



# **Corpus, Ontology and Annotation: Mapping Natural Language Expressions with Facts**

Jin-Dong Kim  
University of Tokyo



# GENIA Event Annotation - Motivation

- Deeper linguistic analysis lowers language barrier
  - ✓ Predicate-Argument Structure
  - ✓ Semantic Role Labeling

**MEDIE** — [See what causes cancer?](#)

MEDIE is a demo system presented by [Tsuji Laboratory](#)

Semantic Search    **Keyword Search**    GCL Search

subject                          verb                          object

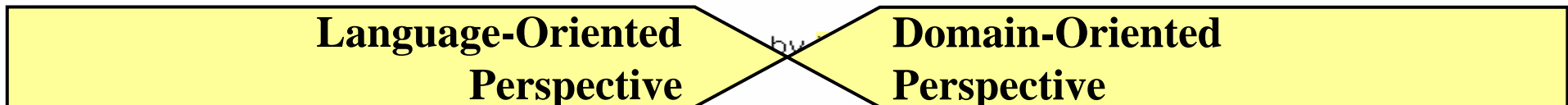
      

       [Help](#)

[»Advanced search](#)

Adenovirus-mediated high dose reexpression of **Peg3 / Pw1** mRNA expression .

One of the mechanisms for **p53** to induce mitochondria-mediated **cell** death events is to activate genes that are directly involved in the initiation of mitochondria-induced apoptosis .





# GENIA event annotation- Motivation

## □ Domain-oriented Information Access

**MEDIE** — [See what causes cancer?](#)

*MEDIE is a demo system presented by [Tsuji Laboratory](#)*

**Semantic Search**   **Keyword Search**   **GCL Search**

**cause**   **event**   **theme**

              [Help](#)

[»Advanced search](#)

- Biological\_process
  - Cellular\_process
    - Cell\_adhesion
    - Cell\_communication
    - Cell\_differentiation
    - Cell\_recognition
    - Cellular\_physiological\_process
  - Physiological\_process
    - Localization
  - Metabolism
    - DNA\_metabolism
      - DNA\_modification
      - DNA\_recombination
      - DNA\_repair



# GENIA Event Annotation - example

Secretion of >TNF<T32, the product of another >>NF-kappa B<T34-dependent gene<T33, was abolished by >BHA<T35 in >PMA-stimulated >U937 cells<T37<T36.

## EVENT E23

TYPE : Localization

THEME : T32

CLUE : >Secretion< >of< TNF, the product of another NF-kappa B-dependent gene, was abolished by BHA in PMA-stimulated U937 cells.

ClueType

LinkTheme

## EVENT E24

TYPE : Negative\_regulation

THEME : E23

CAUSE : T35

CLUE : Secretion of TNF, the product of another NF-kappa B-dependent gene, was >abolished< >by< BHA >in PMA-stimulated U937 cells<.

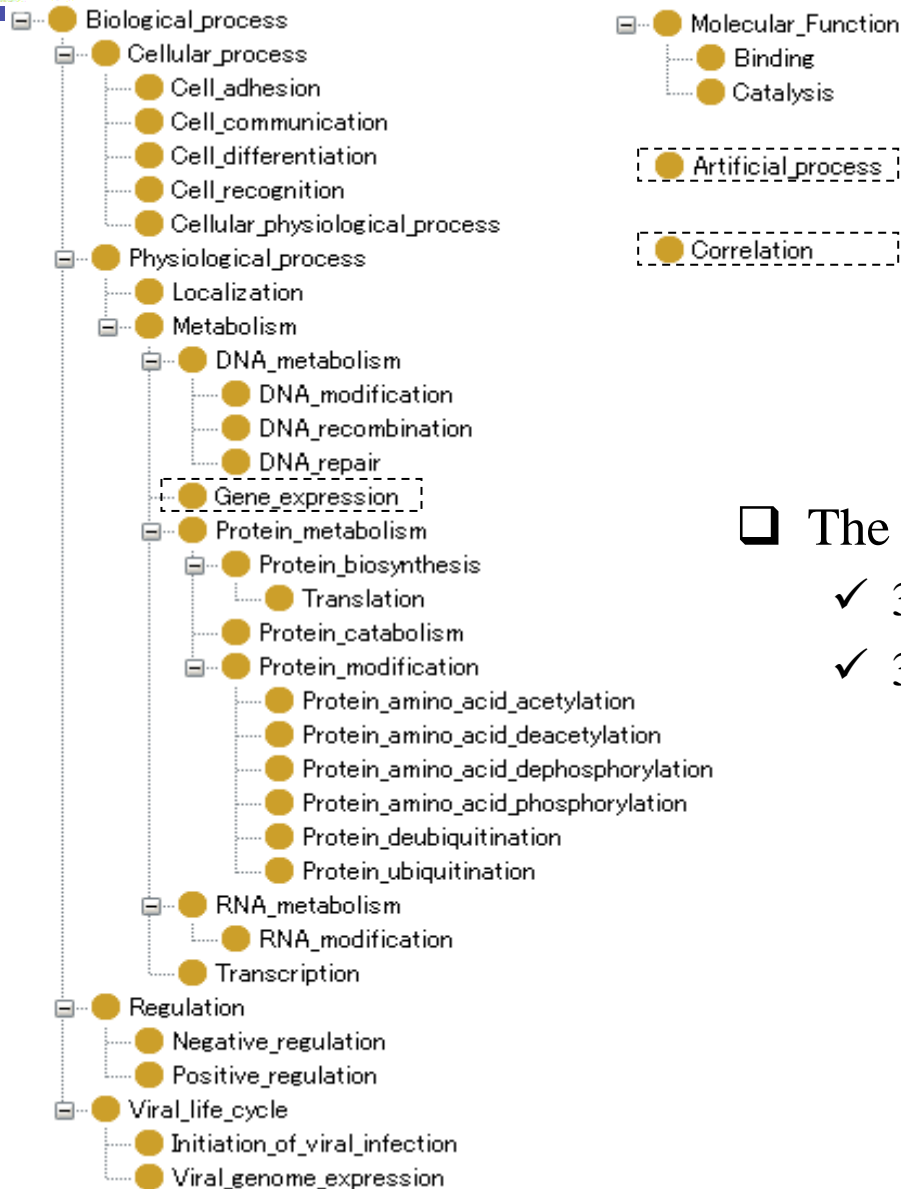
LinkCause

ClueType

- ✓ For an identified event in the given sentence,
  - ➔ classify the **type** of events and record the text span giving the clue of it (ClueType).
  - ➔ identify the **theme** of the events and record the text span linking the theme to the event (LinkTheme).
  - ➔ identify the **cause** of the events and record the text span linking the cause to the event (LinkCause).
  - ➔ record the environment (location, time) of the events (ClueLoc, ClueTime).



# GENIA event ontology



□ The current GENIA event ontology consists of

- ✓ 34 hierarchical concepts taken from GO.
- ✓ 3 newly introduced concepts.

⇒ Correlation

- meaning ‘some’ relation between events.

⇒ Artificial\_process

- Artificially performed processes.
- Transfection, treatment, ...

⇒ Gene\_expression

- Transcription + Translation



# Types of IE Tasks

## □ Recognition of mentions

- ✓ Find expressions (sentences) mentioning certain facts.
- ✓ Do not care the modality or mood of the expressions.
- ✓ Form a basis for further semantic processing

**GENIA Event Annotation**

## □ Recognition of supports

- ✓ Find expressions (sentences) supporting certain facts.
- ✓ A task which is one step beyond the mention recognition task.
  - ➔ Have to recognize not only the expressions mentioning a fact but also expressions supporting the relation

**Mining Disease-Gene Association**  
**[Hong-Woo Chun 2007]**

## □ Recognition of facts

- ✓ Find facts.
- ✓ collect potential evidences and make a decision based on the collection.
- ✓ do not care about the exact location where the facts are mentioned.

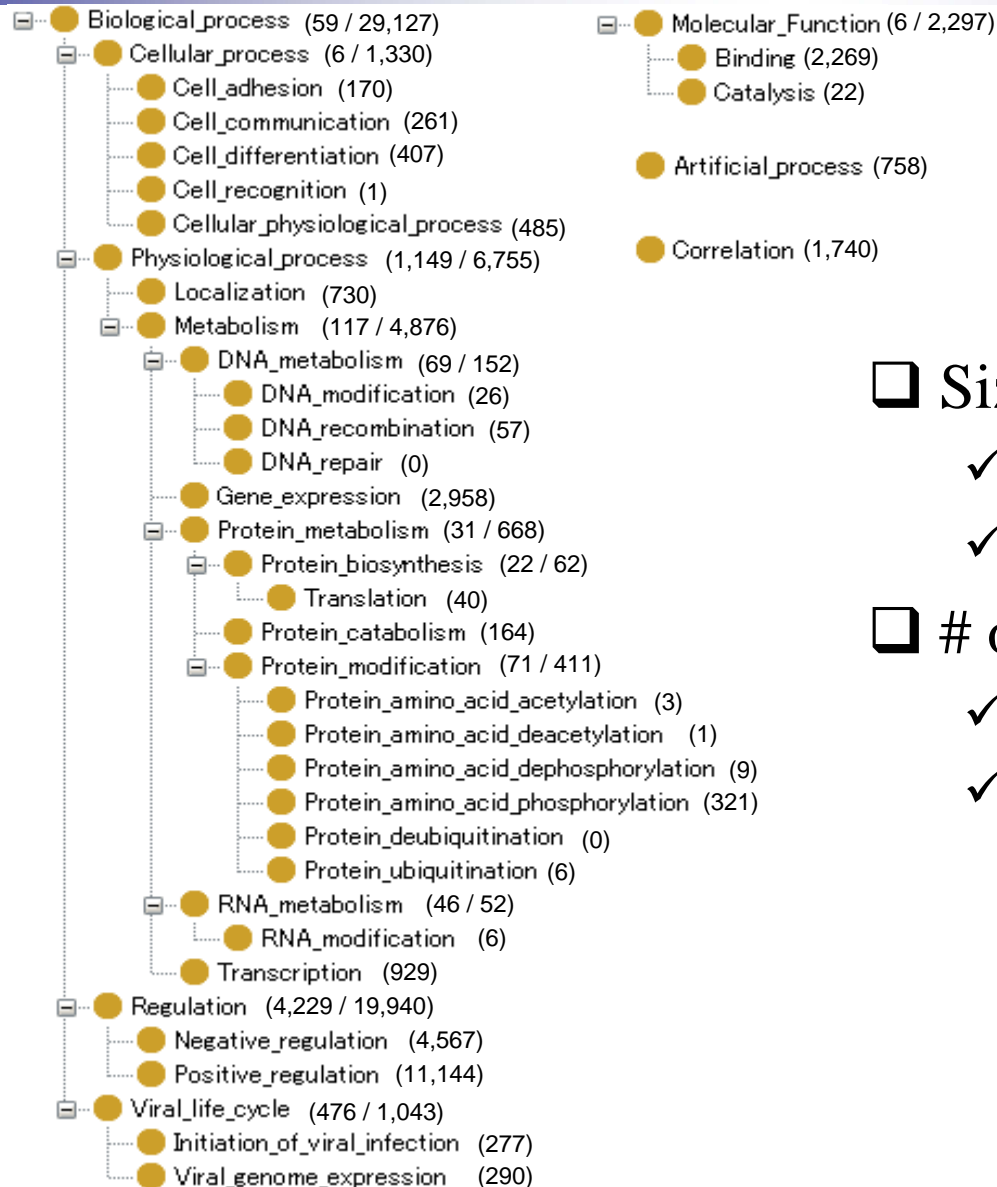
**BioCreative Protein Interaction Pairs Task**  
**[Rune Sætre 2007]**

NLP-oriented task

IE-oriented task



# Statistics



## Size of the corpus

- ✓ 958 abstracts
- ✓ 7,992 sentences

## # of events annotated

- ✓ 29,127 biological processes
- ✓ 2,297 molecular functions

# Theme distribution for Regulation

□ Theme patterns observed more than 3 times (statistics out of 19,940)(958 abstracts)

✓ Regulation EVENT	11,572
✓ Regulation Protein	3,402
✓ Regulation Other_name	2,357
✓ Regulation DNA	1,263
✓ Regulation RNA	308
✓ Regulation Cell_type	127
✓ Regulation Virus	62
✓ Regulation Cell_line	40
✓ Regulation Other_organic_compound	16
✓ Regulation Tissue	15
✓ Regulation Lipid	15
✓ Regulation Inorganic	15
✓ Regulation EVENT EVENT	15
✓ Regulation Body_part	11
✓ Regulation Peptide	9
✓ Regulation Nucleotide	9
✓ Regulation CONS	8
✓ Regulation Atom	8
✓ Regulation Amino_acid_monomer	7
✓ Regulation Multi_cell	6
✓ Regulation Protein Protein	5
✓ Regulation Cell_component	5
✓ Regulation Other_artificial_source	4
✓ Regulation DNA DNA	4

**Regulation of an event**

**Regulation of a function of a gene(product)**

**Regulation of a function of a virus**

**Regulation of a cellular process of a cell**

**Regulation of the amount of a cell**

**16,774/19,940  
= 84%**



# Linguistic manifestation – Regulation

## □ Language patterns

- |                       |                                      |
|-----------------------|--------------------------------------|
| ✓ 195 effects [on]    | ✓ 35 sensitive                       |
| ✓ 164 regulation [of] | ✓ 35 modulating                      |
| ✓ 157 dependent       | ✓ 35 changes [in]                    |
| ✓ 156 regulated       | ✓ 33 role                            |
| ✓ 131 involved [in]   | ✓ 30 affected                        |
| ✓ 126 role [in]       | ✓ 28 responsive                      |
| ✓ 125 effect [on]     | ✓ 27 important [for]                 |
| ✓ 97 regulate         | ✓ 26 transcriptional regulation [of] |
| ✓ 77 not affect       | ✓ 25 play * role [in]                |
| ✓ 65 regulates        | ✓ 25 affect                          |
| ✓ 61                  | ✓ 24 involvement [in]                |
| ✓ 60 regulating       | ✓ 24 are key regulators [of]         |
| ✓ 52 not affected     | ✓ 22 no effect [on]                  |
| ✓ 51 Regulation [of]  | ✓ 21 plays * role [in]               |
| ✓ 48 regulation       | ✓ 19 modulation [of]                 |
| ✓ 44 response         | ✓ 19 influence                       |
| ✓ 44 independent      | ✓ 18 responsiveness                  |
| ✓ 41 controlled       | ✓ 18 modulate                        |
| ✓ 39 control          | ✓ 18 controlling                     |
| ✓ 36 control [of]     | ✓ ...                                |

## ☐ Inter-annotator Agreement

### ✓ Level 1 (strict match)

- ➔ Any two annotations are identical if
  - Their event types are the same.
  - Their clue expressions are overlapped.
  - Their themes are the same.



**56%**

### ✓ Level 2 (soft match)

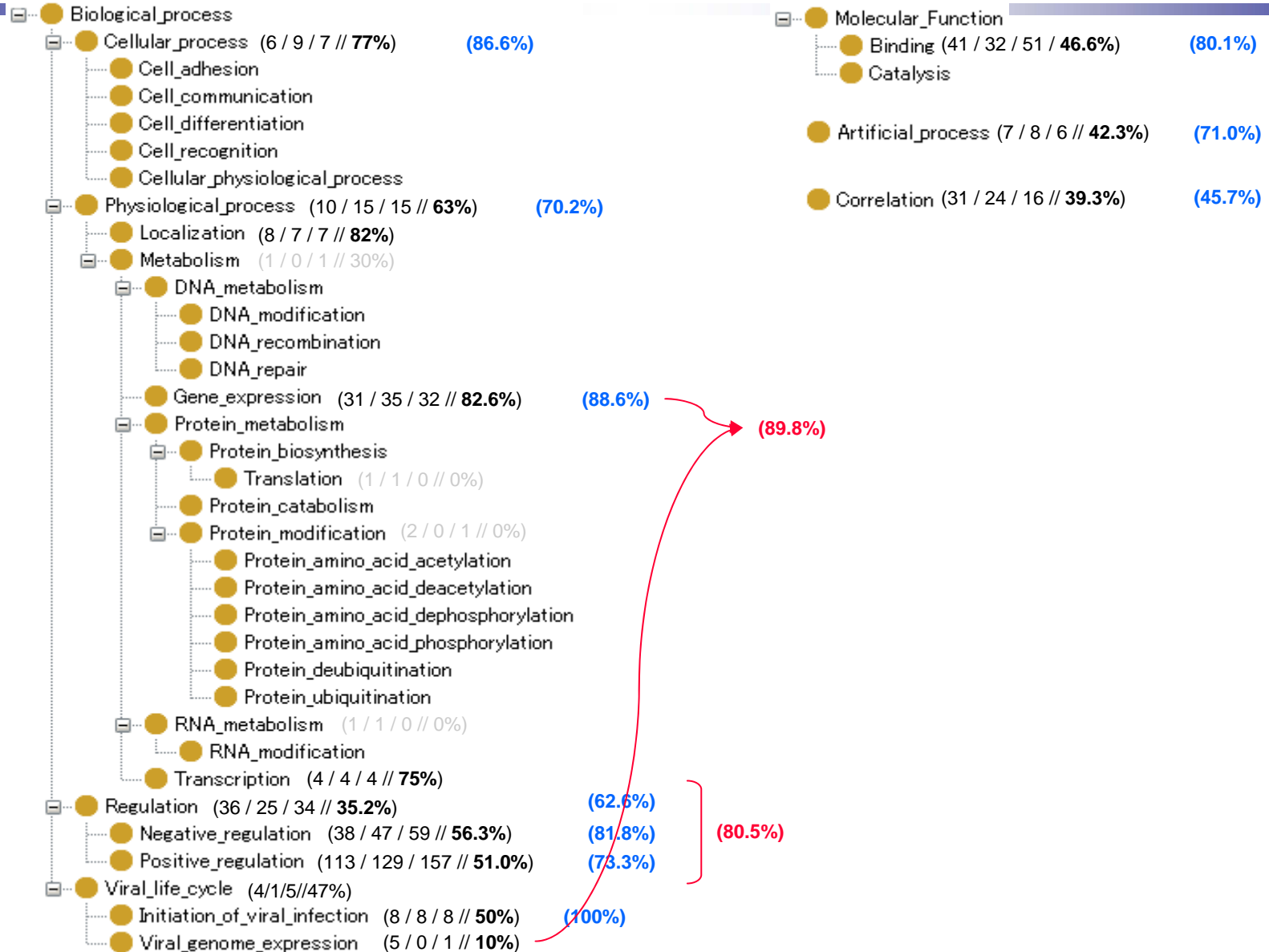
- ➔ Any two annotations are identical if
  - Their event types are the same
  - Their clue expressions are overlapped
  - They share at least one theme.  
(allows incomplete theme list)



**77%**



# Inter-Annotator Agreement - Detail



# Analysis

- ❑ Viral\_life\_cycle (4 / 1 / 5 // 47%)
  - ✓ Initiation\_of\_viral\_infection (8 / 8 / 8 // 50%)
    - ➡ Incomplete theme list → 100% by soft match

Thus, reduced levels of PKC-induced nuclear NF-κB activity in two T cell subclones did not affect their normal cell growth, but correlated with a pronounced reduction in their susceptibility to HIV-1 infection.

EVENT E39 (assertion: exist, uncertainty: certain)  
 TYPE : Initiation\_of\_viral\_infection  
 THEME : A16  
 THEME : T62  
 CLUE : Thus, reduced levels of PKC-induced nuclear NF-κB activity in two T cell subclones did not affect their normal cell growth, but correlated with a pronounced reduction in their susceptibility to HIV-1 infection.

- ✓ Viral\_genome\_expression (5 / 0 / 1 // 10%)
  - ➡ Confusion with Gene\_expression and Transcription
  - ➡ Problem of ontology
    - Overlapped definitions

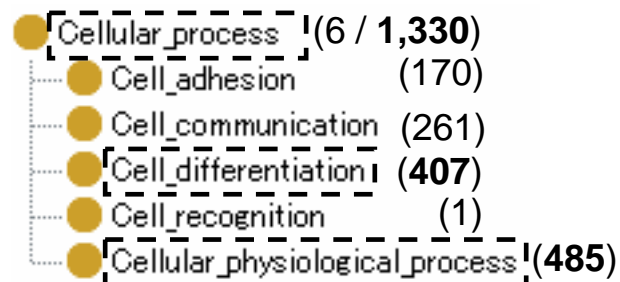


# Use Case - Event Sentence Retrieval

## ❑ Problem definition

✓ Sentence classification problem

➔ Does a given sentence mention a specific type of biomedical event? (YES/NO)



❑ Training/Evaluation Samples were extracted from GENIA event annotated corpus

✓ 7,992 sentences



# Feature Representation (1/3)

- Bag-of-Words
  - ✓ Porter Stemmer
  - ✓ Stopword list



# Feature Representation (2/3)

## □ Cluewords

- ✓ Extracted from clue expressions (multi-word expressions)
- ✓ Purpose
  - ➔ To raise the coverage.
- ✓ Process
  - ➔ Collecting words in clue expressions
  - ➔ Get the stem of each words
  - ➔ Get the count of each stems
  - ➔ stems with count 1 are filtered out
  - ➔ Get a ordered-list of stems by precision.

# Clue Expressions

- Clue expressions:
  - ✓ expressions giving a clue for the type of event.
  - ✓ Have been marked by human annotators.

The effects of ▶prostaglandin E2◀T6 (▶PGE2◀T7) on ▶▶ cytokine◀T9 production◀T8 and ▶proliferation◀T10 of the ▶CD4+ human helper T cell clone SP-B21◀T11 were investigated.

## EVENT E7

TYPE : Cellular\_physiological\_process

THEME : T11

CLUE : ▶The effects of prostaglandin E2 (PGE2) on cytokine production and ◀  
▶proliferation◀ ◀◀of◀ the CD4+ human helper T cell clone SP-B21 were investigated.◀

clue expression for  
the type of event



## Clue words

### Cell\_differen.(9)

- ✓ differenti
- ✓ fate
- ✓ matur
- ✓ hemoglobin
- ✓ granulocyt
- ✓ commit
- ✓ develop
- ✓ phenotyp
- ✓ growth

### Cell\_phy\_proc(39)

- ✓ diapedes
- ✓ transmigr
- ✓ renew
- ✓ roll
- ✓ nondivid
- ✓ divis
- ✓ outgrowth
- ✓ cytolyt
- ✓ expans
- ✓ prolif
- ✓ mitogenesi
- ✓ lysi
- ✓ syncytia
- ✓ ...

### Cell\_proc (76)

- ✓ transmit
- ✓ diapedes
- ✓ transmigr
- ✓ renew
- ✓ roll
- ✓ paracrin
- ✓ nondivid
- ✓ divis
- ✓ cytolyt
- ✓ outgrowth
- ✓ prolifer
- ✓ prolif
- ✓ lysi
- ✓ ...

## □ Syntactic Patterns

✓ BNP[clueword ... theme...]

✓ BNP[clueword ... ]  $\xleftarrow{\text{ARG}}$  in/of  $\xrightarrow{\text{ARG}}$  [... theme...]

✓ VP[clueword ... ]  $\xrightarrow{\text{ARG}}$  NP[... theme...]

□ Bag-of-words from selected BNPs according to the syntactic patterns



# Performances

Event type	Bag-of-words	+Clue words	+Syn. Structures
Cellular_process	51.7/62.2/56.5	56.9/67.9/61.9	59.4/70.5/64.5
Cell_differentiation	63.1/68.9/65.8	70.8/71.6/71.2	82.7/74.2/78.2
Cellular_phy._process	41.5/53.3/46.7	55.0/71.0/61.4	61.5/65.2/63.4

Event type	Classification with preference toward high recall	+Syn. Structures
Cellular_process		91.0/40.0/55.5
Cell_differentiation		92.2/50.8/65.5
Cellular_phy._process		86.1/44.6/58.7



# Conclusion

- GENIA event annotated corpus
  - ✓ The first stage will be finished in this month
    - ➔ 1,500 abstracts
  - ✓ There are much room for improvement.
  - ✓ Has not yet reached at a gold-standard level.
  - ✓ But, is already a rich source of information for language processing system for biomedical domain.