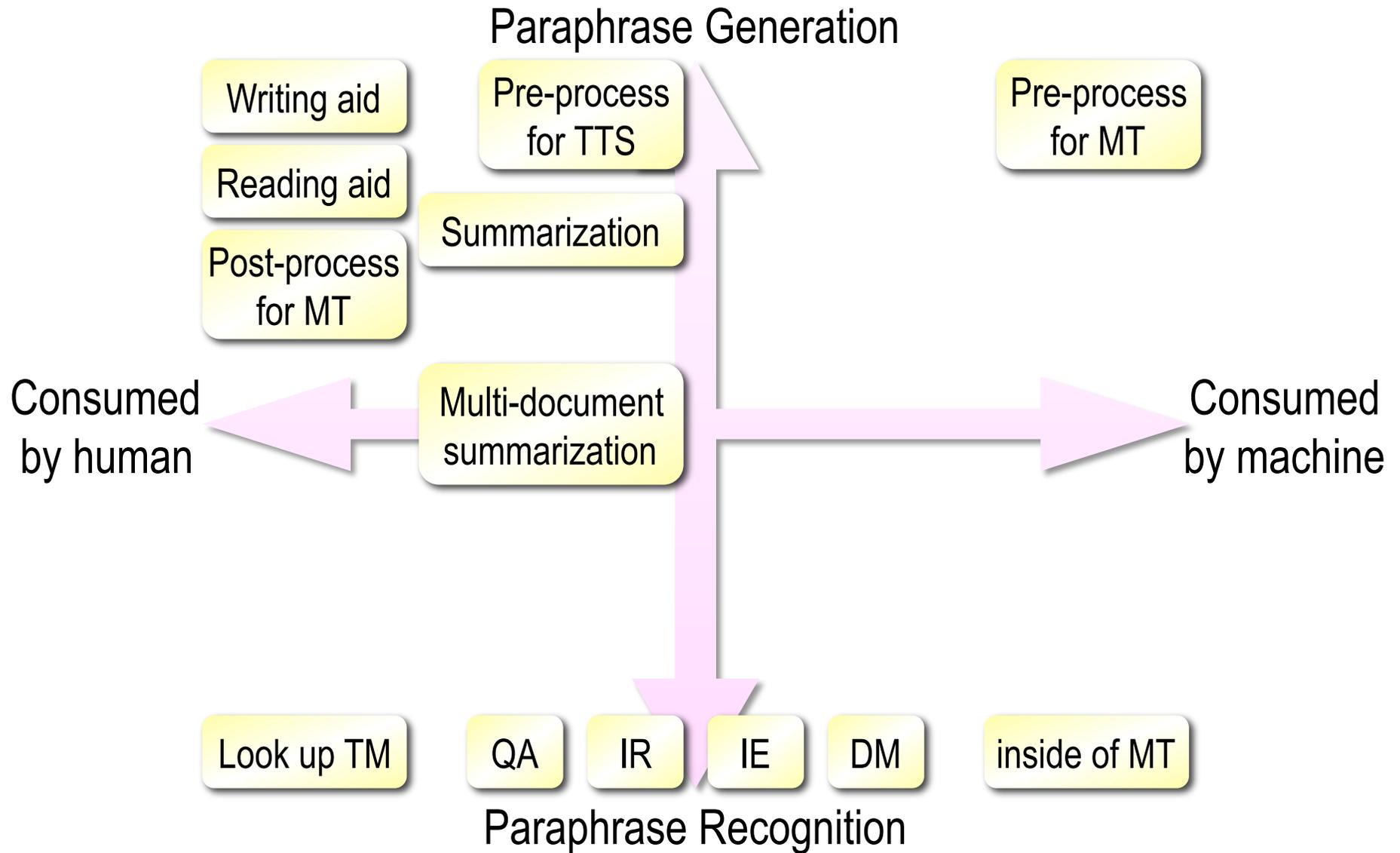
□ Lexical semantics for [D]

- FrameNet [Baker+, 98]
- VerbNet [Kipper+, 00]
- Lexical Conceptual Structure [Jackendoff, 91]
- Generative Lexicon [Pustejovsky, 95]

Outline

1. Sameness of meaning
2. Linguistically-motivated typology
- ▶ 3. Paraphrases in apps
4. Computation
5. Future directions

Map of apps



Requirements for application

- ❑ Target types of paraphrases
- ❑ Differences accepted
 - Connotation
 - ◆ Theme/Rheme
 - ◆ Formality
 - ◆ Emotion (attitude)
 - Denotation
 - ◆ Entailment
 - ◆ Inference
- ❑ Full-auto / consumed by human

Computing paraphrases for machine

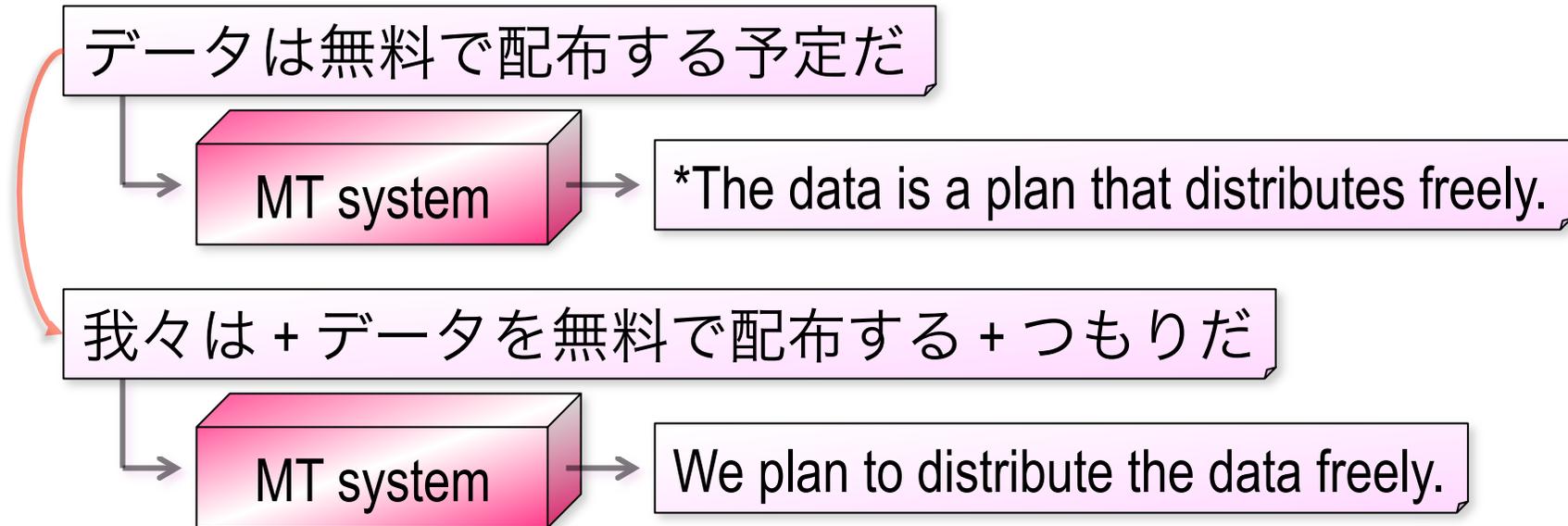
Multi-document summarization [Barzilay, 03]

The image illustrates the process of multi-document summarization. It features several browser screenshots:

- USATODAY.com - Syrian troops to begin pullback Monday**: The main article page. Key paragraphs are highlighted in pink:
 - "The withdrawal from central and northern Lebanon toward the Bekaa Valley will begin right after a meeting in Damascus, Syria, of the presidents of the two countries, Lebanese Defense Minister Abdul-Rahim Murad told The Associated Press. Assad and Lebanese President Emile Lahoud will decide on the timetable of the pullback and repositioning of forces."
 - "The Syrian withdrawal will begin Monday directly after the meeting in Damascus of the Syrian and Lebanese leaderships," Murad said.
- Columbia Newsblaster: Syria ready for Lebanon pullback**: A summary page titled "Summarizing all the news on the Web". It contains a condensed version of the news, with a paragraph highlighted in pink:
 - "Lebanese have poured into the streets to protest Syria's military presence in their small Mediterranean country. Syrian troops will begin pulling back to the Lebanese border following a Monday meeting of the countries' leaders."
- Other snippets**: Smaller screenshots at the bottom left show related news items, such as "Syria must 'fully withdraw ... secret services and intelligence officials that really keep the clamp of fear in the Lebanese people.'" and "15,000 troops after nearly 30 years in Lebanon."

Computing paraphrases for machine

□ Pre-edit for machine translation [Shirai+, 98]

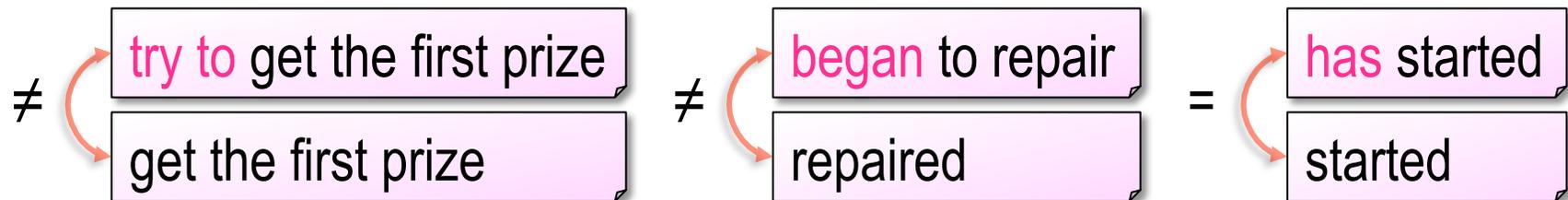


- Not only paraphrase, but also anaphora resolution
- Entailment / inference cannot be not applied

Computing paraphrases for machine

□ Data mining

- Summary of events [Izumi+, 10]
 - ◆ Light-verb construction
 - ◆ Keep factuality, but not some aspectual info.



- Collecting instances of plausible events
 - ◆ Discover unknown unknowns [Torisawa+, 08]
 - ◆ Build statement maps [Murakami+, 09]

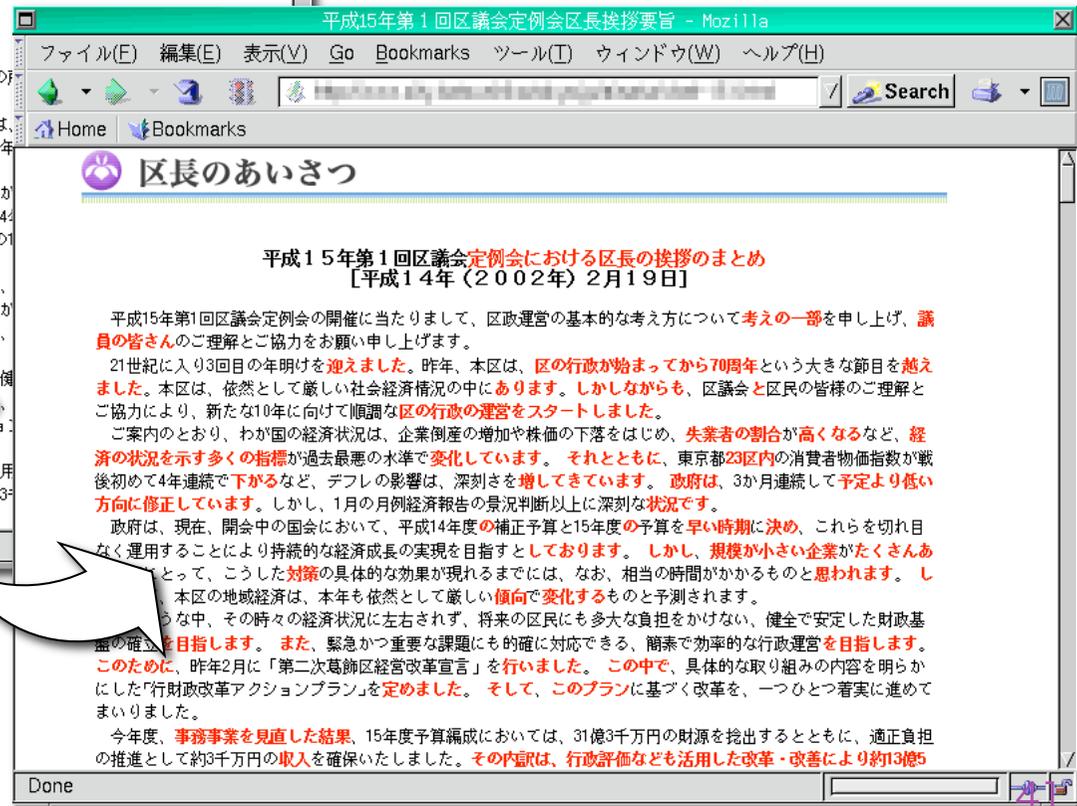
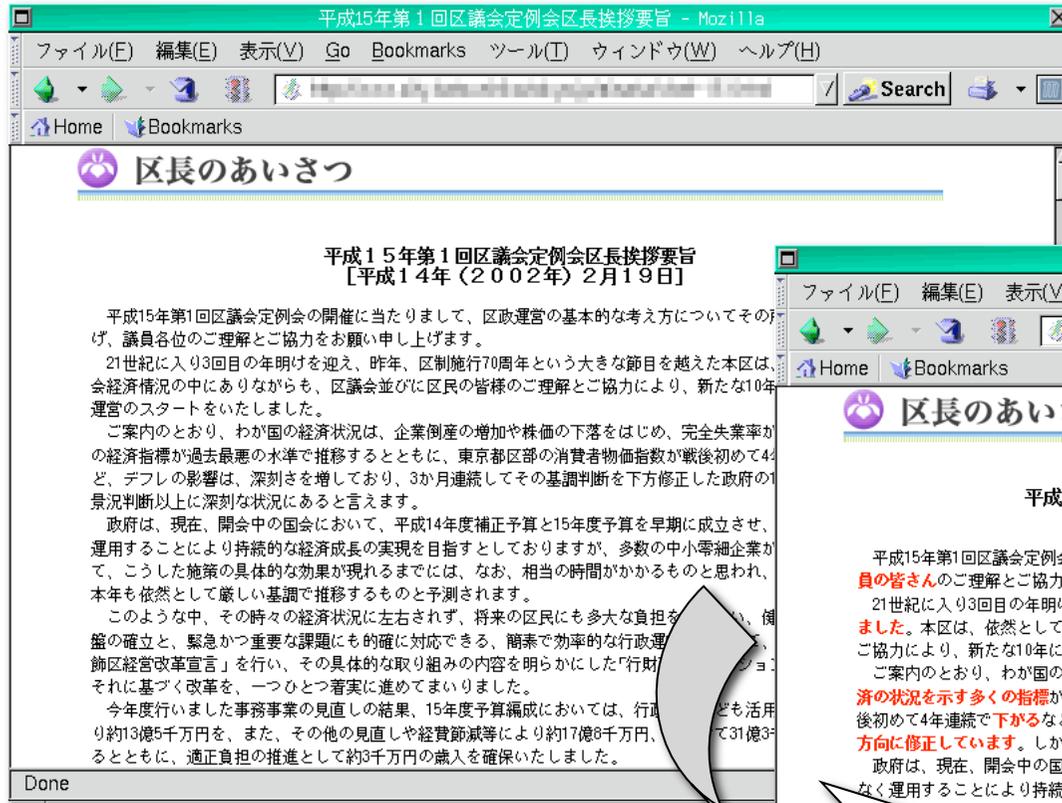
Computing paraphrases for human

- Writing aid (information dispatching aid)
 - Showing alternatives [Max+, 08]
 - ◆ Easier, clearer, more-decorative, etc.
 - Automatic rewrite
 - ◆ Normalization of specific documents
 - e.g., technical manuals, health reports

- Reading aid (information consuming aid)
 - Simplifying texts [Carroll+, 98][Canning+, 99][Inui+, 03]
 - Adding explanatory information
 - ◆ e.g., gloss of words, related terms

Computing paraphrases for human

□ Text simplification for reading aid [Inui+, 03]



Paraphrases in apps

- Typology and modularization are necessary

	IR	IE	DM	MT	Writing	Reading	
[A] Extra-sentential							
[B] Extra-clausal							
[C] Pure syntactic							
[D] Morpho-syntactic							
[E] Lexical							
Focus							
Formality							
Emotion							
Entailment							
Inference							

Outline

1. Sameness of meaning
2. Linguistically-motivated typology
3. Paraphrases in apps
- ▶ 4. Computation
5. Future directions

Toward full-automation of paraphrasing

Phase 1. Knowledge development

Acquisition

- Handcrafting patterns
- Automatic acquisition (corpus, Web)

Phase 2. Use of knowledge

Recognition
Generation

- Segmentation and disambiguation
- Applicability check in the given context
 - ◆ Grammaticality
 - ◆ Semantic appropriateness
 - ◆ Equivalency of meaning

Phase 3. Tuning for apps

- e.g., simplification, reduction of homonyms, etc.

Paraphrase Acquisition

1st phase toward
automatic paraphrasing

Previous work

□ Handcrafting patterns

- Transformation rules [Mel'cuk+, 87][Dras, 99][Jacquemin, 99]
- Thesaurus (of words) [A lot of work]

□ Automatic acquisition

- Distributional similarity in a single corpus [Lin+, 01][Torisawa, 01][Hagiwara+, 06], etc.
- Alignment of parallel/comparable/bilingual corpus [Barzilay+, 01][Shinyama+, 02][Pang+, 03][Ibrahim+, 03][Dolan+, 04][Bannard+, 05], etc.
- From the Web [Szpektor+, 04]

□ Implicit modeling

- Statistical translation model [Quirk+, 04][Bannard+, 05]
- Tree kernel [Collins+, 01][Takahashi, 05]

Handcraft rules/patterns

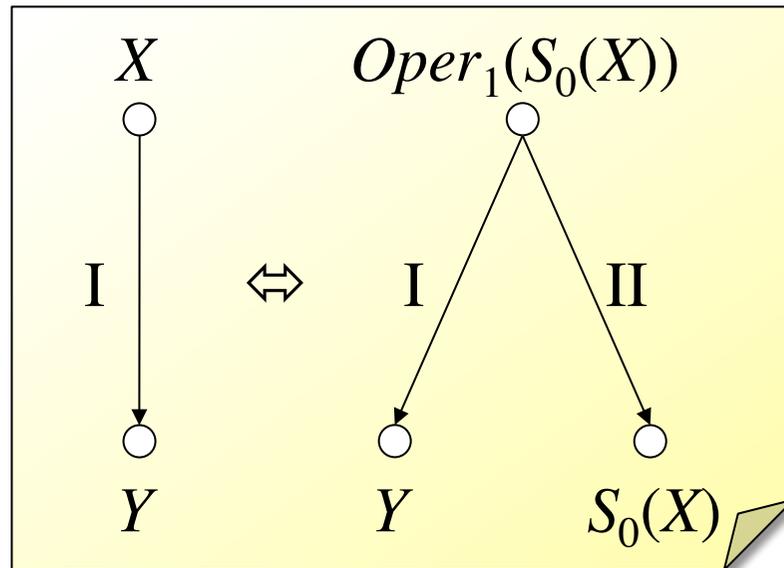
□ For a sentence

- Transformation grammar [Harris, 81]

NP1 V1 (+AUX) V2 (-AUX) NP2
 → NP2 V1 BE V2-PP by NP1

Active → Passive

- Meaning-text Theory [Mel'čuk+, 87]



VP ⇔
 Light-verb construction

- Various types of rules [Takahashi+, 01]

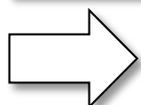
Extract from thesaurus

□ Near-synonyms: words within the same synset

- e.g., WordNet [Miller+, 85]

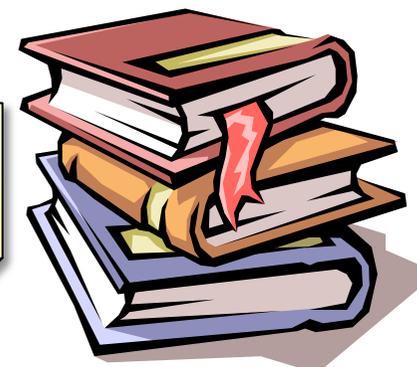
02526085: achieve, accomplish, attain, reach

05793554: basis, base, cornerstone, foundation, ...



achieve ⇔ accomplish

base ⇔ basis



- Just near-synonym [Clark, 92]

- ◆ Subtle difference [Edmonds, 99]

- ◆ Static synonymy apart from context [Fujita+, 00]

- How to enlarge thesaurus?

- ◆ Neologisms

google (v) ⇔ search Web using Google

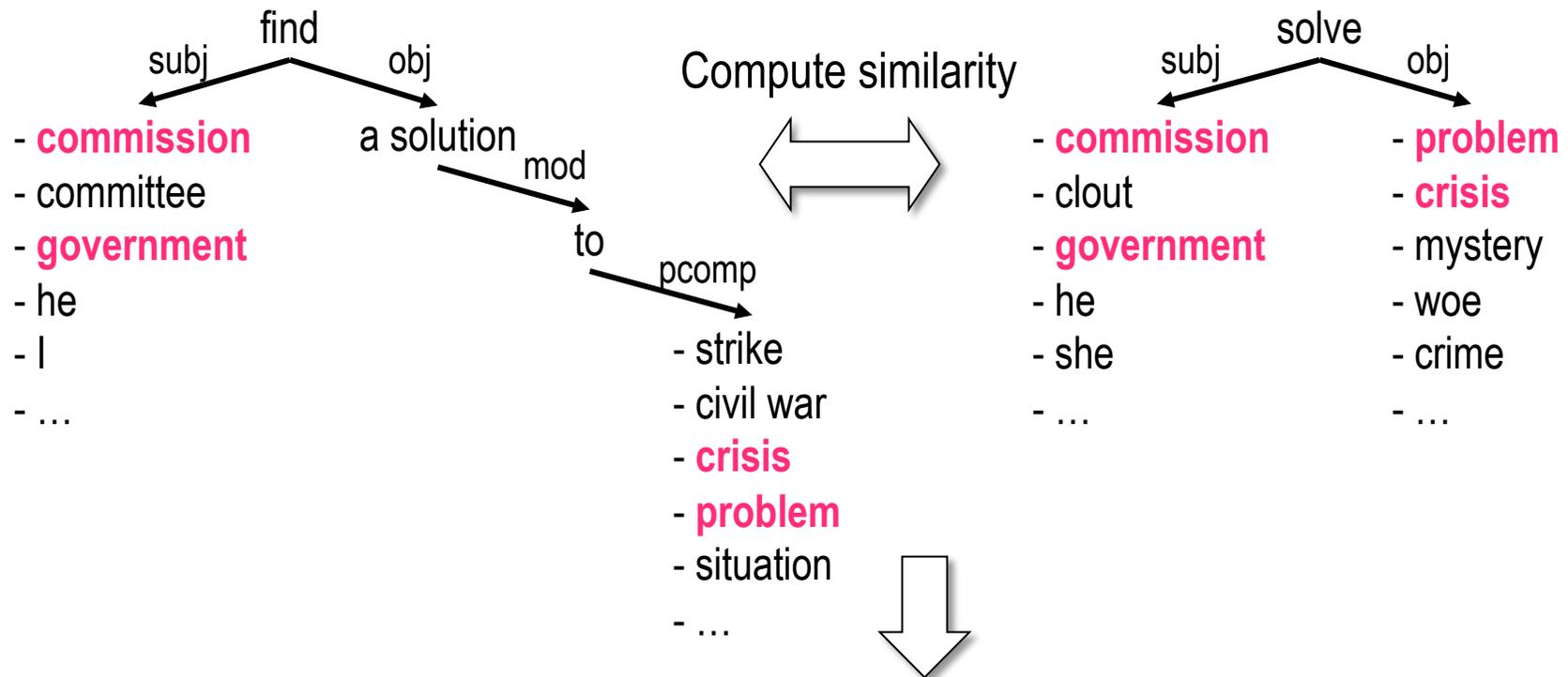
- ◆ Named entities

Future University Hakodate ⇔ FUN

Extract from single corpus

□ Distributional hypothesis [Harris, 64]

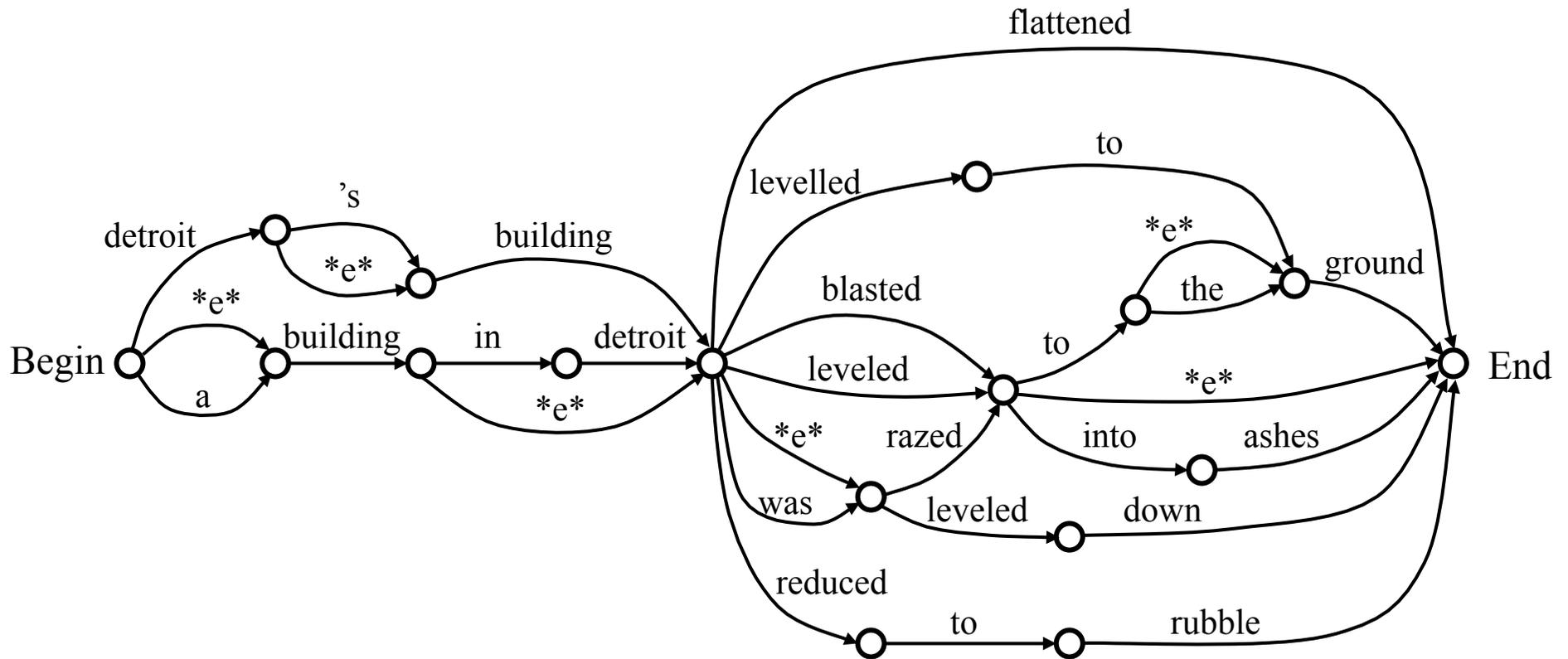
- Semantically similar words tend to appear in similar contexts.
- e.g., VP ← NP [Lin+, 01][Torisawa, 02]



X find a solution to Y ⇔ X solve Y

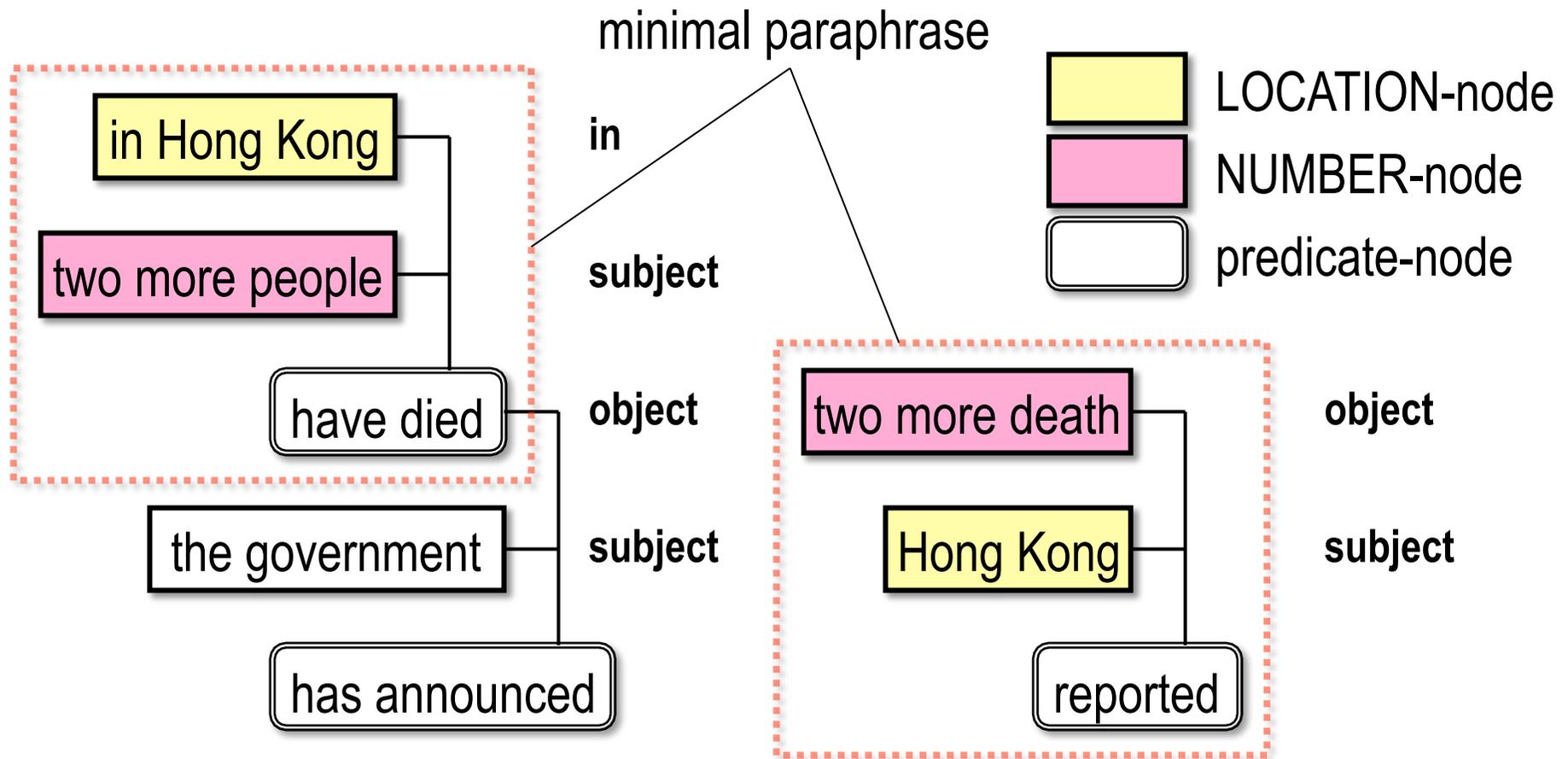
Extract from parallel corpus

- With multiple-sequence alignment
 - Multiple verbalizations of proofs [Barzilay+, 03]
 - Multiple translations [Pang+, 03]



Extract from comparable corpus

- News articles reporting the same event
 - Named entities as anchor [Shinyama+, 02]



Extract from bilingual corpus

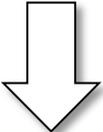
- Phrases translated into the same phrase
 - Translation table of SMT [Bannard+, 05]

what is more, the relevant cost dynamic is completely under control

im ubrigen ist die diesbezugliche kostenentwicklung vollig unter kontrolle

wir sind es den steuerzahlern schuldig die kosten unter kontrolle zu haben

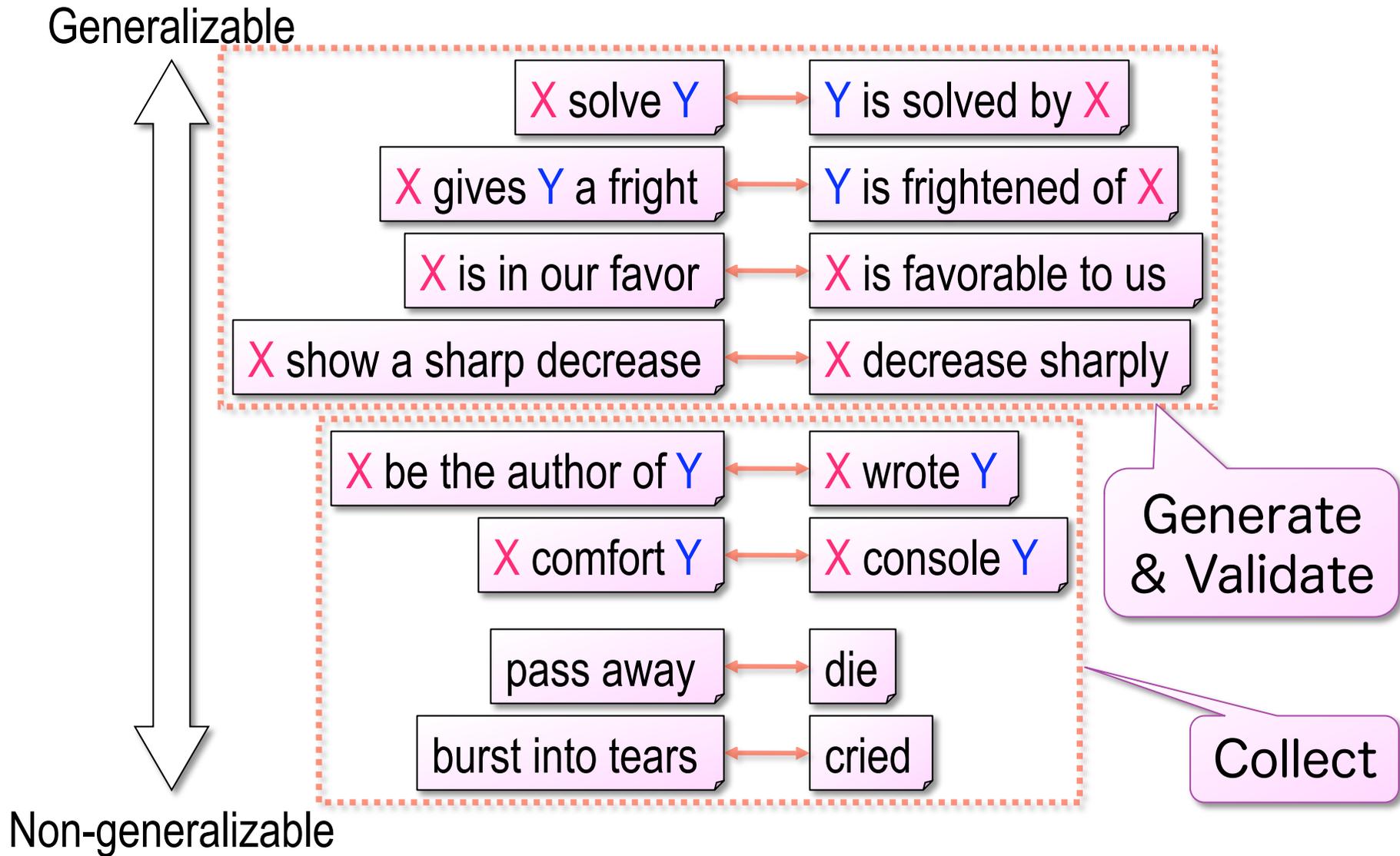
we owe it to the taxpayers to keep the costs in check



under control ⇔ in check

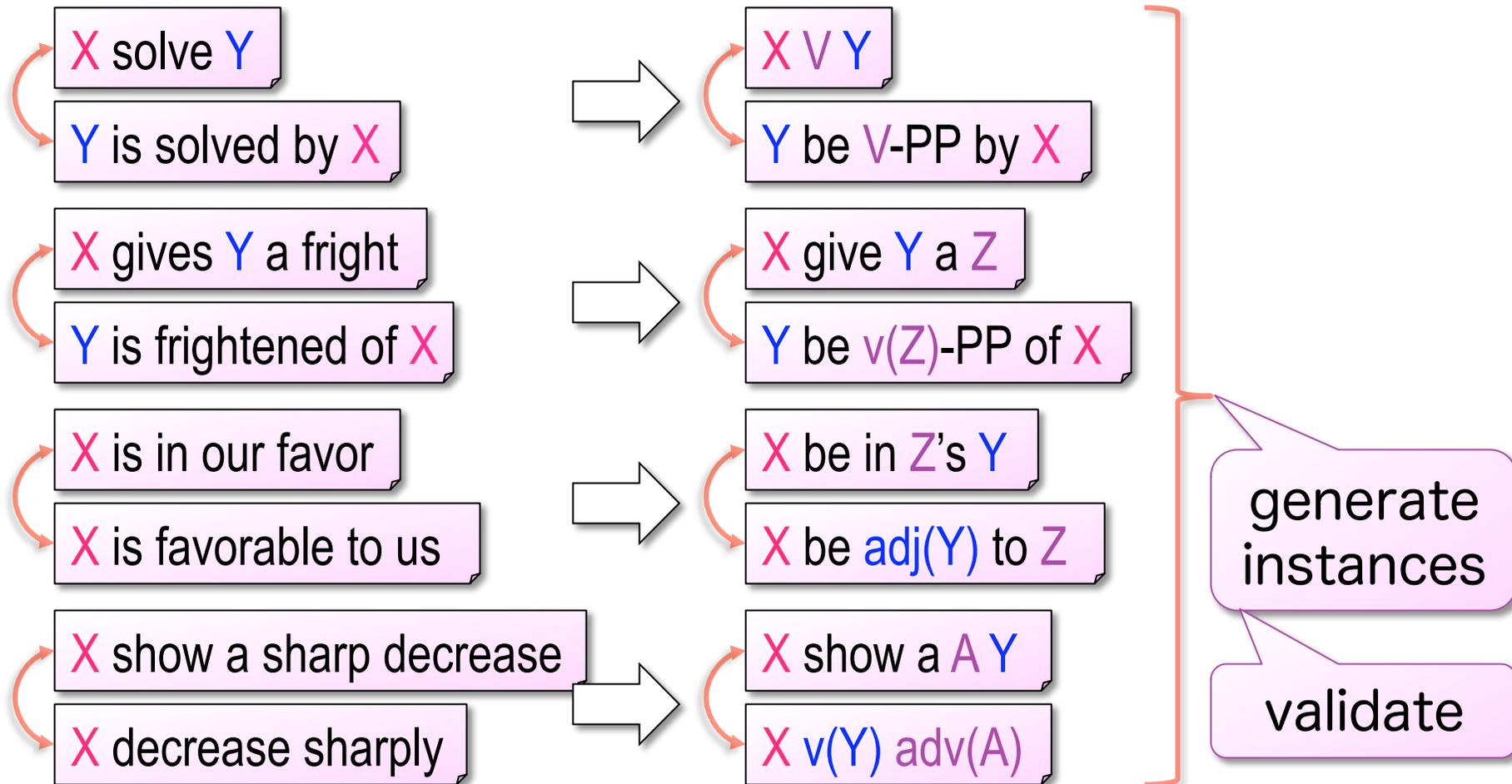
Knowledge for Intra-clausal paraphrases

[C-E]



Generate morpho-syntactic paraphrases

- Generation of knowledge [Fujita+, 07;08]
 - Syntactic transformation + Lexical derivation



Issues and current status

□ Issues

- How to cover various types of paraphrases?
→ e.g., knock off each type (typology-based)

□ Current status

Type	Handcraft	Corpus	Combi
[A] Extra-sentential	○	—	—
[B] Extra-clausal	○	△	—
[C] Pure syntactic	○	△	—
[D] Morpho-syntactic	△	△	○
[E] Lexical	—	○	—

Annotations:

- Manageable: points to Handcraft column (rows [A]-[D]).
- Too noisy: points to Handcraft column (row [D]).
- Promising: points to Corpus column (rows [B]-[D]).
- Promising: points to Combi column (row [D]).
- Low coverage: points to Combi column (row [E]).

Toward full-automation of paraphrasing

Phase 1. Knowledge development

Acquisition

- Handcrafting patterns
- Automatic acquisition (corpus, Web)

Phase 2. Use of knowledge

Recognition
Generation

- Segmentation and disambiguation
- Applicability check in the given context
 - ◆ Grammaticality
 - ◆ Semantic appropriateness
 - ◆ Equivalency of meaning

Phase 3. Tuning for apps

- e.g., simplification, reduction of homonyms, etc.

Two aspects of paraphrasing

□ Paraphrase recognition/identification

- Given pair of linguistic expressions \rightarrow label $\in \{=, \neq\}$
 - ◆ Theme of machine learning research

\langle give an advice , advise $\rangle \rightarrow =$

\langle give a copy , copy $\rangle \rightarrow \neq$

\langle make a copy , copy $\rangle \rightarrow =$

□ Paraphrase generation

- Numerous outputs

investigate the cause of a fire

- ◆ incl. unseen expressions

investigate why there was a fire

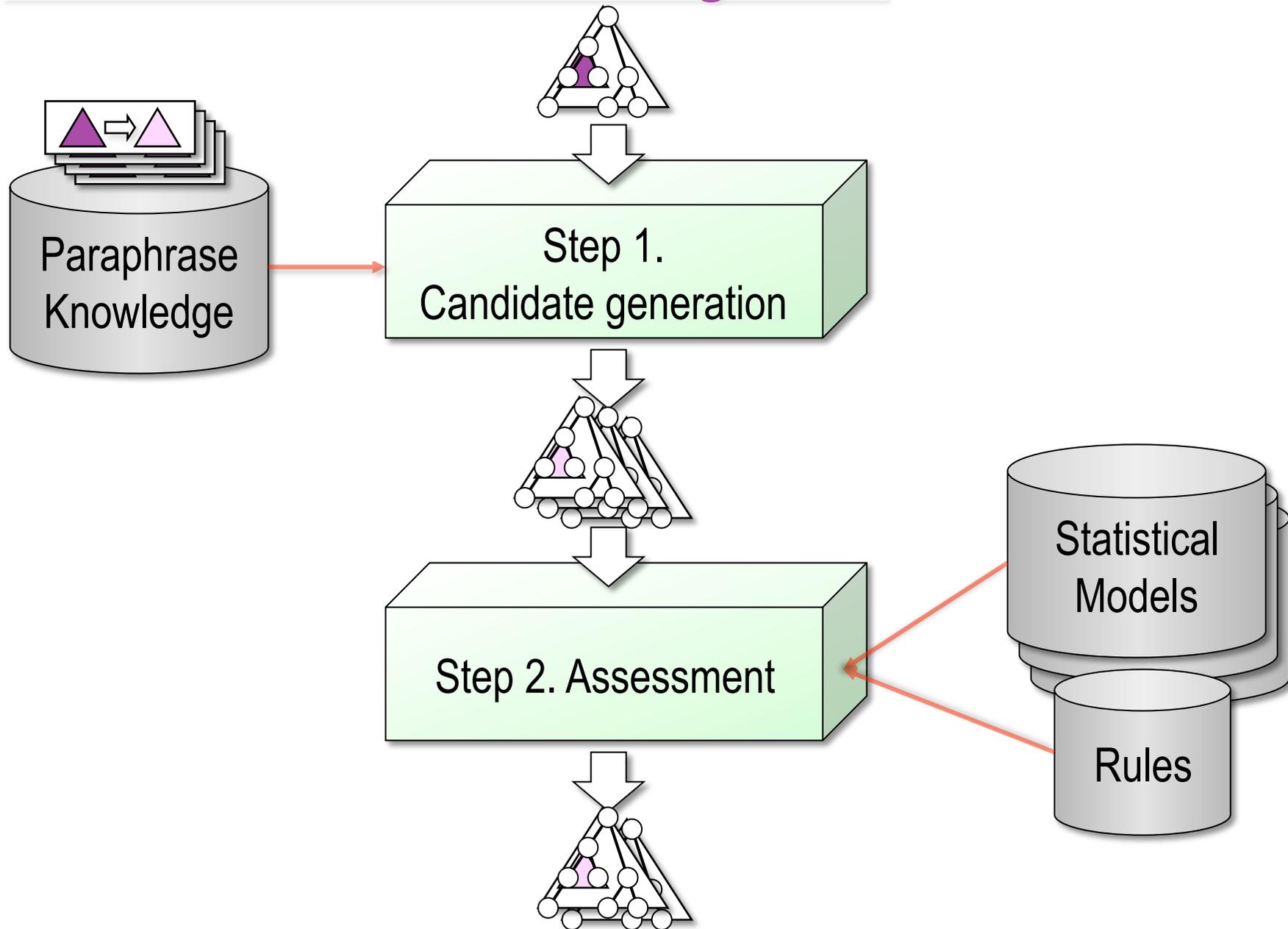
investigate what started a fire

make an investigation into the cause of a fire

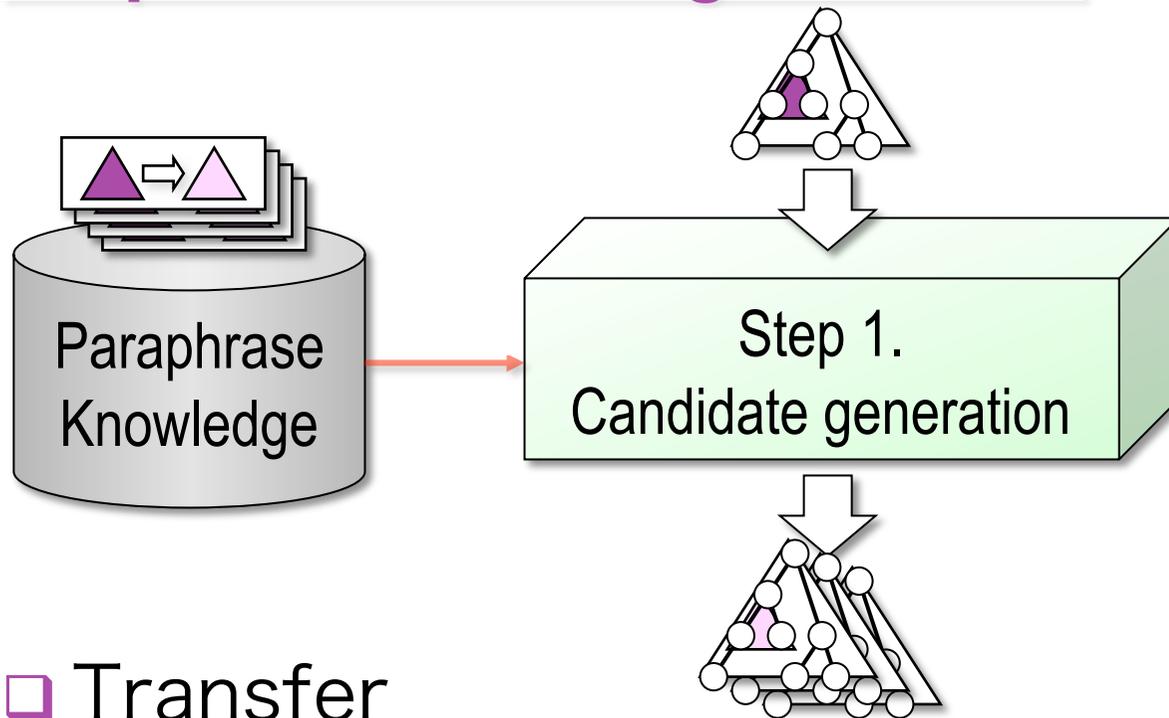
Paraphrase Generation

Example of 2nd phase toward
automatic paraphrasing

Generation includes recognition



Step 1. Candidate generation



□ Transfer

- Approach to MT in '70~'80
 - ◆ Assume compositionality
 - ◆ Substitute parts of input structure

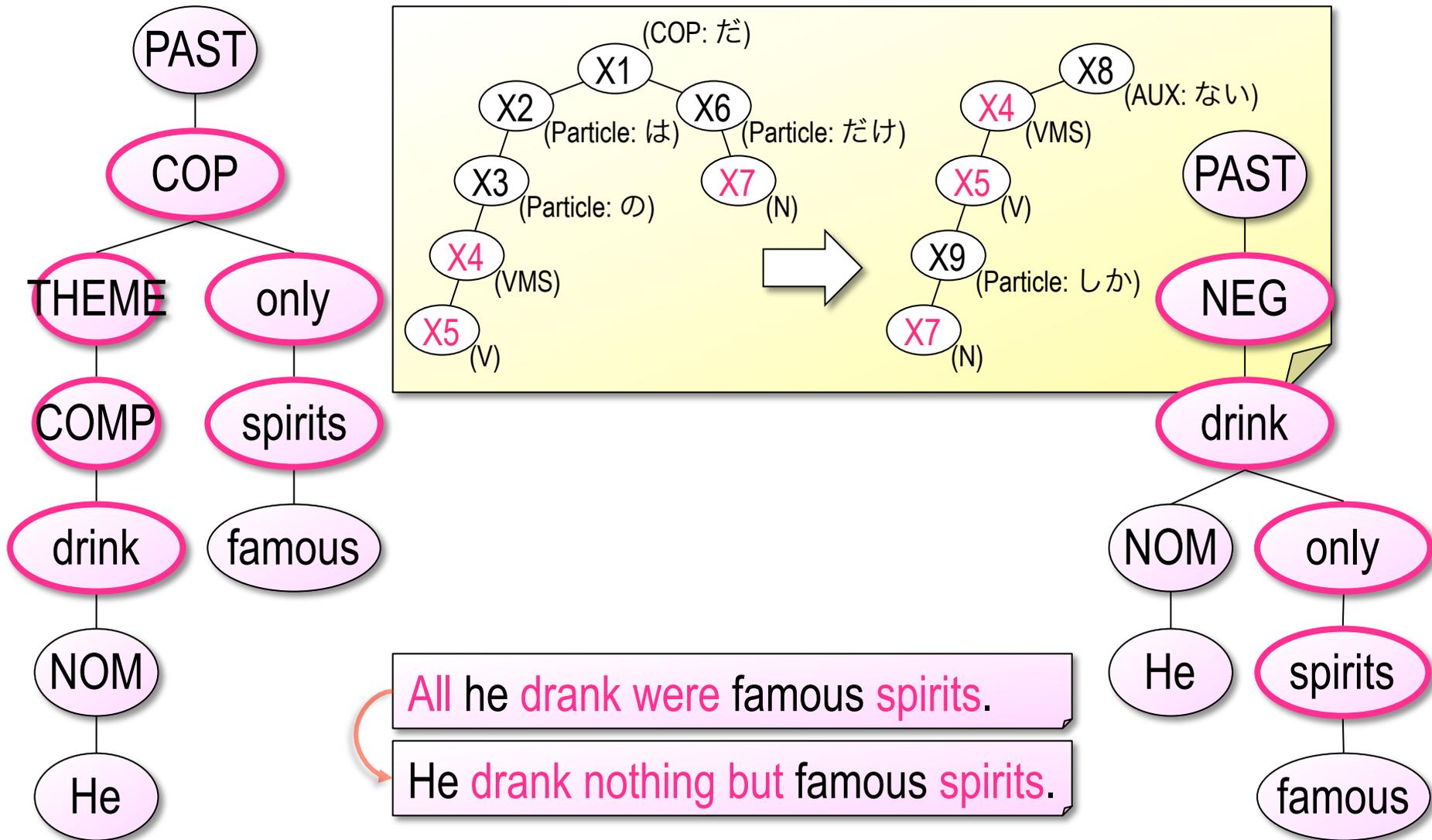
□ Transducer

- Accept sequence (structure is encoded)

Transfer over dependency tree

[A-E]

[Takahashi+, 01]



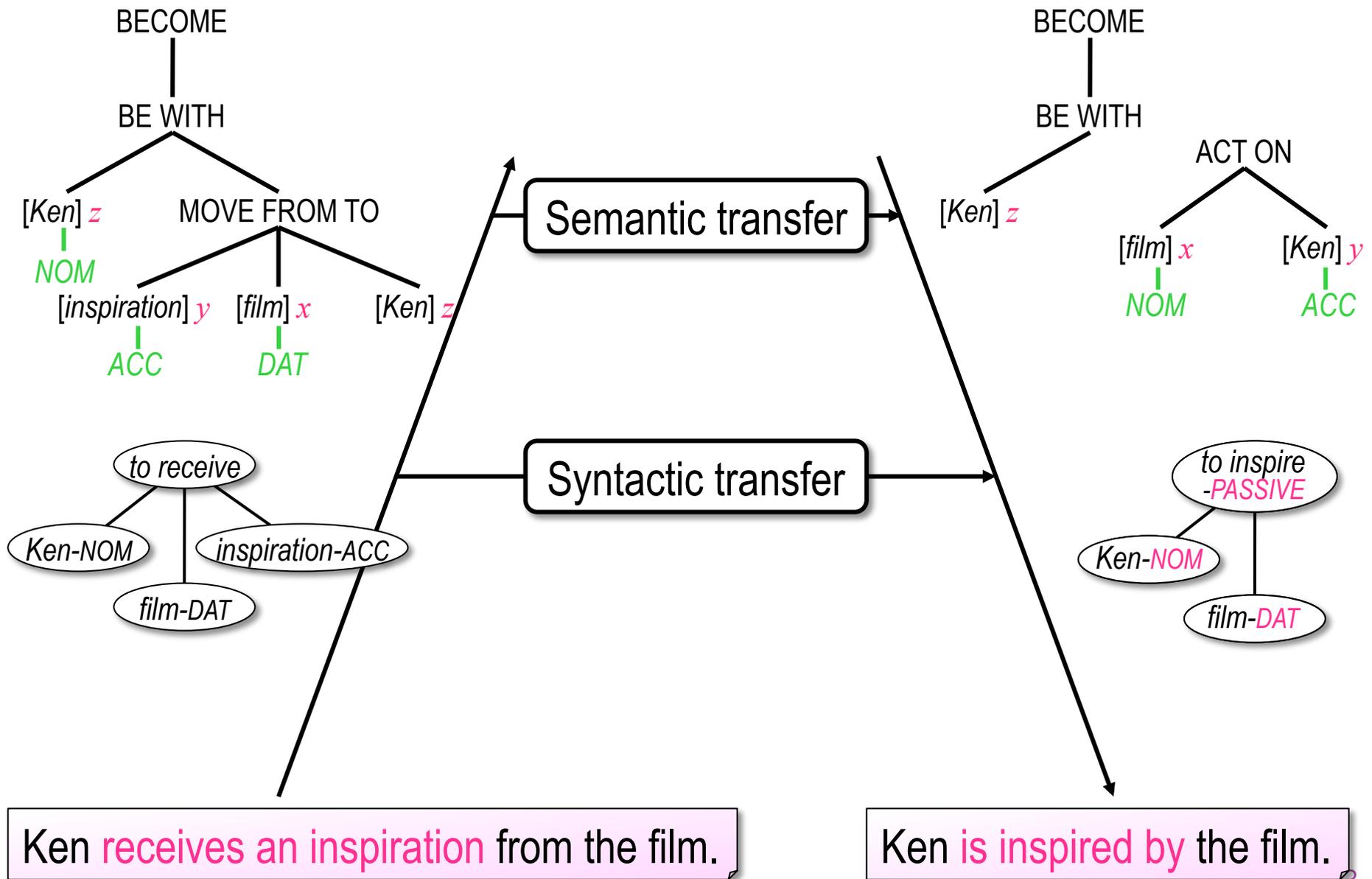
Transfer

- At the (shallow) syntax level
 - Minimal standard for various apps
 - Backed up by matured parsing technology
 - Many acquisition methods work at the same level

- Discussion
 - How wide range can be realized at this level?
 - How semantic constraints are incorporated?
 - ◆ e.g., lexical semantics for [D]
 - ◆ Leave until the assessment step?

Transfer guided by lexical semantics

[Fujita+, 04]



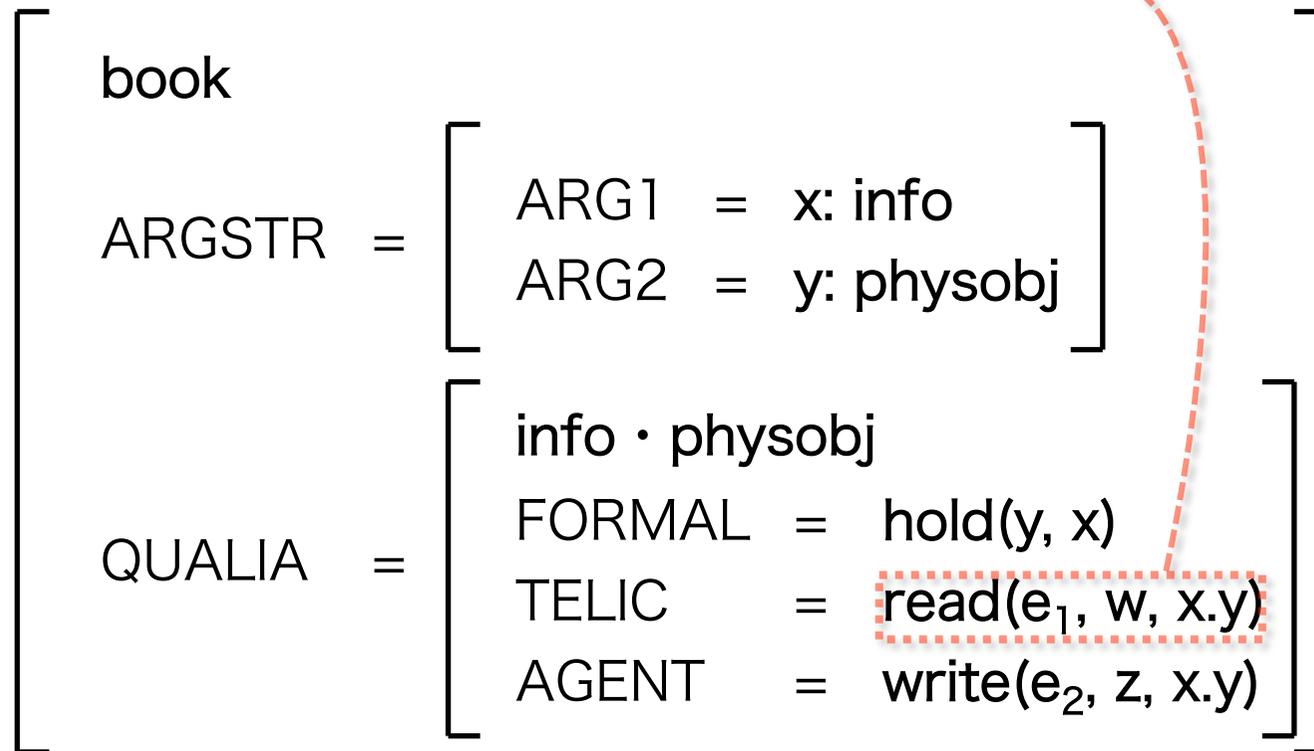
Equivalence explained by lexical semantics

□ Recovering meaning using GL framework

- Computing metonymy and default

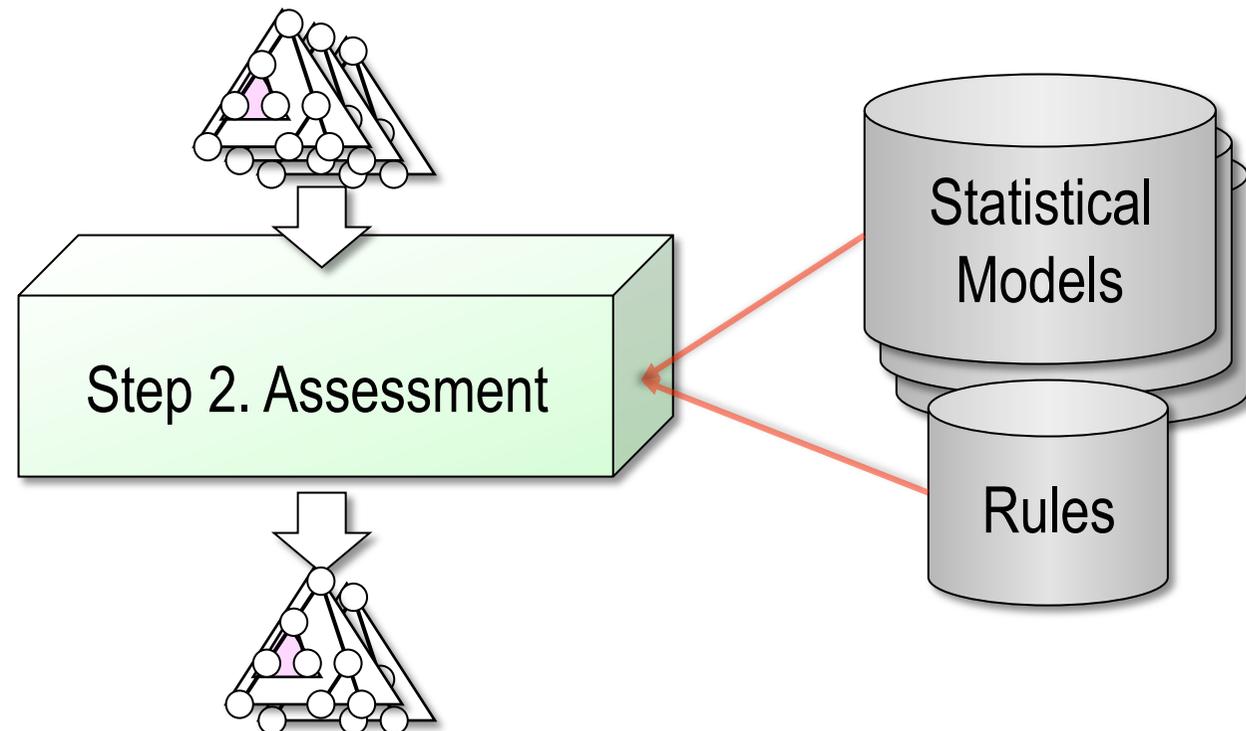
[Vila+, soon]

John began the book.
John began **reading** the book.



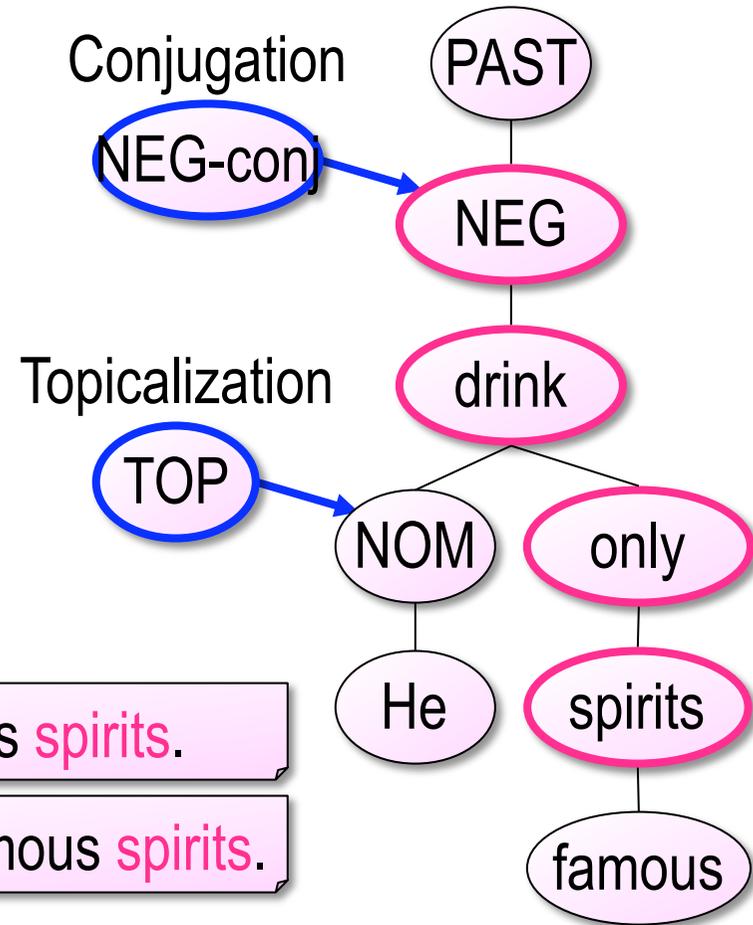
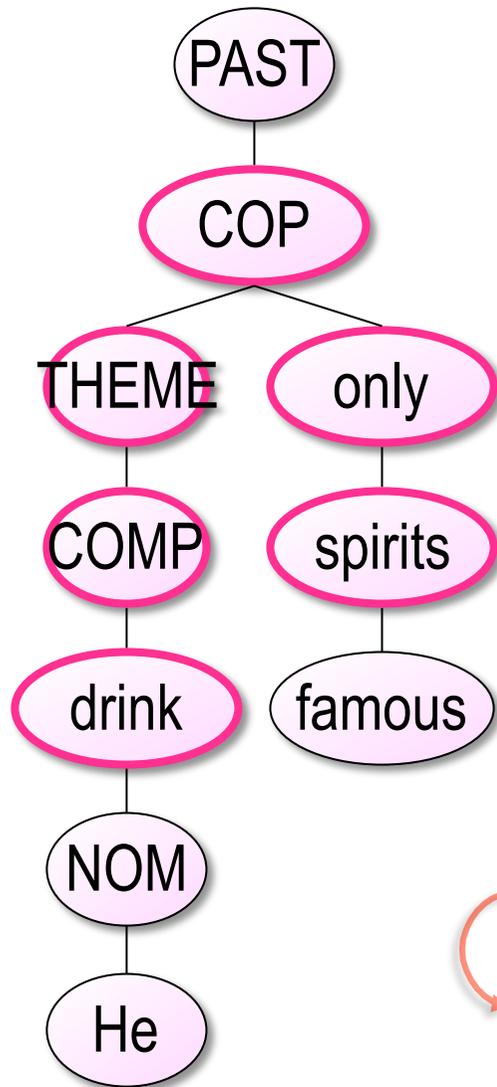
Step 2. Assessment

- Because knowledge is static
 - Grammaticality
 - Semantic appropriateness
 - Equivalency of meanings in the context
- Filtering, correction, ranking
 - Rule-based
 - Statistical approach



Rule-based correction

[Takahashi+, 01]



All he drank were famous spirits.

He drank nothing but famous spirits.

Statistical assessment

- Grammaticality: statistical language model
 - Collocation
 - ◆ e.g., <V, Slot, N> [Fujita+, 04][Pantel+, 07]
 - Global grammaticality of sentences [Wan, 05]

- Semantic appropriateness
 - Compare gloss and context [Okamoto+, 03]

- Equivalency of meanings in the context
 - Suitability for the given context
[Pantel+, 07][Szpektor+, 08]

Issues and current status

- Application of knowledge to a certain context
 - Influence of paraphrase to the context
 - How to deal with generality and idiosyncrasy?

- Two approaches
 - Transfer + assessment
 - Transducer

- Viewpoints of assessment
 - Grammaticality } Discussed
 - Semantic appropriateness
 - Equivalency of meanings in context } Not yet explored

Outline

1. Sameness of meaning
2. Linguistically-motivated typology
3. Paraphrases in apps
4. Computation
- ▶ 5. Future directions

Future directions (technical points of view)

Phase 1. Knowledge development

- How to cover various types of paraphrases?
 - ◆ → Not enough
 - Need a formalism and a resource repository

Phase 2. Use of knowledge

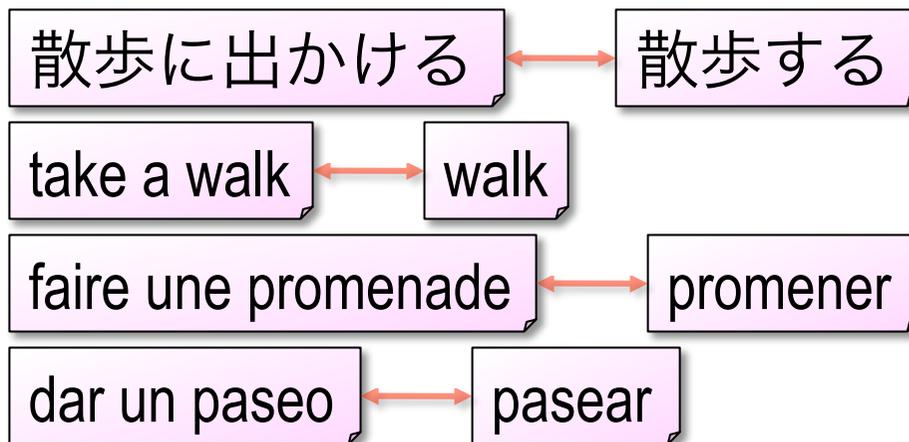
- How to deal with generality and idiosyncrasy?
 - ◆ → Some levels on grammaticality
 - ◆ → More studies on “paraphrase in context”
 - We ask users in generation-type apps

Phase 3. Tuning for apps

- How to selectively use each type of paraphrases?
 - ◆ → No cross-application platform. Modularization!!

Future directions (linguistic points of view)

- ❑ Establishing the way to compile the typology
 - incl. infrastructure: community, portal
- ❑ Parallelism



Thank you

Acknowledgment

My ex-supervisor: Prof. Kentaro INUI

My ex-boss: Prof. Satoshi SATO

<http://paraphrasing.org/~fujita/>