



The GENiA corpus – Linguistic and Semantic Annotation of Biomedical Literature

Jin-Dong Kim
Tsujii Laboratory,
University of Tokyo



Contents

- ❑ Ontology, Corpus and Annotation for IE
- ❑ Annotation and Information Extraction
- ❑ Linguistic and Semantic Annotation
- ❑ The GENIA corpus
- ❑ The GENIA annotation
- ❑ XConc Suite – an integrated corpus annotation environment
- ❑ GENIA event annotation
- ❑ GENIA corpus – current status
- ❑ Conclusions

□ Ontology

✓ In Philosophy

➔ a particular theory about the nature of being or the kinds of existents

✓ In Computer Science

➔ a data model that represents a domain and is used to reason about the objects in that domain and the relations between them.

- a controlled vocabulary of terms
- a body of knowledge
- ...

□ Semantic Annotation

✓ A mapping between knowledge pieces written in ontology language and natural language.

□ Corpus

✓ From a linguistic perspective

➔ a collection of texts that are representative of a language

✓ From a perspective of IE

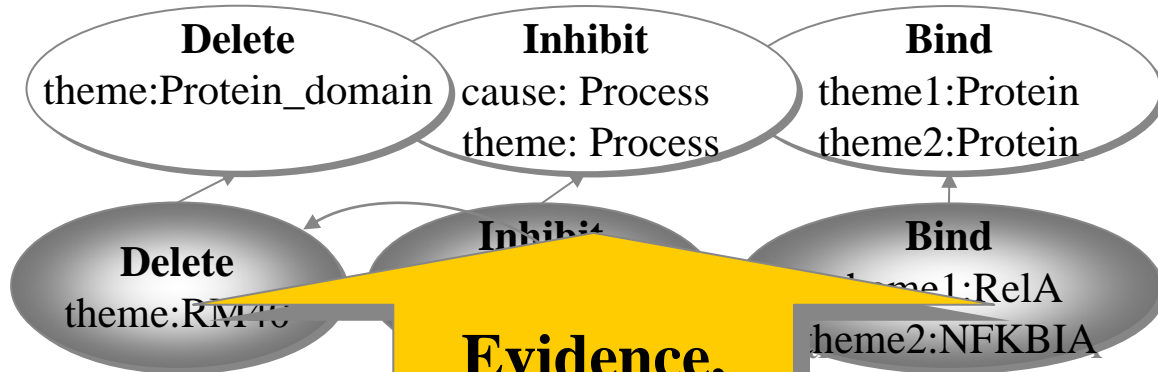
➔ a sample of knowledge source written in natural language



Ontology Corpus and Annotation for IE (2/2)

ONTOLOGY

contain
domain: Protein|Protein_domain
region: Protein_domain



Domain Knowledge, Vocabulary

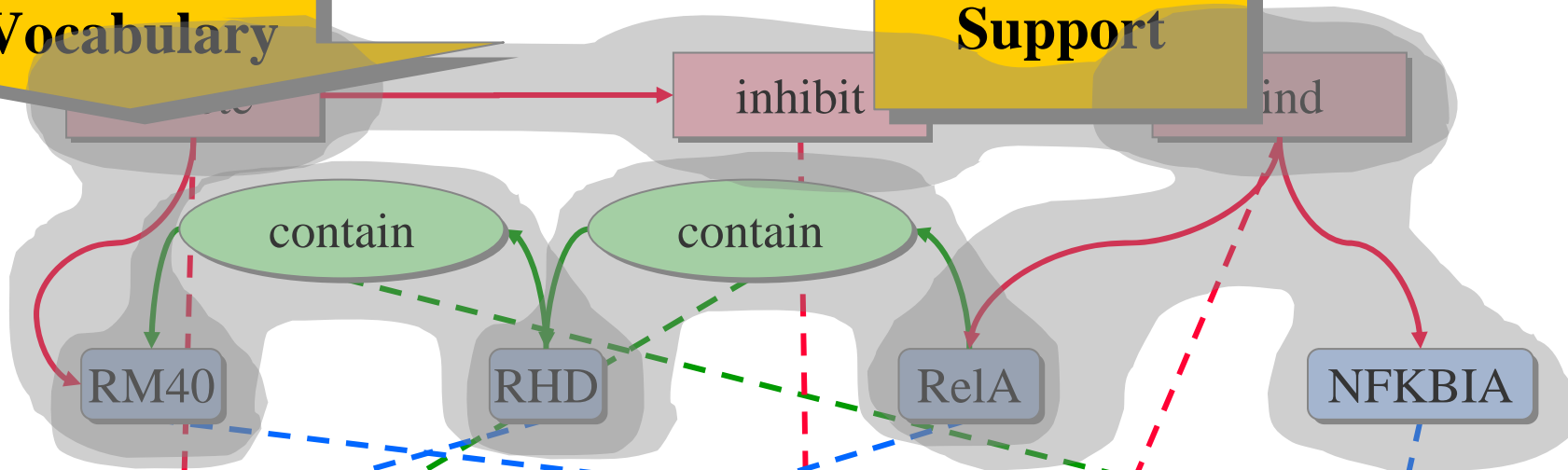
Evidence, Support

ANNOTATION

events

relations

elements

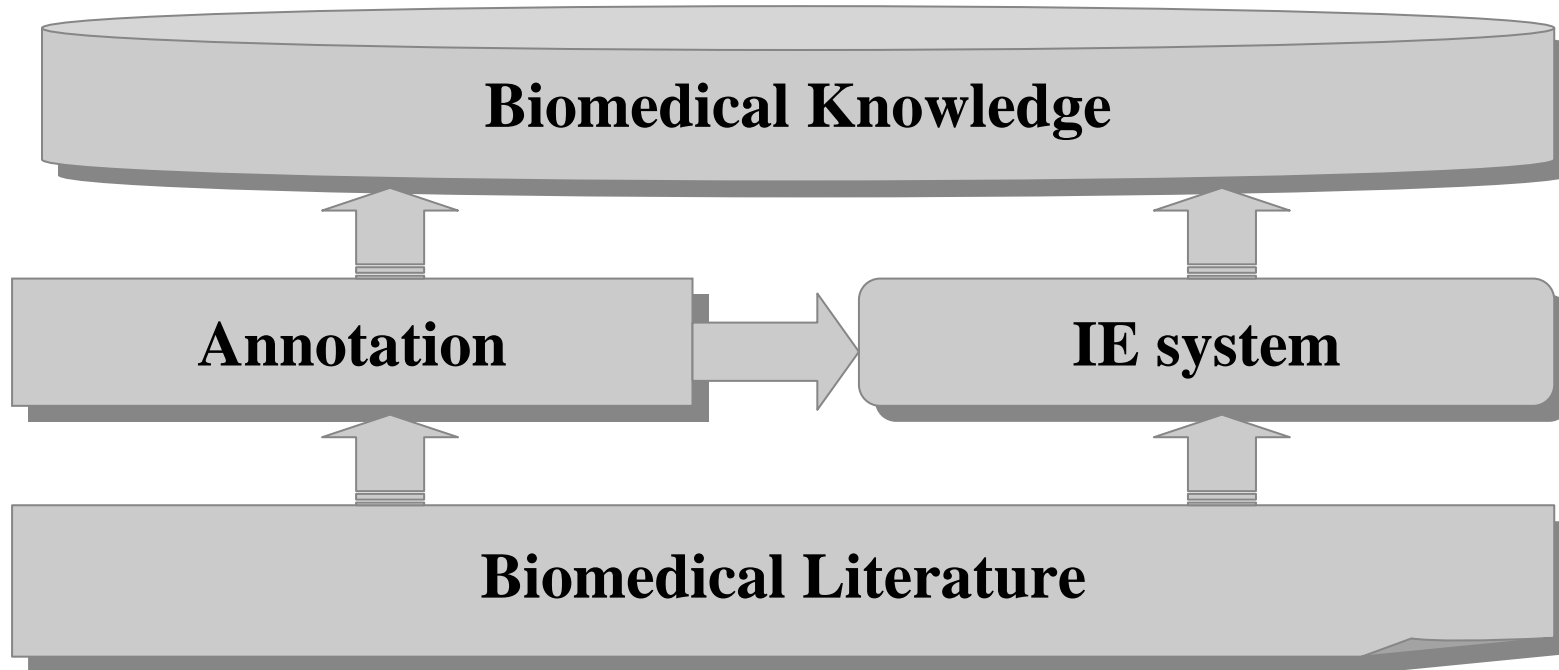


TEXT

... 3) selective [deletion of] the functional [nuclear localization signal] [present in] the [Rel homology domain] [of] [NF-kappa B p65] [disrupts its ability] [to engage] [I kappa B/MAD-3], and 4) ...

PMID:1493333

Annotation & Information Extraction



- Semantic annotation simulates an ideal performance of IE system.
 - ✓ IE systems can be developed by referencing annotated corpus.
 - ✓ The performance of IE systems can be evaluated by being compared to the annotated corpus.



The GENIA corpus

WSJ corpus

- ✓ Journalistic newswire style of writing on financial domain.

MEDLINE corpus

- ✓ Scientific writing covering the general or a specific domain of life sciences.

GENIA corpus

- ✓ Is a subset of MEDLINE.
- ✓ Covers the specific subject domain of “biological reactions concerning transcription factors in human blood cells”.
 - ⇒ Search Query: “Human”[MeSH] and “Blood Cells”[MeSH] and “Transcription Factors”[MeSH]
- ✓ Has been being annotated with linguistic and semantic information.



The GENIA annotation

□ Linguistic annotation

✓ Reveals linguistic structures behind the text

⇒ Part-of-speech annotation

– annotates for the syntactic category of each word.

⇒ Syntactic Tree annotation

– annotates for the syntactic structure of sentences.

**Follow PTB II guidelines
with a small modifications**

□ Semantic annotation

✓ Reveals knowledge pieces delivered by the text.

⇒ Term annotation

– annotates domain-specific terms

⇒ Event annotation

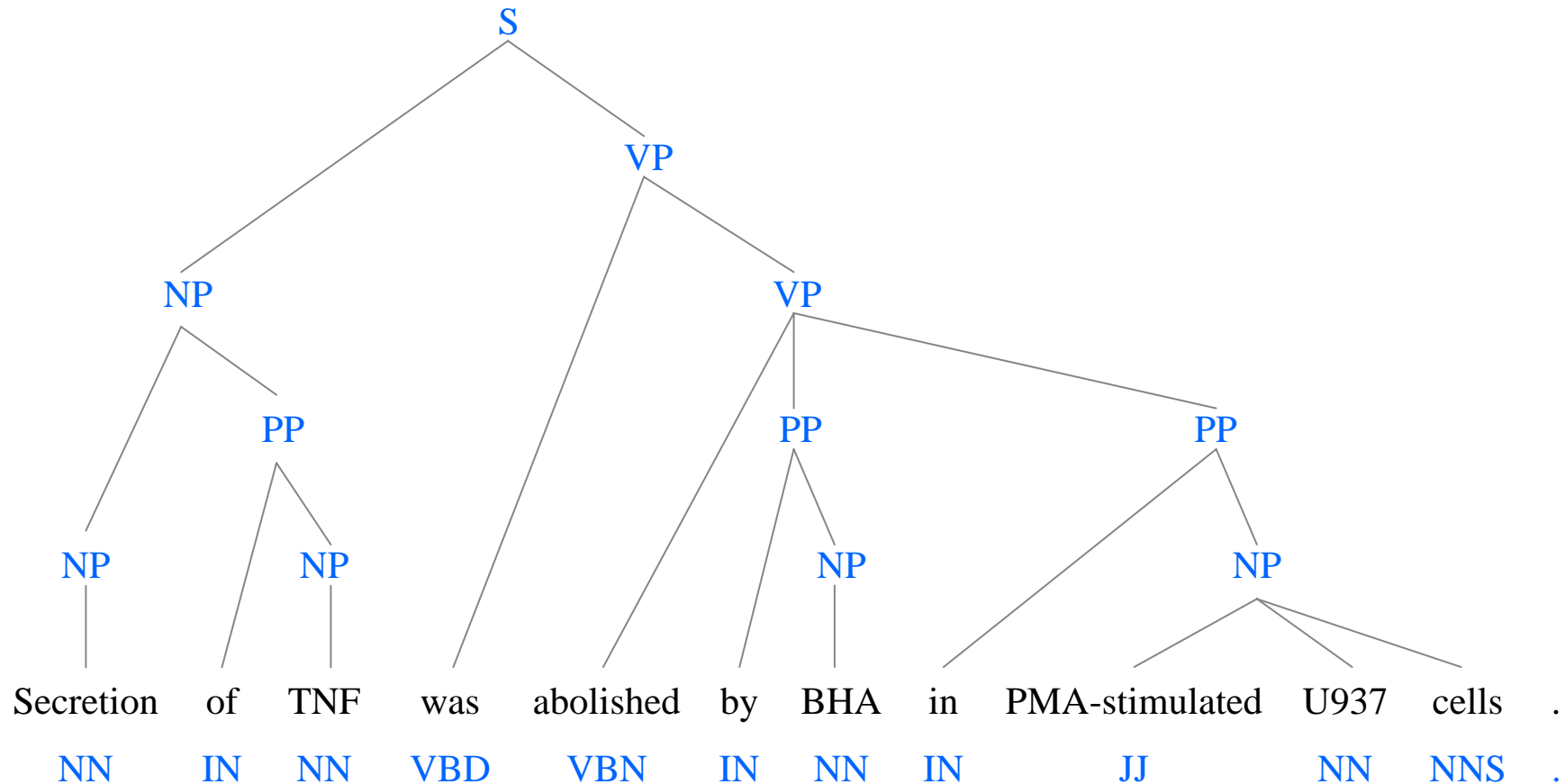
– annotates events on biological entities.

**Ontology-driven
annotation**



GENIA linguistic annotation - example

Syntactic Tree annotation



Part-of-speech annotation



GENIA term annotation - example

Protein_molecule

Other_organic_compound

Cell_line

Secretion of **TNF** was abolished by **BHA** in PMA-stimulated **U937** cells .



GENIA term annotation - Application

- ❑ Training & evaluation of bio-entity recognition systems
 - ✓ Shared task of bio-entity recognition at Colin 2004 JNLPBA workshop
 - ➡ Participating systems were trained with a part of GENIA corpus, and
 - ➡ the performance was evaluated against the other part of the corpus.

Bio-entity recognition system	Recall	Precision	F-score
SVM+HMM (Zhou et al., 2004)	76.0	69.4	72.6
Semi-Markov CRFs (in prep.)	72.7	70.4	71.5
Two-Phase (Kim et al., 2005)	72.8	69.7	71.2
Sliding Window (in prep.)	71.5	70.2	70.8
CRF (Settles, 2005)	72.0	69.1	70.5
MEMM (Finkel et al, 2004)	71.6	68.6	70.1
:	:	:	:

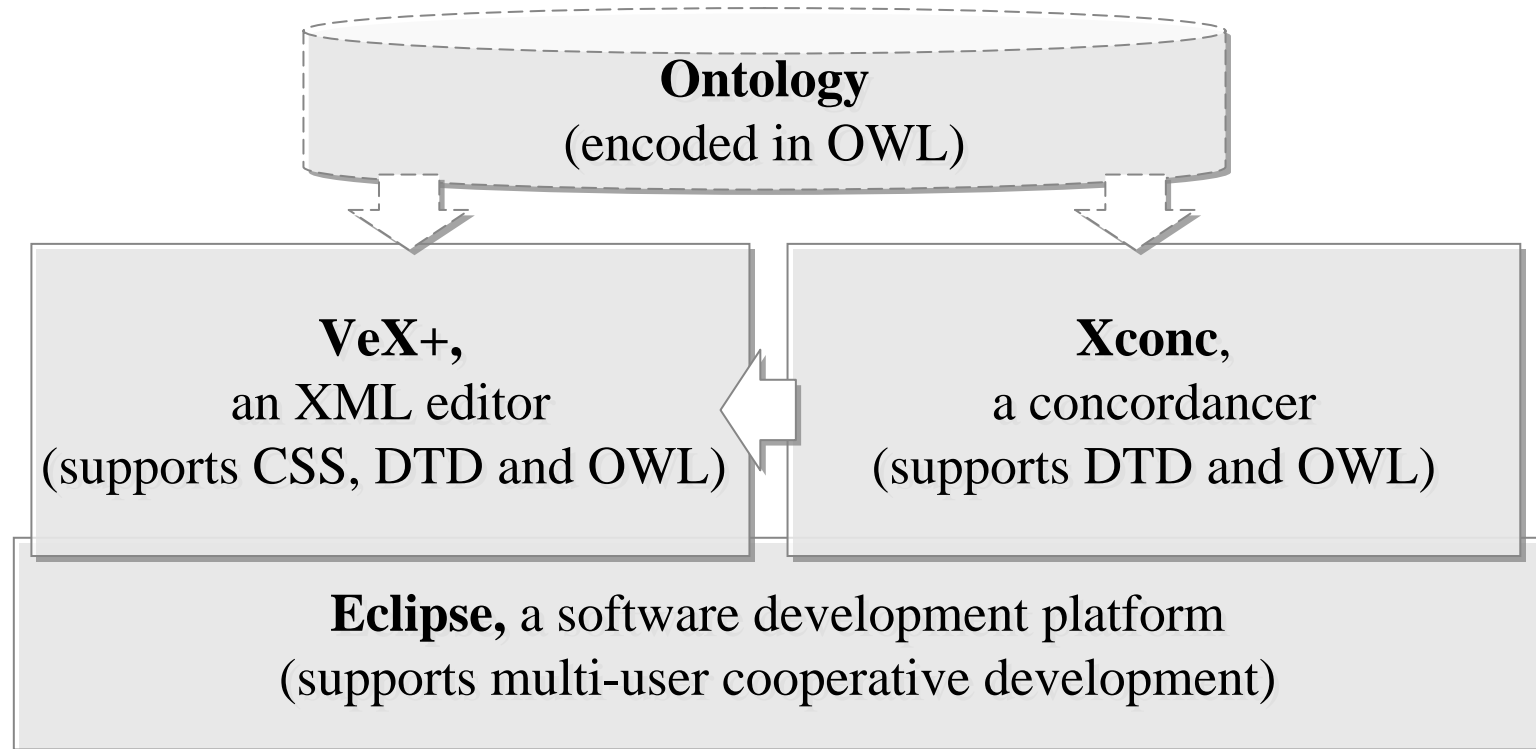


XConc Suite – An integrated annotation environment

- ❑ What can we do with it
 - ✓ Editing corpus files
 - ➔ Full-featured XML editor
 - ✓ Searching patterns
 - ➔ Regular expression + XML element selection
 - ✓ Version control, history management
 - ➔ CVS support
- ❑ Organization
 - ✓ VeX+
 - ➔ A general purpose XML editor
 - ✓ XConc
 - ➔ An XML-based concordancer
 - ✓ Eclipse
 - ➔ A general purpose software development platform



XConc Organization





How does it look

The screenshot shows the XConc Annotation software interface with several red annotations pointing to specific features:

- Navigation View +**: Points to the Navigator pane on the left.
- XML Editor +**: Points to the main text area displaying XML content.
- Property View/Editor +**: Points to the Properties pane on the right.
- Query Editor +**: Points to the XConc Search pane on the right.
- Ontology View +**: Points to the Ontology Tree pane on the left.
- Concordance View +**: Points to the XConc Search Result table at the bottom.
- Element Selector Editor +**: Points to the XConc Selector Editor pane at the bottom right.

The XML Editor displays the following content:

```
1653056.xml
1653056.xml
##PMID1653056
>
##TITLE
>
NF-kappa B activation by tumor necrosis factor alpha in the Jurkat T cell line is independent of protein kinase A, protein kinase C, and Ca(2+)-regulated kinases.
>
##ABSTRACT
>
NF-kappa B is a DNA-binding regulatory factor able to control transcription of a number of genes, including human immunodeficiency virus (HIV) genes.
>
In T cells, NF-kappa B is activated upon cellular treatment by phorbol esters and the cytokine tumor necrosis factor alpha (TNF alpha).
```

The XConc Search Result table shows the following data:

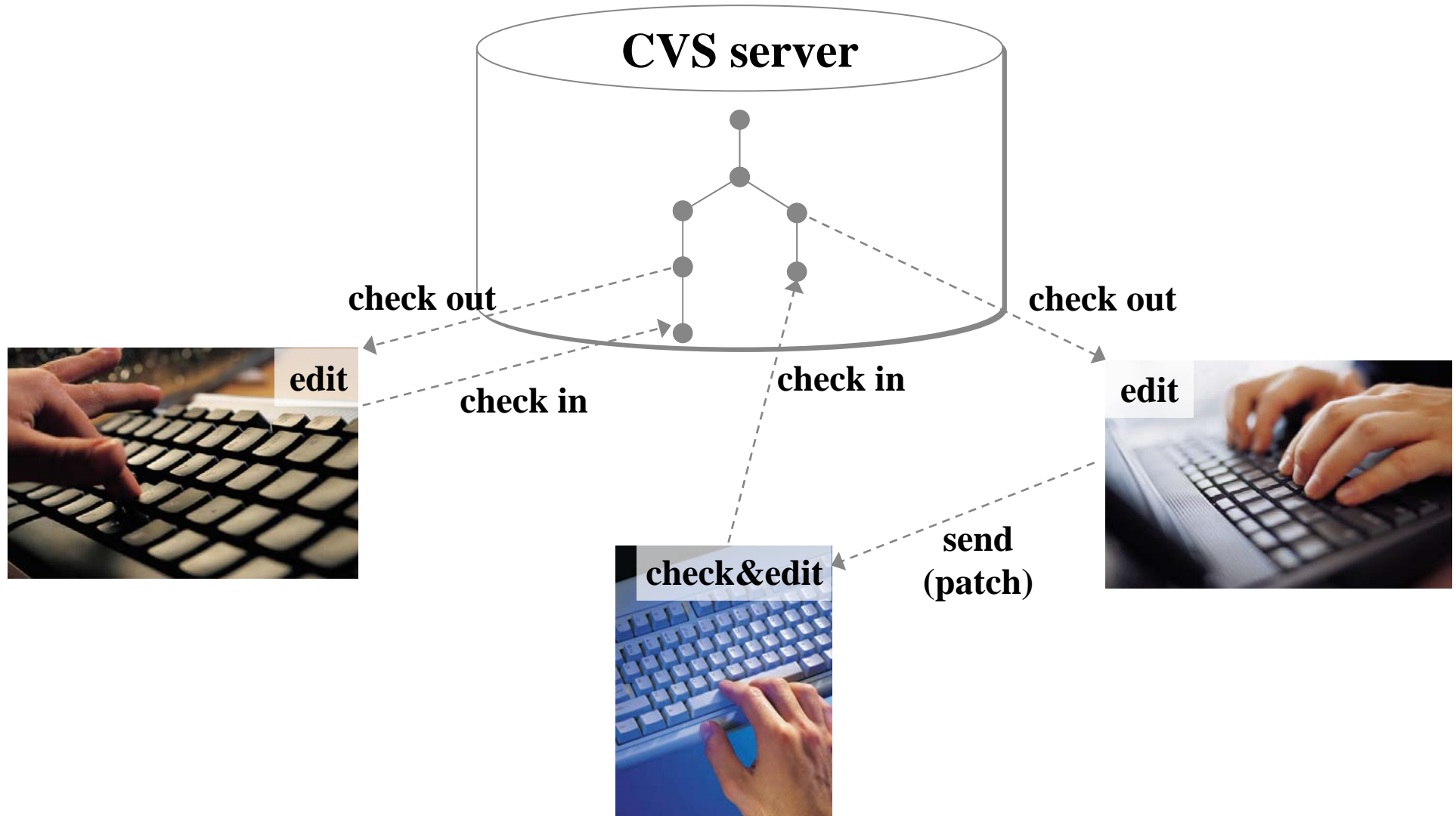
No	Ele...	Attributes	Left	Center	Right	C.	FileName
21	term	sem=Prote...	In T cells,	NF-kappa B	is activated upon ce...	<input type="checkbox"/>	D:\XConc\WGENI...
22	term	sem=Prote...	...ar events leading to	NF-kappa B	activation by TNF al...	<input type="checkbox"/>	D:\XConc\WGENI...
23	term	sem=Prote...	...by Ca2+ influx wh...	TNF alpha	activation was not.	<input type="checkbox"/>	D:\XConc\WGENI...
24	term	sem=Prote...	Thus, TNF alpha-ind...	NF-kappa B	activation was four...	<input type="checkbox"/>	D:\XConc\WGENI...
25	term	sem=Prote...	...fication facilitated	NF-kappa B	activation by both T...	<input type="checkbox"/>	D:\XConc\WGENI...
26	term	sem=Prote...	...e have detected t...	crossreactin...	in activated normal ...	<input type="checkbox"/>	D:\XConc\WGENI...
27	term	sem=Prote...		Anti-CD2 re...	activate the HIV ion...	<input type="checkbox"/>	D:\XConc\WGENI...
28	term	sem=Prote...	The	c-erbA	-dependent activatio...	<input type="checkbox"/>	D:\XConc\WGENI...
29	term	sem=Prote...	...a component of the	AP-1 transc...	, activates transcript...	<input type="checkbox"/>	D:\XConc\WGENI...

The XConc Selector Editor shows the following configuration:

TARGET: term
ELEMENT: term
ATTRIBUTE: @sem
VALUE: Protein|Pr
CONNECTIVITY: &

term(@sem="Protein|Protein_Domain_or_region|Protein_Substructure)|Protein

CVS supporting





XConc Suite - Demo

- Term annotation demo
- Concordancing demo



Annotation with XConc - screenshot

□ Part-of-speech annotation

```
##PMID1431113
##TITLE
>Redox<NN>status<NN>of<IN>cells<NNS>influences<VBZ
>constitutive<JJ>or<CC>induced<VBN>NF-kappa<NN>B<NN
>translocation<NN>and<CC>HIV<NN>long<JJ>terminal<JJ
>repeat<NN>activity<NN>in<IN>human<JJ>T<NN>and<CC
>monocytic<JJ>cell<NN>lines<NNS>.<PERIOD
##ABSTRACT
>We<PRP>have<VBP>tested<VBN>the<DT>hypothesis<NN>that<IN
>cellular<JJ>activation<NN>events<NNS>occurring<VBG>in<IN
>T<NN>lymphocytes<NNS>and<CC>monocytes<NNS>and<CC
>mediated<VBN>through<IN>translocation<NN>of<IN>the<DT
>transcription<NN>factor<NN>NF-kappa<NN>B<NN>are<VBP
>dependent<JJ>upon<IN>the<DT>constitutive<JJ>redox<NN
>status<NN>of<IN>these<DT>cells<NNS>.<PERIOD
```



Annotation with XConc - screenshot

□ Term annotation

```
1431113.xml
<<<
##PMID1431113
<
##TITLE
<Redox status of cells influences constitutive or
induced NF-kappa B translocation and HIV long
terminal repeat activity in human T and monocytic
cell lines.
<
##ABSTRACT
We have tested the hypothesis that cellular activation
events occurring in T lymphocytes and monocytes and
mediated through translocation of the transcription
factor NF-kappa B are dependent upon the constitutive
redox status of these cells.

We used phenolic, lipid-soluble, chain-breaking
antioxidants (butylated hydroxyanisole (BHA
```

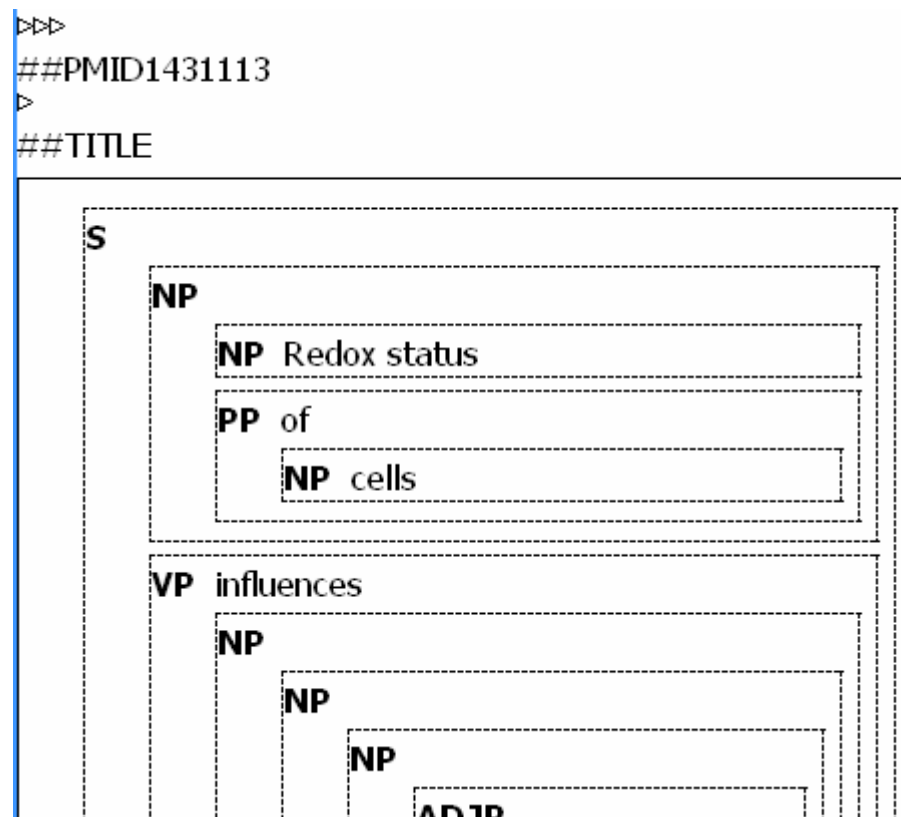


Annotation with XConc - screenshot

□ Part-of-speech & Term annotation

```
>>>
##PMID1431113
>
##TITLE
>>Redox<NN>status<NN>of<IN>cells<NNS>influences<VBZ>
>constitutive<JJ>or<CC>induced<VBN>NF-kappa<NN>B<NN>
>translocation<NN>and<CC>HIV<NN>long<JJ>terminal<JJ>
>repeat<NN>activity<NN>in<IN>human<JJ>T<AA>and<CC>
>monocytic<JJ>cell<AA>lines<AMS>.PERIOD
>
##ABSTRACT
>>We<PRP>have<VBP>tested<VBN>the<DT>hypothesis<NN>that<IN>
>>cellular<JJ>activation<NN>events<NNS>occurring<VBG>in<IN>
>>T<NN>lymphocytes<NNS>and<CC>monocytes<NNS>and<CC>
>>mediated<VBN>through<IN>translocation<NN>of<IN>the<DT>
>>transcription<NN>factor<NN>NF-kappa<NN>B<NN>are<VBP>
>>dependent<JJ>upon<IN>the<DT>constitutive<JJ>redox<NN>
>>status<NN>of<IN>these<DT>cells<NNS>.PERIOD
```


□ Syntactic tree annotation – alternative



- ✓ The editing screen can be easily customized by writing CSS stylesheets and DTD document type definition.

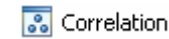
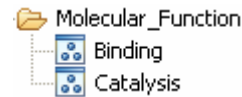
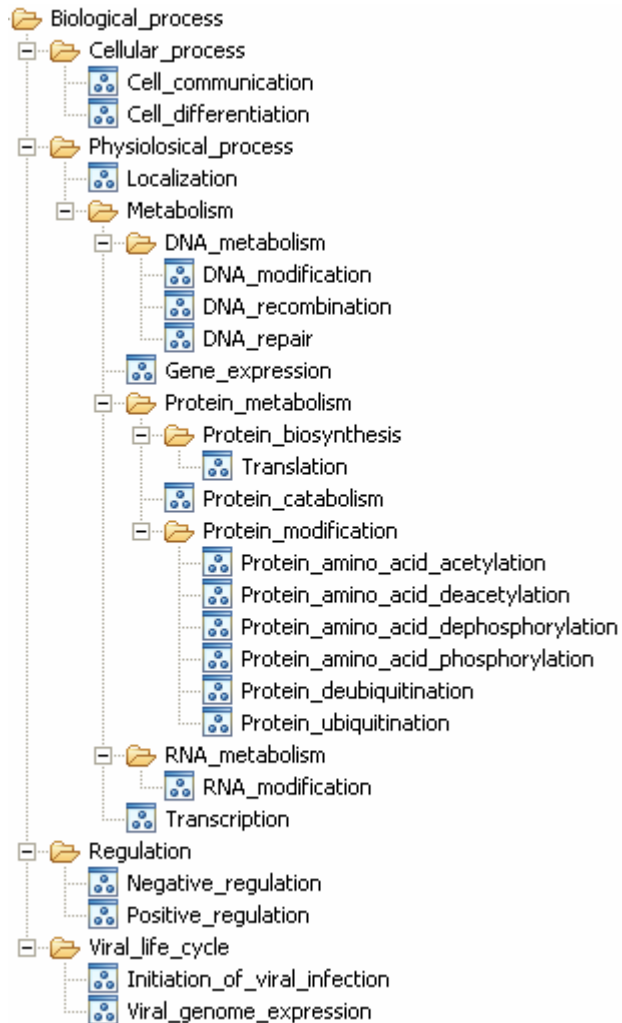


GENIA event annotation

- ❑ A biological event
 - ✓ is defined as a temporal occurrence
 - ✓ that happens to one or more biological entities.
- ❑ The GENIA event ontology
 - ✓ defines a number of biological events which cause some specific change on genes or gene products (proteins)
- ❑ 1st step of GENIA event annotation
 - ✓ will cover the NF κ B pathway
 - ➔ From the GENIA corpus, articles with the MeSH term, NF κ B, have been selected (571 abstracts).
 - ➔ From the Gene ontology, concepts required for describing NF κ B pathway have been selected (34 terms).
- ❑ Next step
 - ✓ will cover other pathways making the approach get generalized.



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



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.

ClueLoc

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 annotation - Policy

Ex.3	... I kappa B/MAD-3 <u>binds</u> directly <u>to</u> NF-kappa B p50 ... (PMID1493333)
	Event #5: Binding (GO:0005488) oTheme : <i>I kappa B/MAD-3</i> and <i>NF-kappa B p50</i>
Ex.4	<u>Expression of M10</u> <u>did not affect</u> <u>induction of HIV</u> transcription ... (PMID1402661)
	Event #6: Gene_expression (-) oTheme : <i>M10</i> Event #7: Transcription (GO:0006350) oTheme : <i>HIV</i> Event #8: (not) Positive_regulation (GO:0048518) oTheme : Event #7 oCause : Event #6

- Most event classes involve one theme,
 - ✓ but some relate to more than one themes (Event #5)
- Usually, events act upon entities, making entities themes of the events,
 - ✓ but some events act upon other events (Event #8)
- Some events are connected with causes.(Event #8)
- Usually, entities cause events,
 - ✓ But sometimes an event may be caused by another event (Event #8)
- An event may be negated (Event #8)

Event annotation – Difficulties (1/3)

Ex.1	Lipopolysaccharide <i>induces</i> phosphorylation of MAD3 ... (PMID8505309)
	Event #2: Positive_regulation (GO:0048518) oTheme : Event #1 oCause : <i>Lipopolysaccharide</i>
Ex.2	<i>Enhancement of</i> human immunodeficiency virus 1 replication in monocytes <i>by</i> 1,25-dihydroxycholecalciferol. (PMID1650477)
	Event #4: Positive_regulation (GO:0048518) oTheme : Event #3 oCause : <i>1,25- dihydroxycholecalciferol</i>
Ex.7	... HS-40 <i>behaved as an authentic enhancer for high-level</i> zeta 2 globin promoter <i>activity</i> ... (PMID8455611)
	Event #15: Positive_regulation (GO:0048518) oTheme : <i>zeta 2 globin promoter</i> oCause : <i>HS-40</i>

❑ Variety of natural language expression

✓ different expressions for the same event class.

➡ Positive_regulation → **X** *induces* **Y**, *enhancement of Y by X*, **X** *behaved as an authentic enhancer for Y*

✓ Hard to determine the text part responsible for the mention of event.

➡ **X** *behaved as an authentic enhancer for* high-level **Y** *activity*

Event annotation - Difficulties (2/3)

Ex.6	Similar to its effect on the <u>induction of AP1</u> by okadaic acid, PMA <u>inhibits</u> the <u>induction of c-jun mRNA</u> by okadaic acid. (PMID1851743)
	<p>Event #11: Positive_regulation (GO:0048518)</p> <ul style="list-style-type: none"> oTheme : <i>c-jun mRNA</i> oCause : <i>okadaic acid</i> <p>Event #12: Negative_regulation (GO:0048519)</p> <ul style="list-style-type: none"> oTheme : Event #11 oCause : <i>PMA</i> <p>Event #13: Positive_regulation (GO:0048518)</p> <ul style="list-style-type: none"> oTheme : <i>API</i> oCause : <i>okadaic acid</i> <p>Event #14: Negative_regulation (GO:0048519)</p> <ul style="list-style-type: none"> oTheme : Event #13 oCause : <i>PMA</i>

- Some events can be identified even if they are not explicitly mentioned in text. (Event #14)



Event annotation - Difficulties (3/3)

Ex.5	In this report, we demonstrate that a novel Ets-related transcription factor, Elf-1, <u>binds specifically to</u> two purine-rich motifs in the HIV-2 enhancer. (PMID1527846)
	Event #9: Binding (GO:0005488) oTheme : <i>Elf-1</i> and <i>HIV-2 enhancer</i> Event #10: Binding (GO:0005488) oTheme : <i>Elf-1</i> and <i>two purine-rich motifs</i> (in HIV-2 enhancer)

- ❑ Sometimes the same event may be mentioned at different granularities. Event #9 and #10



GENIA event annotation – Stat (1/2)

□ Annotation

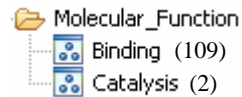
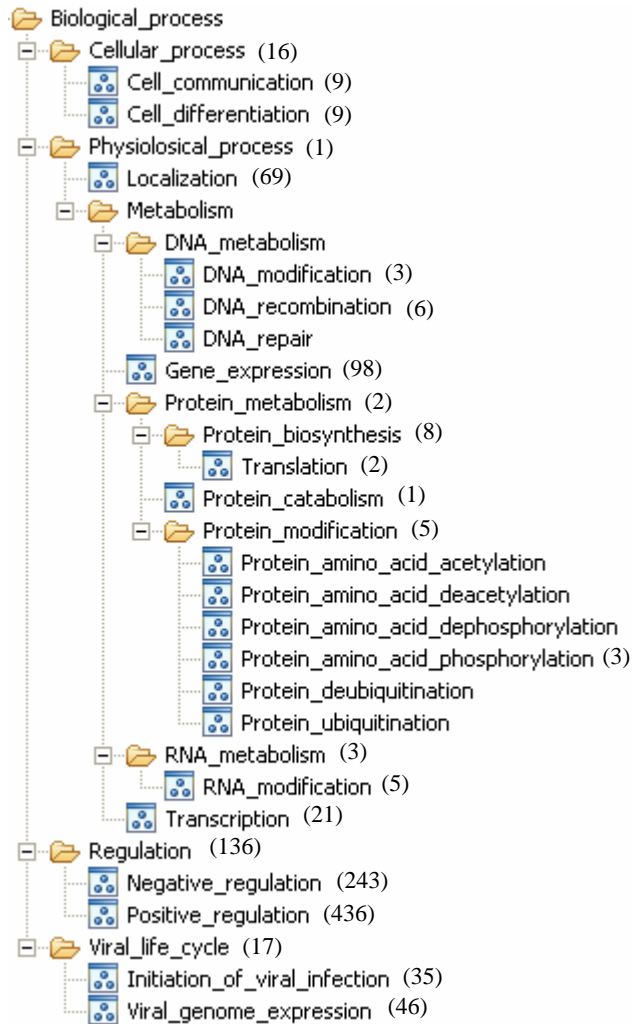
- ✓ 5 annotators + 1 manager with biology background.
- ✓ using XConc Annotation tool

□ 53 abstracts have been annotated

- ✓ # of sentences: 506
- ✓ # of sentences with events: 466
 - ➔ 92%
- ✓ # of events: 1,408
 - ➔ Avg. 3.02 events/sentence



GENIA event annotation – Stat (2/2)



Correlation (107)

Artificial_process (11)

Event Class	Freq.
Positive_regulation	436
Negative_regulation	243
Regulation	136
Binding	109
Correlation	107
Gene_expression	98
Localization	69
Viral_genome_expression	46
Initiation_of_viral_infection	35
Transcription	21
Cellular_process	16
Artificial_process	11
...	



GENIA annotation – current status

Publicly available

Internally available

Term	2000 abs.	390 abs.	} done
Part-of-speech	2000 abs.		
Syntactic Tree	600 abs.	700 abs.	} on-going
Event	55 abs.		
Coreference	(is being performed by Infocomm Research, 288 abs.)		} future
Relation			

- ❑ New release will be available soon.
 - ✓ Including internally available sets of annotation (except event annotation).
 - ✓ With many bugs fixed.
 - ✓ Together with XConc Suite



Conclusion

- ❑ Corpus annotation simulates automatic analysis of text.
 - ✓ Automatic analyzer can be developed by mimicking and generalizing the performance of manual annotation.
 - ✓ Automatic analyzer can be evaluated by being compared to the performance of manual annotation.
- ❑ Linguistic annotation
 - ✓ addresses the problem of finding linguistic structure of text.
- ❑ Semantic annotation
 - ✓ addresses the problem of finding knowledge pieces delivered by the text.
- ❑ GENIA annotation of biomedical literature
 - ✓ Semantic annotation
 - ➔ Entities, Functions&Events, Relations
 - ✓ Linguistic annotation
 - ➔ Part-of-speech, Syntactic structure, coreferences
- ❑ XConc Suite – an integrated corpus annotation tools