MANCHESTER 1824

The University of Manchester

Annotated Texts and Resources for Written Language

Junichi TSUJII

National Centre for Text Mining
Computer Science
University of Manchester

UK

Computer Science
University of Tokyo

JAPAN

GENiA

Speech and Written Text

- Speech
  - Observables
  - Obvious applications inside speech proper, Speech recognition, Speech synthesis
- Written Text
  - Non-observables, Semantics/Pragmatics
  - Machine Translation (inside written text)
  - Information Extraction, Semantic web, e-Publishing, Terminology, Ontology and Knowledge Management, Intelligent IR and Meta-data Extraction

- Multi-Linguality
- Multi-Domains
  - Biology/Medicine/Health
  - Financial Markets
  - Legal documents
- Multi-Tasks
  - Machine Translation, CLIR
  - Information Extraction, Automatic Summarization, etc.
  - Ontology and Knowledge Management, Semantic Web
  - E-Journal, Knowledge Discovery

Text Annotation

- Task-oriented Annotation
  - Bio-Creative annotated text
  - System development
  - Defined by specific tasks
    - Specific curation tasks in specific environments
    - Mapping of Protein names to database IDs in specific text types
    - Specific event types such as Protein-Protein Interaction, in specific text types
    - Disease-Gene Association of specific sets of diseases

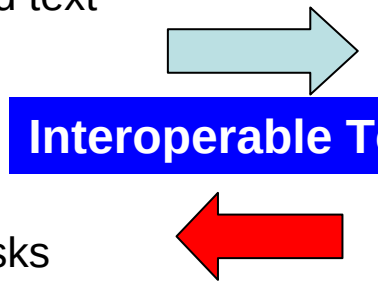
Market failures

Interoperable Tools

- Task-neutral Annotation
  - GENIA Corpus [U-Tokyo, NaCTeM]
  - Development of generic tools
  - Defined by theories
    - Linguistics
      - Tokens
      - POS
      - Phrase Structure
      - Dependency Structure
      - Deep Syntax (PAS)
    - Biology/Medicine/Health
      - Named Entities of various semantic types
      - Events
    - Linguistics + Biology/Medcine/Health
      - Co-references

Text Annotation

- Task-oriented Annotation
  - Bio-Creative annotated text
  - System development
  - Defined by specific tasks
    - Specific curation tasks in specific environments
    - Mapping of Protein names to database IDs in specific text types
    - Specific event types such as Protein-Protein Interaction, in specific text types
    - Disease-Gene Association of specific sets of diseases



- **Task-neutral Annotation**
  - **GENIA Corpus [U-Tokyo, NaCTeM]**
- **Development of generic tools**
  - **Defined by theories**
    - **Linguistics**
      - **Tokens**
      - **POS**
      - **Phrase Structure**
      - **Dependency Structure**
      - **Deep Syntax (PAS)**
    - **Biology/Medicine/Health**
      - **Named Entities of various semantic types**
      - **Events**
    - **Linguistics + Biology**
      - **Co-references**

Text Annotation

- Task-oriented Annotation
  - Bio-Creative annotated text
  - System development
  - Defined by specific tasks
    - Specific curation tasks in specific environments
    - Mapping of Protein names to database IDs in specific text types
    - Specific event types such as Protein-Protein Interaction, in specific text types
    - Disease-Gene Association of specific sets of diseases

Interoperable Tools

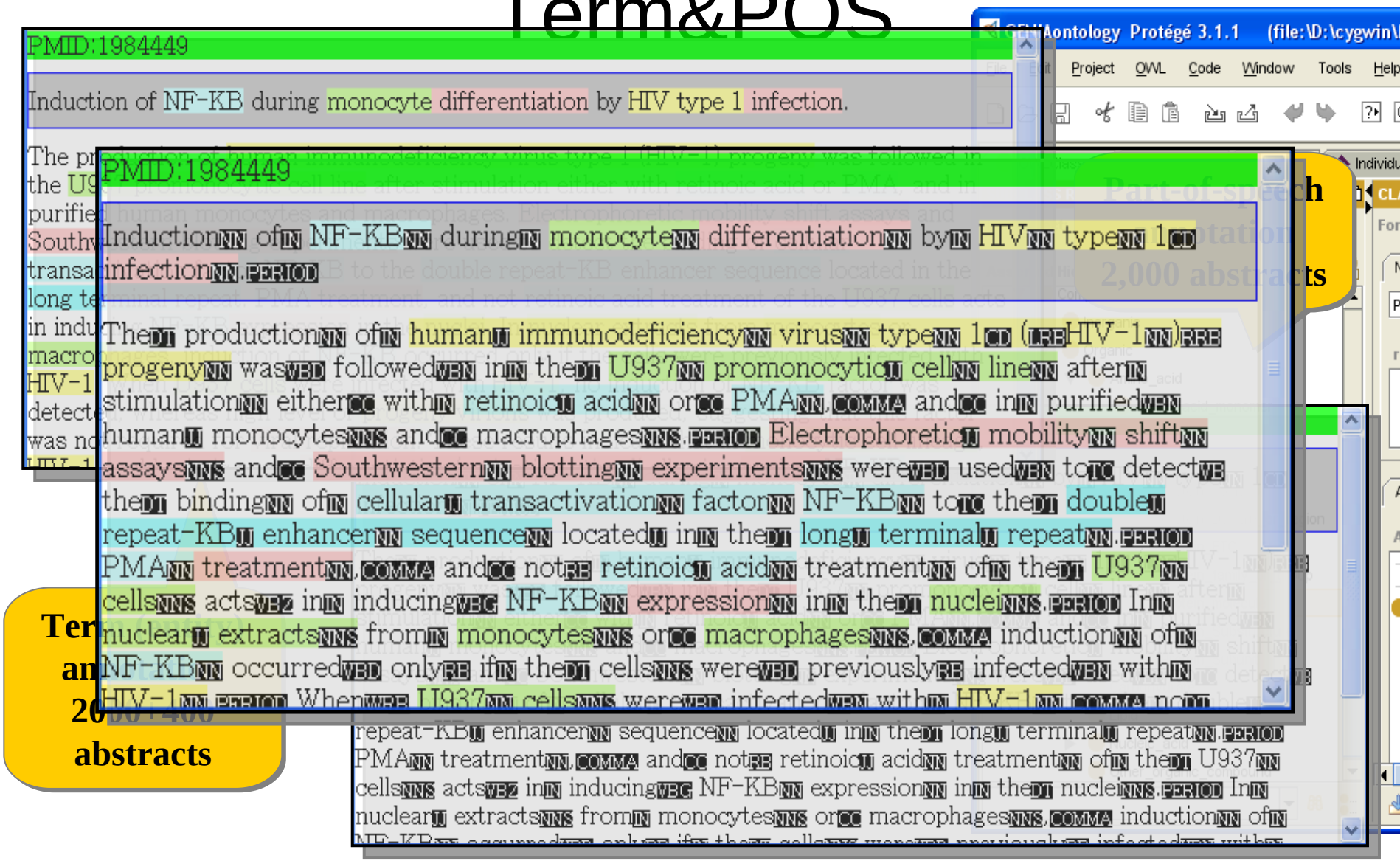
- **Task-neutral Annotation**
  - **GENIA Corpus [U-Tokyo, NaCTeM]**

Development of generic tools

- **Defined by theories**
  - **Linguistics**
    - **Tokens**
    - **POS**
    - **Phrase Structure**
    - **Dependency Structure**
    - **Deep Syntax (PAS)**
  - **Biology/Medicine/Health**
    - **Named Entities of various semantic types**
    - **Events**
  - **Linguistics + Biology**
    - **Co-references**

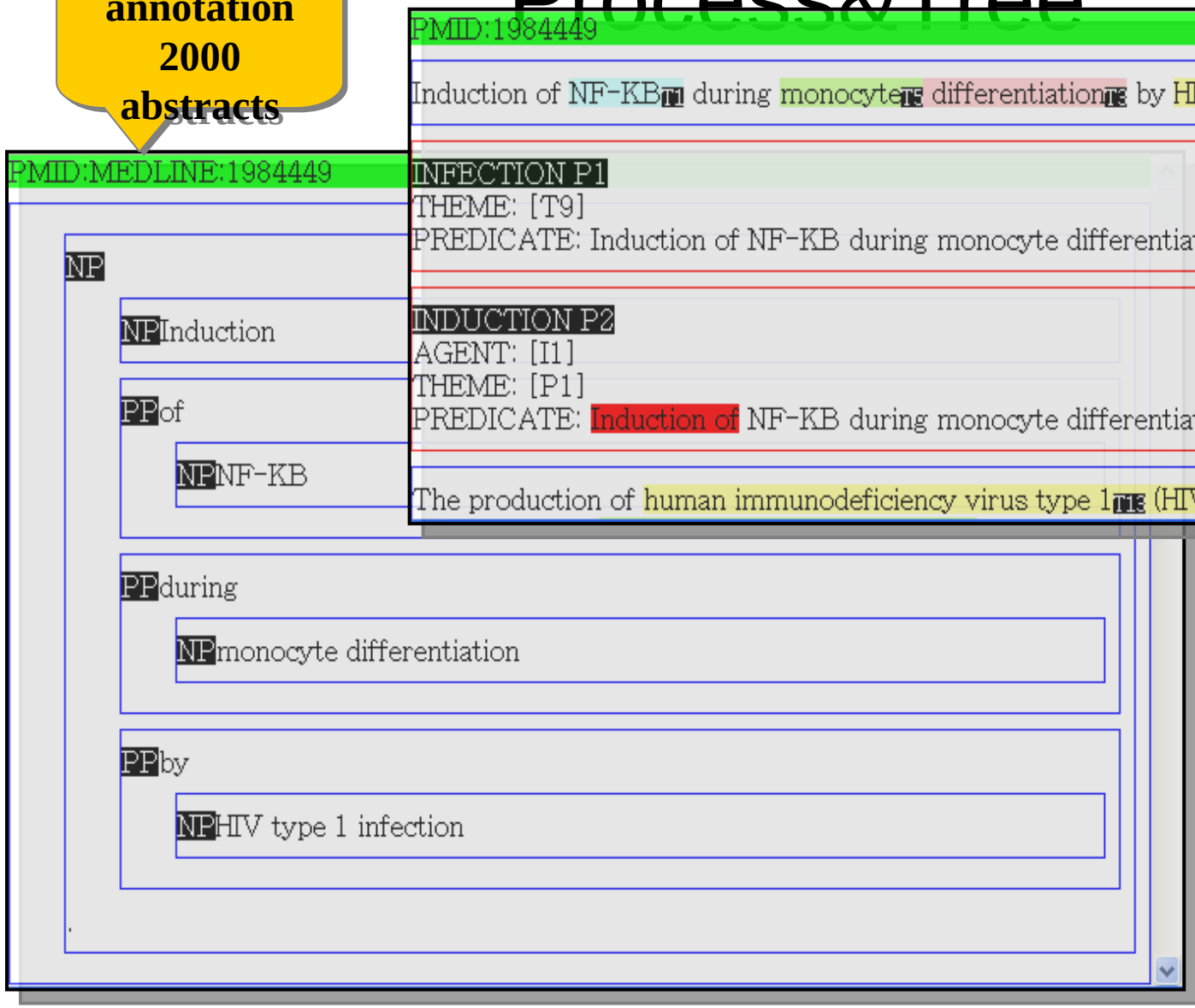
How far should we go from linguistics to conceptual contents ?

Annotation of GENIA corpus – Term&POS



Annotation of GENIA corpus – Process&Tree

Tree annotation 2000 abstracts



Process annotation 500 abstracts by May 2006

1000 abstracts by Dec. 2006

MANCHESTER 1824

The University of Manchester

Task-neutral Annotation

- Tool Development
  - Training, Development, Test
  - **Domain Adaptation**

GENIA

Tool1: POS Tagger

The peri-kappa B site mediates human immunodeficiency
DT NN NN NN VBZ JJ NN
virus type 2 enhancer activation in monocytes …
NN NN CD NN NN IN NNS

- General-Purpose POS taggers, trained by WSJ
  - Brill's tagger, TnT tagger, MX POST, etc.
  - 97%
- General-Purpose POS taggers do not work well for MEDLINE abstracts

Errors seen in TnT tagger (Brants 2000)

A chromosomal translocation in …

DT JJ NN IN

… and membrane potential after mitogen binding.

CC NN NN IN NN ~~JJ~~

… two factors, which bind to the same kappa B enhancers…

CD NNS WDT ~~NN~~ TO DT JJ NN NN NNS

… by analysing the Ag amino acid sequence.

IN VBG DT ~~VBG~~ JJ NN NN

… to contain more T-cell determinants than …

TO VB ~~RBR~~ ~~JJ~~ NNS IN

Stimulation of interferon beta gene transcription in vitro by

NN IN JJ JJ NN NN ~~IN~~ ~~NN~~ IN

Performance of GENIA Tagger

- GENIA tagger

Training corpus	WSJ	GENIA
WSJ	97.0	84.3
GENIA	75.2	98.1
WSJ+GENIA	96.9	98.1

(Ref.) TnT tagger

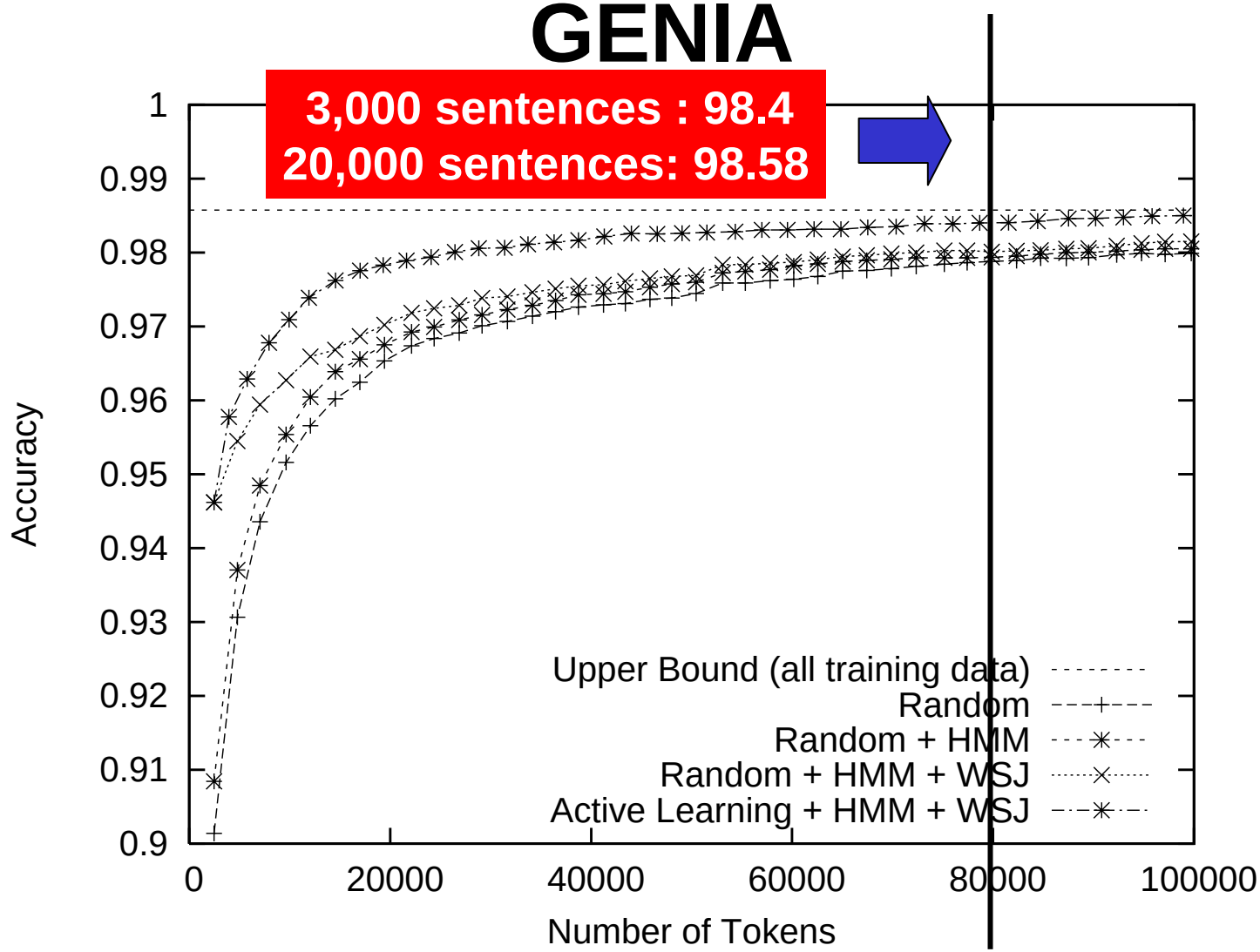
Training corpus	WSJ	GENIA
WSJ	96.7	84.3
GENIA	80.1	97.9
WSJ+GENIA	96.5	97.5

No degradation of the tagger trained by the mixed corpus

Some degradations (0.2 ~ 0.4) were observed, compared with the taggers trained by “pure” corpora

CRF-based POS + Active Learning

GENIA



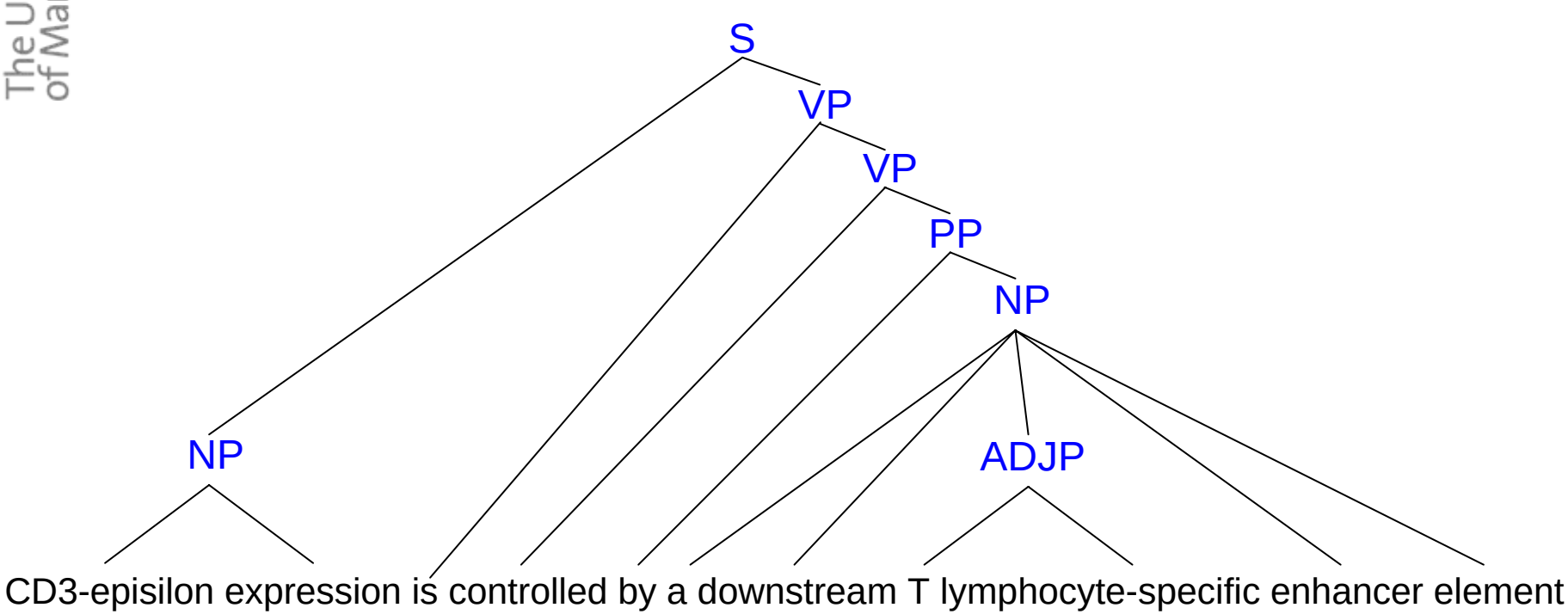
MANCHESTER 1824

The University of Manchester

Tool2　Full &Deep Parser Enju

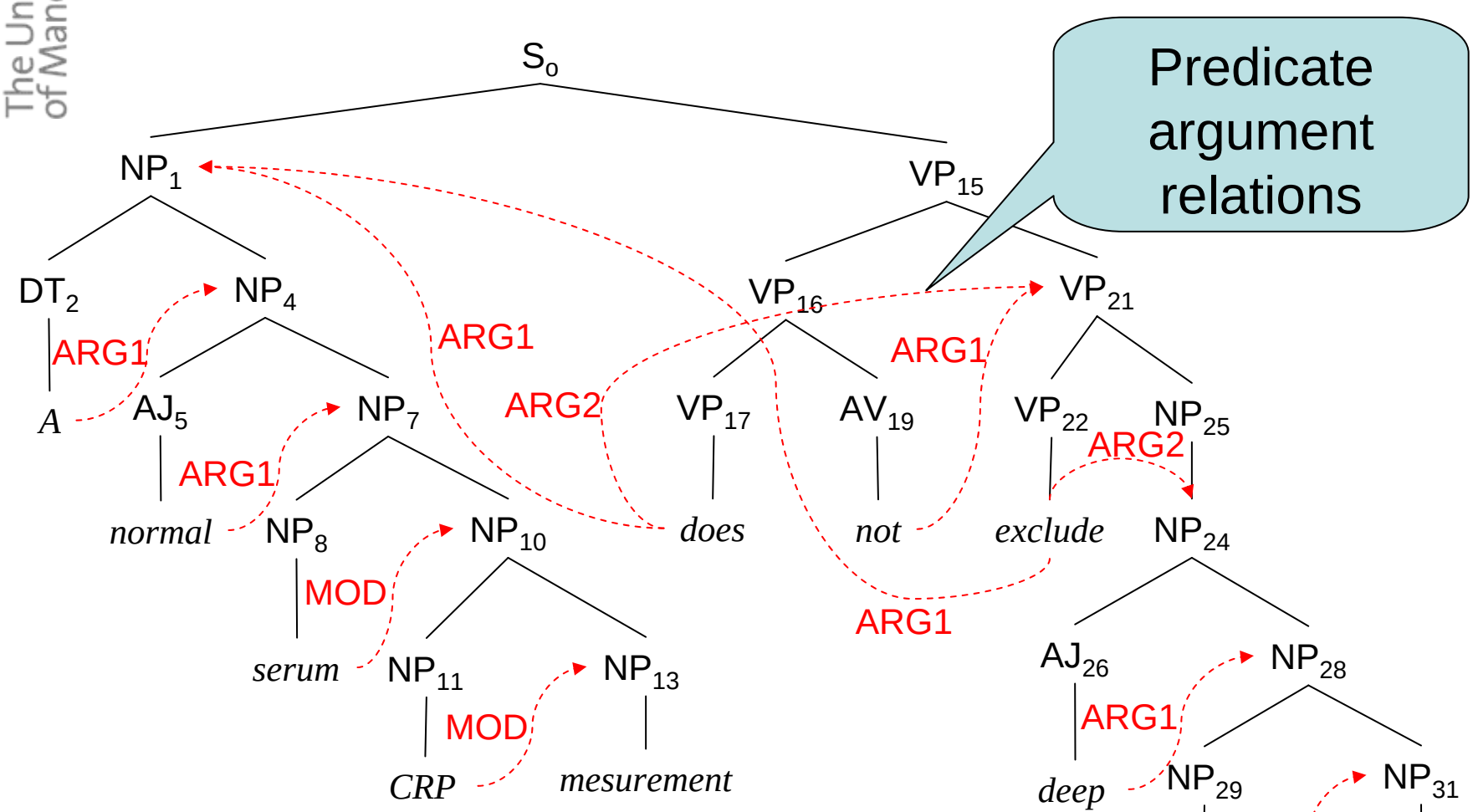
GENIA

GENIA treebank



- Based on the standard of the Penn Tree Bank

Semantic structure



MANCHESTER 1824

The University of Manchester

Performance of Semantic Parser

	Penn Treebank	GENIA
Coverage	99.7%	99.2%
F-Value (PA relations)	87.4%	86.4%
Sentence Precision	39.2%	31.8%
Processing Time	0.68sec	1.00sec

GENIA

MANCHESTER 1824

The University of Manchester

Adaptation with Reference Distribution

Lexical Assignment

Syntactic Preference

$$p_E(t \mid \mathbf{w}) = \frac{1}{Z_{\mathbf{w}}} \prod_{w_i \in \mathbf{w}} p_{lex}(l_i \mid w_i) \cdot q_{syn}(t \mid \mathbf{l}),$$

$$Z_{\mathbf{w}} = \sum_{t \in T(\mathbf{w})} \prod_{w_i \in \mathbf{w}} p_{lex}(l_i \mid w_i) \cdot q_{syn}(t \mid \mathbf{l})$$

$$p_M(t \mid s) = \frac{1}{Z'_s} p_0(t \mid s) \exp\left(\sum_j \rho_j g_j(t \mid s) \right)$$

Feature function

Feature weight

Original model

GENiA

MANCHESTER 1824

The University of Manchester

	F-score	Training Time (Sec)
Baseline (PTB-trained, PTB-applied)	89.81	0
Baseline (PTB-trained, GENIA-applied)	86.39	0
Retraining (GENIA)	88.45	14,695
Retraining (PTB+GENIA))	89.94	238,576
Structure with RefDist	88.18	21,833
Lexical with RefDist	89.04	12,957
Lex/Structure with RefDist	90.15	31,637

GENIA

	P	R	F	σ_F	AUC	σ_{AUC}
without evaluating self	60.4	69.3	64.3	4.3	0.879	0.026
without predicting self	60.4	65.6	62.7	3.5	0.834	0.032
with the prediction of self	57.8	66.1	61.4	3.9	0.914	0.020
(Miyao et al., 2008)	54.9	65.5	59.5			
(Airola et al., 2008)	52.9	61.8	56.4	5.0	0.848	0.023
(Sætre et al., 2007)	64.3	44.1	52.0			
(Mitsumori et al., 2006)	54.2	42.6	47.7			
(Yakushiji et al., 2005)	33.7	33.1	33.4			

Table 5: Comparison with previous results of the PPI extraction methods with the abstract-wise 10-fold cross validation on the AImed corpus

MANCHESTER 1824

The University of Manchester

Event Annotation

GENIA

GENIA event annotation

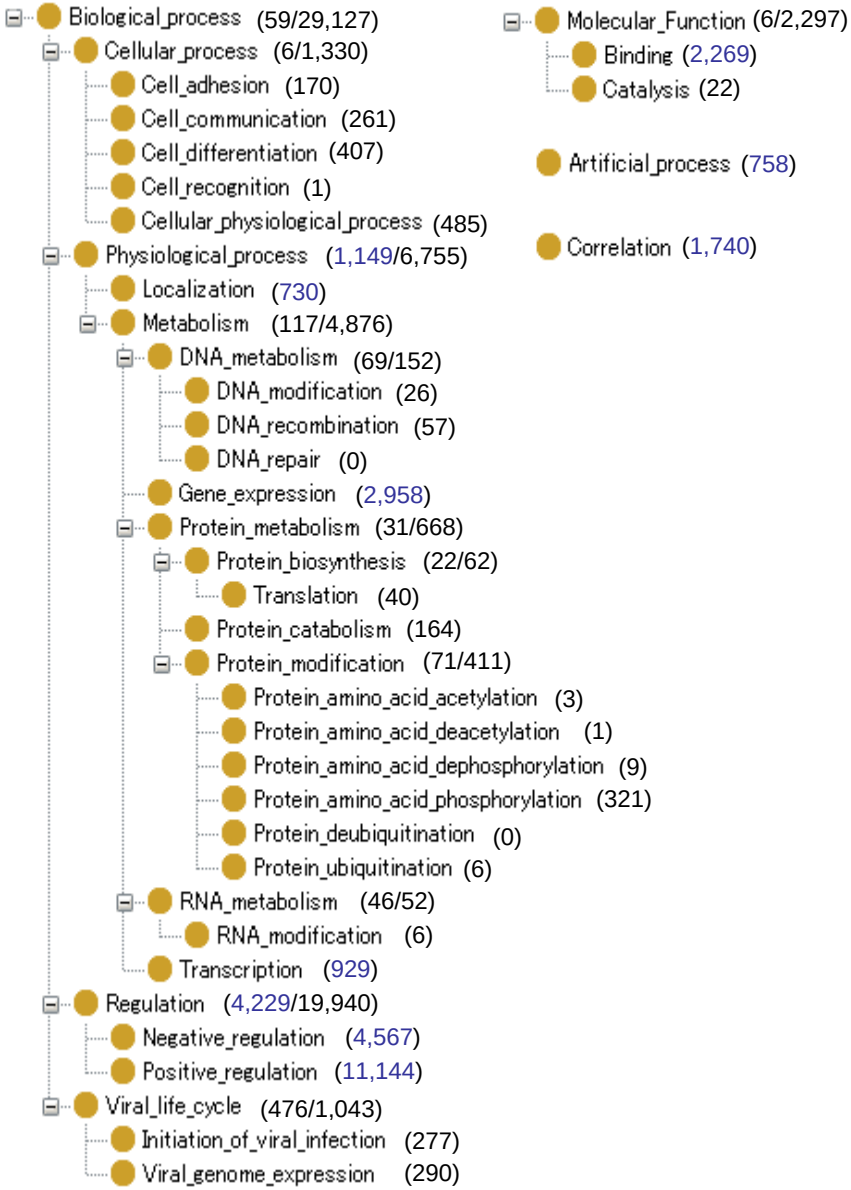
- Target of GENIA event annotation
  - Corpus
    - Part of GENIA corpus which is taken from PubMed using the MeSH terms, Human, Blood Cells and Transcription Factors.
  - Ontology
    - From the Gene ontology, concepts required for describing NF*k*B pathway have been selected (34 terms).
    - 3 additional concepts have been defined
      - Gene_expression
      - Artificial_process
      - Correlation

GENIA event annotation – Stat (1/2)

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

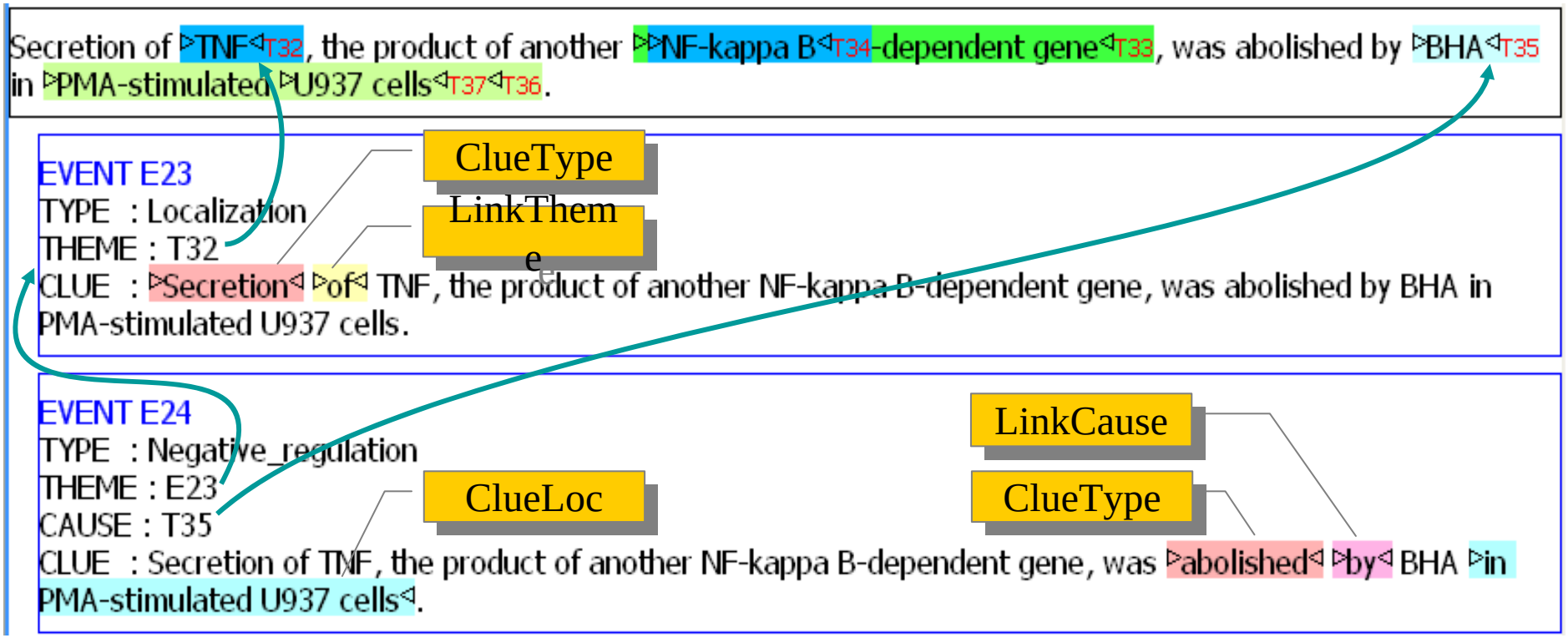
- 1,000 abstracts have been annotated
  - # of sentences: 8,981
  - # of sentences with events: 8,265
    - 92.0%
  - # of events: 34,065
    - Avg. 4.15 events/sentence

GENIA event annotation – Stat (2/2)



- Correlation
  - meaning ‘some’ relation between events.
- Artificial_process
  - Artificially performed processes.
  - Transfection, treatment, …
- Gene_expression
  - Transcription + Translation

GENIA Event Annotation - example

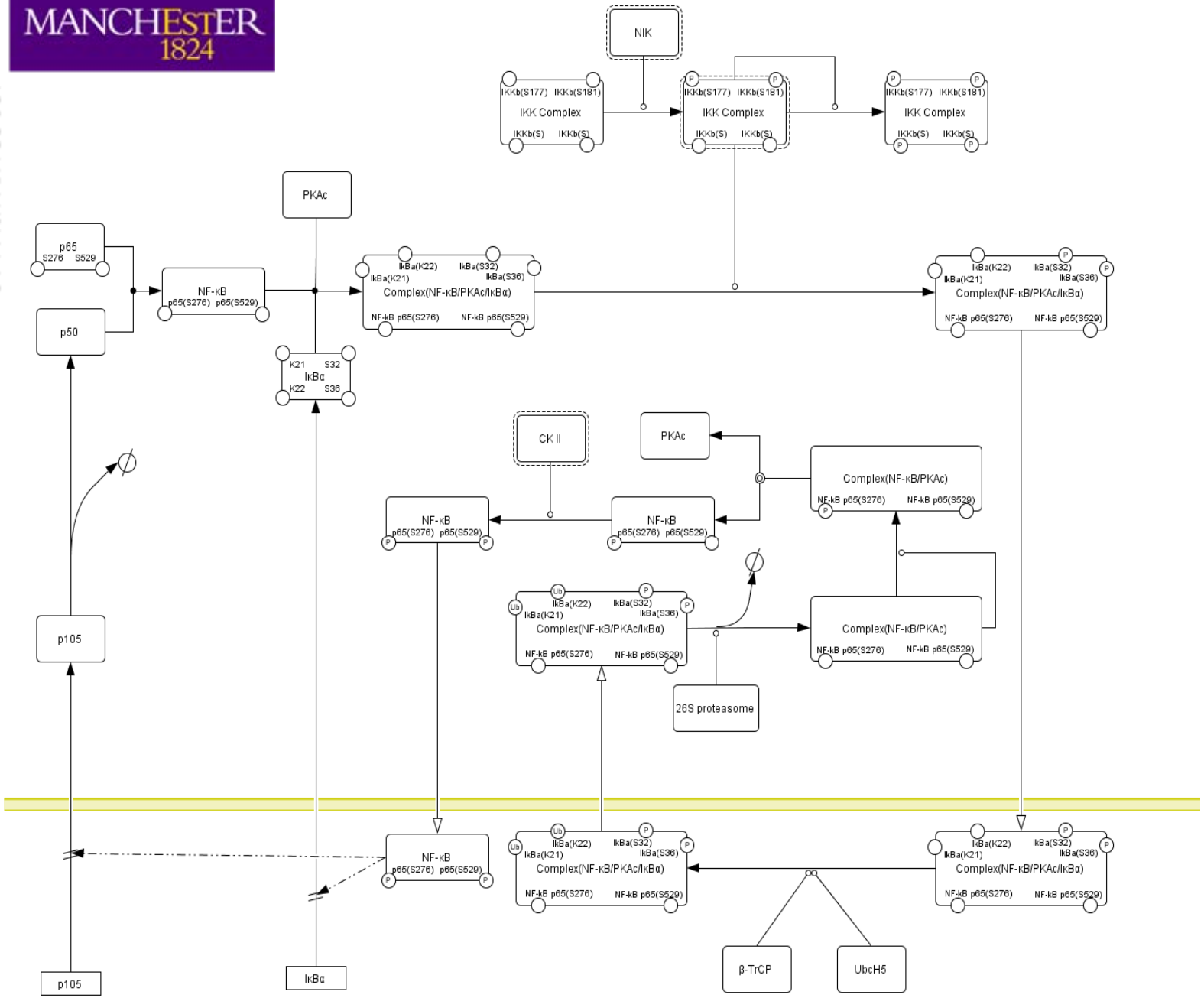


- For an identified event in the given sentence,
  - classify the **type** of events and record the text span giving the clue of it (ClueType).
  - identify the **theme** of the events and record the text span linking the theme to the event (LinkTheme).
  - identify the **cause** of the events and record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

Gene_expression

- Theme patterns observed (2,958)
  - Protein 2,308
  - DNA 591
  - RNA 25
  - Peptide 4
  - Protein Protein 2
  - Erroneous 27
- Keywords
  - coexpress, nonexpress, overexpress, express, biosynthesis, product, synthesize, constitute, …

Metabolic Pathways and Text Annotation


NIK
IKK Complex
IKK Complex
IKK Complex
PKAc
p65
p50
NF-κB
Complex(NF-κB/PKAc/IκBα)
Complex(NF-κB/PKAc/IκBα)
IκBα
CK II
PKAc
Complex(NF-κB/PKAc)
NF-κB
NF-κB
p105
Complex(NF-κB/PKAc/IκBα)
Complex(NF-κB/PKAc)
26S proteasome
NF-κB
Complex(NF-κB/PKAc/IκBα)
Complex(NF-κB/PKAc/IκBα)
β-TrCP
UbcH5
p105
IκBα

Discrepancy between event annotation and pathway annotation

React.	Type	expected annotation class	○	×
R1	Binding	Binding	11	56
R2	Binding	Binding	16	20
R3	Phosphorylation	Protein_amino_acid_phosphorylation	19	1
	+ site information	Protein_amino_acid_phosphorylation	2	5
	+ modifier information	+ (any) Regulation	4	1
R4	ubiquitination	Protein_ubiquitination	5	0
R5	degradation	Protein_catabolism	27	1
	+ modifier information	+ (any) Regulation	6	3
R6	translocation	Localization	26	0
R7	Binding	Binding	23	12
R8	gene expression	Gene_expression (or Transcription + Translation)	5	9
R9	gene expression	Gene_expression (or Transcription + Translation)	1	10
R10	processing	Protein_processing	5	12
		total	150	130

Recommendations

- Think globally and act locally:
  - Semantics/Pragmatics: Domain and Task Specific
  - Processing/Application Guided Approach
- Tools Development
  - Tools based on standard
  - Adaptation methods for different domains
  - Annotation Tools for active learning
- Theory guided standardization
  - Semantics/Pragmatics: Theories on un-observables
  - Non-trivial Ontologies

Tools and Corpora Available at

- National Centre for Text Mining
  - http://www.nactem.ac.uk/
- Tsujii Lab, Univ. of Tokyo
  - http://www-tsujii.is.s.u-tokyo.ac.jp/

- TerMine, AcroMine, AcroDisamb, GENIA Tagger, Shallow Parser, Deep Parser, Named Entity Recognizers, Event Recognizer (to be released),
- MEDIE, Info-Pubmed
- Interoperable Software Platforms (UIMA)
  - Contact: kano@is.s.u-tokyo.ac.jp